

Assessing individual lessons using a generic teacher observation instrument: how useful is the International System for Teacher Observation and Feedback (ISTOF)?

Daniel Muijs¹ · David Reynolds² · Pamela Sammons³ · Leonidas Kyriakides⁴ · Bert P. M. Creemers⁵ · Charles Teddlie⁶

Daniel.Muijs@ofsted.gov.uk

¹ Ofsted, London, UK

² University of Swansea, Swansea, UK

³ University of Oxford, Oxford, UK

⁴ University of Cyprus, Nicosia, Cyprus

⁵ University of Groningen, Groningen, The Netherlands

⁶ Louisiana State University, Baton Rouge, USA

Abstract

Teacher effectiveness, which impacts student attainment even when controlling for student characteristics, is of key importance as a factor in educational effectiveness and improvement. Improving the quality of teaching is thus the primary means by which we can enhance student learning outcomes. Thus there has long been great interest in the development of classroom observation measures in the field of educational effectiveness research (EER). The International System for Teacher Observation and Feedback (ISTOF) is a unique instrument in the field, as it was developed by a team from 20 countries using an iterative Delphi process to ensure cross-cultural relevance and validity. While previous studies have looked at psychometric properties of the instrument, they have not interrogated the extent to which ISTOF is useful for evaluating individual lessons and providing feedback to teachers. In this study, we observed three grade 4 mathematics lessons taken from the NCTE video library at Harvard University for this purpose. Findings show that ISTOF can provide a highly differentiated and fine-grained picture of individual lessons, but that the strengths of the generic approach in terms of breadth are to an extent counterbalanced by limitations such as the lack of attention to content richness.

Keywords Classroom observation · Teacher effectiveness · ISTOF · Mathematics · Educational effectiveness

1 Introduction

Classroom practice and teacher effectiveness have long been major areas of study within the field of educational effectiveness, as empirical studies have shown that the classroom level explains more of the variance in student attainment than the school level, and that within the classroom it is what the teacher does that matters (Muijs et al. 2014; Coe et al. 2014). In educational effectiveness research, therefore, classroom observation has traditionally been closely connected to attempts to uncover those factors that distinguish more from less effective teachers in terms of pupil attainment, and it is this tradition from which the observation instrument discussed in this paper has emerged. Four main methods have been proposed to evaluate teacher effectiveness: value added attainment measures, classroom observation, teacher surveys and student feedback (Coe et al. 2014), all of which have specific advantages and disadvantages, with observation seen as particularly useful as it allows detailed data on teacher behaviours to be collected (Chapman et al. 2015).

Early effectiveness research borrowed instruments from the US process–product teacher effectiveness studies of the early 1970s, such as the Flanders Interaction Analysis observation schedule (Flanders 1970), which were, however, somewhat limited as they tended to focus mainly on direct instruction approaches. For this reason educational effectiveness researchers increasingly developed bespoke instruments, the high-inference ‘Classroom Observation Instrument’ (COI) developed for the Louisiana School Effectiveness Study being one early example (Teddlie et al. 1989). This was also true outside the US, where Mortimore et al. (1988), for example, developed a bespoke instrument for their groundbreaking school effectiveness study in London, while in continental Europe the 1995 TIMSS video study inspired a range of work evaluating teaching using the three-factor model (e.g., the Pythagoras study, Klieme et al. 2009). While generally, these were generic, some instruments were adapted to specific subjects, for example the MECORS instrument, used in a large-scale evaluation and teacher effectiveness study of mathematics in primary schools (Muijs and Reynolds 2000). What characterised most of these instruments and studies is that they were typically developed by a small group of researchers in a limited set of contexts, which led to problems when attempting to use them in

international studies or broader contexts. A key example of this, which directly led to the development of the ISTOF instrument described in this paper, was the ambitious International School Effectiveness Research Project (ISERP), an international comparative study that utilised the US-developed Virgilio Teacher Behavior Inventory (VTBI) and Quality, Appropriateness, Incentives and Time (QAIT) observation instruments to look at classroom practices in the USA, Canada, Australia, Hong Kong, Taiwan, The Netherlands, Norway, the UK and Ireland. These instruments proved highly problematic outside of an Anglo-American context, as it became clear that countries defined key components and measures of effectiveness differently, and that many items suffered from ceiling effects, leading to a lack of differentiation between teachers. There was also a tendency for the instruments to focus very strongly on direct instruction, with little emphasis on self-regulated learning (Reynolds 2006; Reynolds et al. 2002). These issues eventually led a group of effectiveness researchers to undertake the process of developing the International Systematic Teacher Observation and Feedback (ISTOF).

In the following sections, we first describe how ISTOF was developed, and review previous studies that used the ISTOF observation instrument. We then describe how the instrument was used to analyse videos of three lessons, describing the methodology and results of these analyses with a view to assessing the usefulness of the ISTOF instrument. We present findings for each of the three lessons, cross-lesson analysis and what ISTOF captured well and less well. This allows us to draw some conclusions on the value of the instrument, in the final section.

2 Developing ISTOF

There were a number of reasons for developing ISTOF (Teddle et al. 2006):

1. To develop an instrument that would work across borders in international school effectiveness studies;
2. To develop classroom observation and the study of teacher effectiveness internationally, especially in those countries in which little research existed in the field;
3. To act as a precursor to international studies of teacher effectiveness.

The aim of ISTOF was therefore to develop a model of effective teaching that was internationally valid, and therefore based on our best knowledge of factors associated with pupil learning. This was to be done by drawing on international expertise on those factors related to better learning outcomes across a range of countries. ISTOF was conceived and designed as a generic teacher effectiveness instrument, aimed at measuring those factors that were hypothesised to generalise across subjects.

ISTOF was from the outset structured as an international, collaborative effort, and was also intended to enable the provision of formative feedback on teaching as well as the collection of research data (Teddle et al. 2006). The development team consisted of members from twenty countries that volunteered to take part (Argentina, Belarus, Belgium (Flanders), Brazil, Canada, Chile, China, Cyprus, Denmark, Finland, Germany, India, Ireland, Japan, Malaysia, The Netherlands, Nigeria, Norway, South Africa, Taiwan, Turkey, the UK and the USA), organised into a number of committees led by a central committee, but all contributing to each phase of instrument development. In each country leading educational effectiveness researchers formed the committee, recruited through personal contacts and at international conferences such as the International Congress on School Effectiveness and Improvement and the American Educational Research Association Annual Meeting. An iterative, multiple step, Internet-based, “modified” Delphi technique (Teddle et al. 2004) was used to construct the instrument. “Modified” here means that the ISTOF queries asked experts their opinions about what constitutes “effective teaching,” whereas the original Delphi technique asked experts to forecast events in the future (e.g., Gordon and Helmer 1964; Heylighen 2003). The process started with a number of queries from the central committee to the country teams, using an iterative process leading to ever more focussed queries. This process was used to generate the components, indicators and items for the final instrument. An example of such a query is the following:

Query 1: There are broad areas of effective teaching that have been identified by researchers and other experts in countries around the world. Please note that when we use the term effective teaching or teacher effectiveness, we are interested in what goes on in the classroom between the teacher and the students. Teacher Observation and Feedback Experts typically identify 3–6 of these components. One common component, for instance, might be Classroom Management. What are the broad components of teacher effectiveness in your country? That is, what do experts in your country consider to be the broad, general components of effective teaching? We recommend that you limit the total number of responses to this question to six components or less (Teddle et al. 2006).

It was made clear to the experts that effective teaching referred to identifying those teaching factors related to pupil learning, most typically attainment on various assessments. Altogether the teams generated 103

components (the range of responses per country was from four to ten). These responses were content analysed by two different teams of analysts, and the resulting lists were reconciled. Members of the central committee and the analysis committee performed separate constant comparative analyses on the 103 components using the qualitative data analysis program ATLAS. The constant comparative method is a qualitative analysis approach that originates in grounded theory. It combines explicit coding of the data with theory generation through a process in which any newly collected data are compared with previous data. This is a continuous ongoing procedure as theories are formed, confirmed, or discarded as a result of any new data that emerge from the study (Glaser and Strauss 1967). The two lists generated by these separate constant comparative analyses were then reconciled via telephone conversations between members of the two teams. Responses were thus distilled into 11 components:

- 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),
- classroom climate (the extent to which the teacher communicates high expectations, 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 practise and contribute to the schools' learning community and the teaching profession), and
- 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 components, planning of single lessons, long-term planning, teacher knowledge, and teacher professionalism and reflectivity are not observable so the actual observation schedule contains seven components.

Once the components had been agreed upon, country teams were invited to submit up to five indicators for each component, and up to 5 items per indicator. The country teams generated 65 indicators and almost 750 items, which were consolidated and reduced to 43 indicators by the central and analysis committees. Country teams were then asked to assess the importance of each of the 43 indicators, and a generalizability study of these ratings was completed.

Four separate analyses were conducted: an analysis of the responses of all 257 members of the country teams to the Likert scales, an analysis of the responses of all 257 participants to the rank ordering, an analysis of the aggregated responses of the countries to the Likert scales, and an analysis of the aggregated responses of the countries to the rank ordering. The results were highly congruent across all the analyses, with rank orders being significantly and strongly correlated within and across countries. An item bank was then developed comprising multiple items for each indicator. Country teams were asked to rate the items in the item bank (which was composed of alternative phrases or "stems" for assessing the indicators). The central team then generated the final version of the teacher observation protocol, which contains 45 items, based on the results of this last query to country teams. Training protocols and guides were produced (Teddle et al. 2006).

The final observation instrument consists of seven components. Each component has between two and four indicators, and each indicator consists of two items. An example can be given for the component 'differentiation and inclusion'. This component consists of the indicators '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 items: '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 item 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), with a 'not applicable' category also available. This was included as not all items 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 (as there may not be any seatwork or practice). The

instrument was translated from English into the host language, and then independently back-translated, followed by expert committee review to check for semantic and conceptual equivalence.

Theoretically, the core international team primarily drew from a teacher effectiveness perspective to develop the original proposal, building on, e.g., Creemers (1994) model, which draws extensively on that of Carroll (1963). The instrument does not, however, draw merely on Direct Instruction approaches as was the case in earlier teacher effectiveness studies, but also incorporates development of self-regulated learning. For example, while an item such as *'The teacher presents the lesson with a logical flow that moves from simple to more complex concepts'* is grounded in a mastery learning perspective similar to that proposed by Carroll (1963), an item such as *'The teacher invites students to explain the different steps of the problem solving strategy that they are using'* relates to the cognitive component of self-regulated learning (SRL), while *'The teacher gives assignments that stimulate all students to active involvement'* reflects the motivational dimension of SRL (Dignath and Buttner 2008).

Overall three of the components (classroom climate, classroom management, clarity of instruction) are explicitly related to established teacher effectiveness models derived from a longstanding tradition of research on the relationship between teacher behaviours and pupil attainment, particularly in basic skills, which has generally supported direct or explicit instruction (Muijs et al. 2014; Brophy 1986), while the promoting active learning and metacognition and differentiation components derive from the more recent constructivist approaches to learning which have strongly influenced approaches to generating self-regulated learning (Tsai 2001). Instructional skills, a factor that focuses on questioning, engaging students and varied instruction, builds on both traditions, with both active learning and engagement and questioning skills. The importance of assessment has been posited by a range of learning theories, and has received a great deal of empirical support (e.g., Black and William 1998). As such, these components illustrate the divergent perspectives of the country teams, but also the instrument's grounding in research on the relationship between teacher behaviours and student learning, which is multifaceted and does not support one approach only.

3 Previous studies on ISTOF

Following development, the validity and reliability of the instrument were successfully established in a range of different contexts internationally (Kyriakides 2008). The instrument has subsequently been used in a number of studies which generally support its reliability and aspects of validity, but not necessarily a factor structure containing 7 factors to match the theoretical components. A study in Flanders, Belgium found that teacher effectiveness measured using the ISTOF observation instrument tended to be quite unidimensional, with the ISTOF items loading onto one 'overall' teacher effectiveness scale (Marciniak and Janssen 2012). An interesting study by Ko (2010) compared the strengths and weakness of ISTOF observation instrument with those of the Quality of Teaching (QoT) instrument (developed by van de Grift 2007) in a sample of teachers in Hong Kong. Confirmatory factor analyses supported the hypothesised multi-factor structure, which showed good model fit, though high correlations between factors suggest the possibility of an overarching higher order 'effectiveness' factor. Variability analysis showed that the instrument had good stability across classroom observations of the same teachers, who were observed between 15 and 23 times each, and was able to distinguish distinct patterns of behaviour between them observed teachers. Reliability tests based on Cronbach's alpha showed internal consistency for the components ranging from 0.70 to 0.97 for each. Overall levels of reliability and validity were similar to those of the QoT. A caution however is that the sample was limited to teachers from a single school. The observation instrument was also validated in Ireland, where discriminant and factorial validity were tested. Teacher effectiveness was found to be higher on most scales in co-educational and girls' schools compared to boys' schools and differed between components. The hypothesised factor structure was supported (Devine et al. 2010, 2013).

A number of studies have been conducted in the UK, which have generally shown the observation instrument to be discriminating, but not necessarily to factor onto the seven proposed components. Ko and Sammons (2008) for example found 8 rather than 7 constructs using EFA, though the sample size was less than twice the number of items, making analysis problematic. Miao et al. (2015) used the instrument to compare instruction in China and the UK. The instrument showed good cross-country validity, with the study supporting the factor structure in both China and the UK. The ratings were strongly positively correlated with attainment in mathematics. The observation instrument was also used for the 'effective classroom practice' project, a mixed methods study of teaching practice (Kington et al. 2009), and in a study of 'Inspiring Teaching' in England (Sammons et al. 2014, 2016) where it was used alongside the QoT instrument, as well as employing qualitative observations, student survey measures and teacher interviews to explore the concepts of effectiveness and inspiring practice from different perspectives. In the Kington et al. (2009) study five factors were extracted, with Cronbach's Alpha's for these factors ranging from 0.55 to 0.84. The 'Inspiring Teaching' study, which looked at

teachers selected by their heads as being particularly effective and inspiring, showed that these teachers scored highly on the ISTOF factors, particularly (over 4.5 out of 5 on average) on classroom climate, assessment and evaluation, classroom management and providing clarity of instruction. It was also adopted in the evaluation of the Inspire Maths programme (Hall et al. 2016) and the evaluation of the Maximum Impact programme of Teach First (Muijs et al. 2010), both of which showed it as discriminating well in the English context, being able both to identify the most effective teachers, and to differentiate between items and factors, with teachers typically scoring higher on classroom climate and clarity of instruction than on promoting metacognition, for example. A study in the Spanish region of Mallorca supported the validity and reliability of the instrument, with the hypothesised factor structure showing good model fit using confirmatory factor analysis, and reliability of the factors ranging from 0.73 to 0.86. The instrument was able to discriminate between teachers in high and low performing schools (Reynolds et al. 2012). In Norway, the observation instrument demonstrated good predictive validity, being correlated 0.31 with student attainment in mathematics, and was seen as useful in professional development (Soderlund et al. 2015).

Overall, then, studies suggest that the ISTOF observation instrument shows sufficient variance. The instrument has been used successfully in studies in a number of countries, though the extent to which it is invariant across contexts is presently unknown. What is less clear is the extent to which it supports a seven-factor structure rather than components loading onto fewer (or even one) effectiveness factor, or whether they form a hierarchical structure with components loading onto an overarching effectiveness factor. The instrument generally shows good *discriminant* validity, allowing observers to distinguish more and less effective teachers well (e.g., Hall et al. 2016; Ko 2010; Muijs et al. 2010), and a number of studies have shown good *predictive* validity when relating the instrument to value added student attainment in a range of subjects (such as test scores in mathematics and home language or number of passing grades in national high stakes assessments) (e.g., Muijs et al. 2010; Kington et al. 2009; Kyriakides 2008), though the number of studies that have looked at this kind of validity is small. Internal consistency of subscales and aggregate scales is typically in the high range, with Alphas of over 0.8 being the norm (e.g., Ko 2010; Muijs et al. 2010; Devine et al. 2010; Kyriakides 2008). The Muijs et al. (2010) study looked specifically at interrater reliability and rater bias, and established that given sufficient training (2 full days, starting with theory and followed by practice on videotaped lessons), a Cohen's Kappa could be achieved of between 0.69 and 0.83 on the different items. What is noteworthy is that across countries teachers tend to score lower on the 'differentiation and inclusion' and 'promoting active learning and developing metacognition' components than on the other components in the model.

4 Research aims

In this paper we explore the use of ISTOF as a means of generating more detailed analysis of a limited number of lessons. Previous studies have typically aggregated ISTOF observations over a relatively large number of lessons, using psychometrics to analyse the validity, reliability and factor structure of the instrument. What they have not done, notwithstanding the fact that this was one of the original aims, is to delve into the detail of individual lessons, which is essential if we are to gauge ISTOF's usefulness for developmental purposes, such as providing feedback to teachers. In this paper we therefore study the advantages and disadvantages of ISTOF as a measure for understanding individual lessons, which should allow us to gain an insight into the extent to which the instrument can be useful for formative purposes.

5 Methods

The lessons analysed are three 4th grade mathematics lessons drawn from the NCTE video library at Harvard University. For a complete description of the lessons see Charalambous and Praetorius (this issue). The lessons rated in this analysis include a geometry lesson taught by a teacher named Mr Smith, a lesson on strategies for multiplication taught by a teacher named Ms Young, and a lesson on multiplying a fraction by a whole number taught by a teacher named Ms Jones (all teacher names are pseudonyms). All three teachers volunteered to have their lessons video-recorded and gave consent for their video-recorded lessons to be used for research purposes. Three observers, one of whom was one of the developers of the instrument, while the other two had not used it previously, observed each lesson and scored it following the observation. The two 'ISTOF novices' were trained in the use of ISTOF by the developer-observer. The training followed the ISTOF protocol, starting with a half-day session to get to understand the instrument and its background (theory) followed by practice, grading videos of classroom teaching followed by discussion and feedback, until a Cohen's Kappa of over 0.7 on each subscale was achieved on three subsequent lesson observations. The lessons shown were, like the target lesson for the study, mathematics lessons, but they were recorded in the UK rather than in the US.

Lessons were scored without the three observers knowing the value-added scores for the teachers involved, to ensure that judgements were unaffected by any prior estimation of teacher effectiveness. As outlined in the observation protocol, the observers took notes while watching the lesson videos, and completed the Likert scale

following observation. As lessons had been filmed from multiple angles, the lesson was watched from each angle (thus each lesson was observed three times) with notes taken before the observation sheet was completed and scored in accordance with the protocol (Teddlie et al. 2006). Where observers disagreed on a rating, the mean was calculated and rounded to the nearest whole rating. The observers then drew on their lesson notes to provide a description of the lesson which was used to inform the discussion of the strengths and weaknesses of ISTOF, not least with regard to those elements of lessons that were (not) addressed by the instrument.

6 Findings

In this section we review the findings for each of the three lessons that were analysed. We discuss the overall ratings on the seven components and what these findings mean with regard to teacher effectiveness. Where appropriate we discuss findings on the individual items.

6.1 Mr Smith's lesson

The ISTOF schedule deliberately does not produce an overall score for a lesson, as the aim of the instrument is in part to focus on differential effectiveness. However, as Table 1 shows, in Mr Smith's lesson three components (clarity of instruction, classroom climate and classroom management) scored close to the midpoint score of 3 on the 5-point scale; one (assessment and differentiation) reached a mean score of 4; and three scales (instructional skills, differentiation and inclusion, and promoting active learning and metacognitive skills) fell well below the midpoint score.

In terms of the assessment component, Mr Smith promptly corrected errors when questioning, and provided correction and explanation to the answers to the mainly closed recall questions he posed at the start of the lesson (both positive indicators in the ISTOF assessment component). At times as the lesson progresses explanation of wrong answers gave way to test preparation tips. Questions at the start of the lesson related to previously covered material and students generally seemed to know the answers. Throughout the lesson, questions clearly referred to what was learnt and to the content of a test to be taken. Much assessment took the form of asking students to come forward to work on the board, during which activity students were adequately corrected when making errors. These actions contributed to the overall positive ratings on the assessment items for Mr Smith. As time was seen to run out, Mr Smith stopped asking students to come forward to work on the board, and instead asked the whole class to shout out answers. Both board work and whole class questioning were therefore used by Mr Smith to assess understanding.

An area which was graded low was differentiation and inclusion. While the lesson was graded 3 (thus at the midpoint) for the item 'All students are actively engaged in learning', lower grades were observed for the other items in this lesson. In terms of engagement, it was noted that both during whole class and seat work the majority of students appeared on task, but there was evidence that some students were bored, sitting with their heads in their hands and yawning or gurning at the camera rather than engaging with the lesson. There was very little evidence of differentiation, with all students given the same task during seatwork, and the majority of questions addressed homogeneously to the whole class. There was limited group work or student to-student communication.

Clarity of instruction was graded above the midpoint. This was largely due to the fact that Mr Smith tended to communicate clearly, giving suitable explanations of key concepts and emphasizing accuracy in work. The lesson was well structured, with a clear progression from review of previous work to new material, and questioning becoming more complex towards the end of the whole class section of the lesson, which finished with individual practice. Transition points were typically well managed, though some time was wasted on distributing materials before seatwork. The start of the lesson consisted in the teacher getting students to repeat what they had done in the previous lesson. Stated objectives, which were scattered throughout the lesson, often focused on test preparation.

Table 1 Mean ratings by component lesson 1-Mr Smith

Component	Mean rating
Assessment and evaluation	4.0
Differentiation and inclusion	1.8
Clarity of instruction	3.3
Instructional skills	2.3
Promoting active learning & developing metacognitive skills	1.7
Classroom climate	3.2
Classroom management	3.0

In terms of instructional skills, Mr Smith generally engaged students in the lesson, and used an appropriate wait time when posing recall questions (with appropriate defined as a short wait time for quick recall questions and a longer period where greater complexity is sought), though in this lesson he was at times quick to provide answers to more complex questions rather than probing students further when wrong answers were provided. The vast majority of questions were closed recall questions, and Mr Smith did not make full use of opportunities to use questions that encourage thinking and elicit feedback, often preferring the shortcut of providing answers himself. The component on which this lesson scored lowest was that of promoting active learning and developing metacognitive skills. This was to a large extent due to the questioning strategies used, where Mr Smith did not tend to ask students to explain their answers or workings, either to each other or to him. During the whole class part of the lesson, which took up two-thirds of it, only once was a student asked to explain their calculations, and here Mr Smith finally provided the correct answer himself. There was little use of real-world examples in the lesson. The lesson was rated higher on the fact that students were presented with a number of different strategies to use, but they were not invited to reflect on the advantages and disadvantages thereof.

Finally, classroom climate and classroom management were both rated at the midpoint. There were some clear strengths in classroom climate, in that the lesson was highly interactive, with Mr Smith attempting to engage all students in question and answer sessions, and students were treated respectfully. On the other hand, expectations did not appear particularly high (for example, when a student attempted a different solution strategy Mr Smith commented that they were not yet expected to do it different ways and discouraged the student), and the lesson, not least the seatwork task, did not appear very demanding. Behaviour in the class was generally good, with very little disruption even at a low level, though Mr Smith did not address off-task behaviours and lack of concentration, which were observed during both the whole class and seatwork parts of the lesson.

6.2 Ms Young's lesson

Ratings for the ISTOF components in Ms Young's lesson are less differentiated than those for Mr Smith, with most clustering around the midpoint, and only assessment and evaluation substantively above the midpoint. As may be seen, however, these overall component scores at times mask large differences on individual items (Table 2).

Assessment and evaluation occurred throughout the lesson, as Ms Young used a lot of questioning during the whole class sections of the lesson and went around checking understanding during seatwork. Assignments given were very clearly related to what students had learnt, with questions early in the lesson relating to learning from previous lessons but then generating new questions where new strategies were practised. Answers were corrected promptly. The picture was more varied when it came to the differentiation and inclusion component. In this lesson Ms Young made extensive use of group work and there was a lot of task-oriented communication between students. Where the lesson was more of a mixed picture was in involving all students in learning. Ms Young tended to focus on those students who were keen to be involved, for example drawing primarily on students who were first to put their hands up to answer questions, but did not make much effort to draw in students who were less keen or may have found the content rather challenging. There was limited differentiation, as students received the same task and also the same homework at the end of the lesson, though it could be argued that by choosing particular students to answer questions (e.g., at the board) the teacher did differentiate, and Ms Young did differentiate in her interaction with the group. Students who had completed the assignment got a new task to do.

In terms of clarity of instruction, the lesson objectives were written on the board at the start of the lesson, and Ms Young generally showed good communication skills and gave detailed and comprehensive explanations. However, an issue here was that her sometimes complex explanations, while rich and challenging, were not always adapted to the level of all students in her class as evidenced by the observation that some students found the explanations hard to follow and appeared confused and in some cases became inattentive. The lesson was well structured, with consolidation of knowledge from previous lessons following the stating of the objectives.

Table 2 Mean ratings by component lesson 2–Ms Young

Component	Mean rating
Assessment and evaluation	4.2
Differentiation and inclusion	2.6
Clarity of instruction	3.2
Instructional skills	3.2
Promoting active learning &	

developing metacognitive skills	2.9
Classroom climate	3.6
Classroom management	3.3

This was followed by increasingly new and complex content, which led to group work tasks and then to whole class consolidation at the end of the lesson. A weakness was observed at the transition points, for example between whole class and seatwork, which tended to be somewhat chaotic and took up too much time. Students did not seem to follow clear routines for transitioning.

A clear strength of this lesson was the high level of questioning. Ms Young frequently asked students to provide explanations and she used open ended questions to develop understanding. Students were asked to solve quite complex problems at various points during the lesson. What again recurred as an issue was the extent to which this high level content was aimed at all students as opposed to the most able learners in the classroom. An interesting incident occurred where Ms Young wrote an error on the flipchart and waited for students to correct it. As they didn't do this, she eventually corrected it herself, which is illustrative both of the extent of open-ended challenges set and the tendency for the lesson to remain firmly teacher-driven.

Of the three lessons observed this was the one that was rated highest in terms of promoting active learning and developing metacognitive skills. Students were encouraged to use a variety of problem solving strategies. Ms Young both demonstrated different methods, and explicitly told students that they have a variety of tools to solve problems. Explicit problem-solving strategies were taught. Where Ms Young was less effective was in the group work part of the lesson, where the task was a bit unstructured, meaning that students were not explicitly encouraged to discuss solutions or correct each other's work. Furthermore, in this essentially teacher-driven lesson, there were limited opportunities for students to develop their own examples. Where real life context was introduced (selling apples in a shop), the example was somewhat contrived and more typical of classroom mathematics problems than actual reality.

The classroom climate component was an interesting one in this lesson. There were clear strengths here. Expectations were high, students were praised when they showed effort in the lesson, and Ms Young made it very clear that she expected effort and commitment from students. On the other hand, Ms Young did not show much warmth or empathy towards students, and did not make much effort to involve students who did not volunteer to come forward. Classroom management showed a similarly mixed picture. Ms Young certainly attended to misbehaviour, and tried to make sure all students were on task. However, her discipline style could be quite harsh, and she very explicitly singled out a number of students for sharp criticism at the end of the lesson. Rules were not always sufficiently clear, and Ms Young had to clarify the working rules of the group task after the task had commenced.

6.3 Ms Jones' lesson

Ms Jones lesson rated highly on clarity of instruction, classroom climate and management, and was scored highest of the three lessons on these aspects. However, the lesson rated less positively on instructional skills and promoting active learning and metacognition, and was the lowest rated of the three lessons on assessment and evaluation (Table 3).

While Ms Jones corrected mistakes, she did not always clarify why an answer was correct or not, at one point just commenting that a wrong answer sounded 'strange'. Assignments were not always clearly related to learning, and sometimes Ms Jones herself went somewhat off task, for example in giving a lengthy explanation of, and task relating to, baby burping that seemed rather tangential to the learning goals of this lesson.

The lesson showed some strengths in terms of inclusion, in that students were actively engaged and encouraged to be so by Ms Jones, and the seatwork task involved students in discussing their work with one another. There was less differentiation, however, with students given the same tasks both during whole class and seatwork parts of the lesson. The objectives of the lesson were clearly described at the start, and related to previously learned content. Ms Jones frequently checked understanding through questioning and by going round the tables during seatwork, again asking a lot of questions. The lesson was well structured, starting with revision of previously learnt content, followed by new content, seatwork practice and consolidation in the whole class setting, more new content, group work and consolidation.

Transitions were managed very well, with students well aware of procedures and expectations. Clarity of instruction was a strength of this lesson. Instructional skills were more problematic, however. While Ms Jones' assignments stimulated involvement of all students, her questioning tended to be focussed on recall and

simple procedures, and there was not much variation with regard to question difficulty. Too many questions did not appear to focus on mathematics content. Ms Jones did provide some explicit instruction on problem solving strategies, and at times invited students to explain their problem-solving strategies, but she did little to encourage them to explain answers to each other or correct each other's work, even during the group work task. The lack of open-ended questions gave students few opportunities to reflect. Ms Jones used real life examples relating to her own life, but these did not appear very connected to the mathematical content of the lesson (see the baby burping example above).

Table 3 Mean ratings by component lesson 2–Ms Jones

Component	Mean rating
Assessment and evaluation	3.5
Differentiation and inclusion	2.5
Clarity of instruction	4.2
Instructional skills	2.8
Promoting active learning & developing metacognitive skills	2.4
Classroom climate	4.1
Classroom management	4.1

While promoting metacognition was therefore not a strong feature of this lesson, there was evidence of good classroom climate and classroom management. Ms Jones was enthusiastic and had a good rapport with students, whom she treated respectfully. The lesson was highly interactive, and Ms Jones attempted to get a range of students to answer questions and get involved. However, while she never explicitly stated any low expectations, the often low level of tasks and questioning did not indicate particularly high expectations either.

In terms of classroom management, students generally showed high levels of on task behaviour. A strength of the lesson was the clear understanding by students of rules and procedures (for example, putting hands on heads to indicate finishing the task at the start of the lesson), and where off-task behaviour occurred Ms Jones corrected it promptly. However, while the lesson started on time in so far as one can tell from the video recording, it took a full six minutes of activities such as writing the title of the lesson in workbooks before any mathematical content occurred.

7 Comparing across lessons

These three lessons all demonstrated significant strengths and weaknesses. Collectively, they were strongest in the area of assessment and evaluation, and weakest in the areas of differentiation and inclusion and encouraging active learning and metacognition. The latter is in line with most of the international studies discussed earlier. The three lessons did demonstrate quite varied patterns, however. Ms Jones's lesson scored well on classroom management and classroom climate, but rather low on instructional skills and developing active learning and metacognition. Mr Smith's lesson tended to score mid to low on most components other than assessment and evaluation, while Ms Young's lesson was an interesting one in that relatively homogeneous component scores masked large differences on individual items. As may be seen below, while capturing a lot of variance between lessons, there were also some elements not well captured by the observation instrument.

8 What ISTOF captured well

A strength of the ISTOF instrument that was apparent in these analyses is its ability to distinguish between different lesson components rather than to draw broad conclusions that are susceptible to halo effects, which can be an issue with observation instruments such as the Virgilio Teacher Behavior Inventory (VTBI) and the Framework for Teaching (e.g., Halpin and Kieffer 2015; Reynolds et al. 2002). Both the components and individual items varied significantly within lessons. This allowed the instrument to make relatively fine-grained distinctions, and to draw useful conclusions regarding aspects of lessons that may be more or less present. For example, in all three lessons the component of differentiation was present only to a limited degree, which could point us in useful directions in terms of feedback and CPD, while all three were well structured, suggesting this aspect is well ingrained in the practice of these teachers.

Another strength of ISTOF in these analyses was the relatively demanding nature of most of the items, which means that the instrument was much less susceptible to ceiling effects than, for example, the instruments used during the ISERP study (see above). This also makes ISTOF more suited to distinguishing average from high quality teaching than is the case for many instruments. The broad nature of the instrument means it captures a variety of aspects of pedagogy, from classroom management to eliciting metacognitive thinking, which allow it to capture at least those elements of pedagogy that are generic or similar across subjects. If these three lessons

had been drawn from one school, it would be easy to envisage whole school professional development activities focussed on greater student involvement and differentiation in lessons, for example.

9 What ISTOF did not capture well

While the generic nature of ISTOF is a strength, it is at the same time a key weakness, in that the aspect that was most weakly captured here was subject content. In observing the three lessons, it is clear that there are substantive differences between lessons in terms of accuracy and richness of mathematical content, with Ms Young's lesson being particularly strong in this respect, while in Ms Jones' lesson mathematical content was low level, and a number of misconceptions were not adequately dealt with. A further issue is that while ISTOF was developed out of a multi-country consensus, there could nevertheless be room for criticism of the concepts and measures in that they do not reflect more teacher-centred or direct instruction approaches well. As these have been found to be effective in terms of the development of basic skills in particular (Muijs et al. 2014) this may mean that teachers using techniques that are effective for particular goals may be unduly penalised by the ISTOF measure.

A further factor that should be taken into account when observing lessons may be the specific developmental level of the teacher. There is some evidence that teachers develop strengths in different areas (such as classroom management and metacognition) at different career stages (Antoniou et al. 2015), so it may be inaccurate to expect teachers at different career stages to rate highly on all components. Finally, of course, there are clear dangers to basing judgements on one observed lesson, which will have particular goals and sit within a sequence of instruction on a particular topic.

10 Discussion and conclusion

Classroom observation instruments have long formed an integral part of educational and teacher effectiveness research and of many systems for teacher development. ISTOF, a relative newcomer to the field, is an interesting example, having been developed through an international iterative process. This study provides a particular form of validation of the instrument, in that where previous studies have generally looked at reliability and validity in the aggregate, this study has allowed us to assess the validity and usefulness of the measure as an instrument for more fine-grained analysis of individual lessons. As such, even though the lessons observed were all from a US context as opposed to the international one the instrument was designed for, and scale score comparisons are not meant to be based on measurement procedures with proven equivalence, they have provided useful findings that are additional to those of the studies reported above. The study suggests that ISTOF can provide useful information on lessons, that is sufficiently differentiated to allow for formative feedback, and sufficiently demanding not to be overly susceptible to ceiling effects. As such, ISTOF can be a useful *component of* a professional development and evaluation framework in schools by providing formative feedback that can inform professional development programmes and priorities.

It is important here that we emphasise the need for ISTOF (and other observation instruments) to be embedded within a broader framework. The issues that ISTOF does not pick up are important. Content richness matters, and is related to both student learning and teacher subject knowledge (Coe et al. 2014). The latter was deliberately not included in the observation instrument, but is, alongside planning and teacher professionalism, an element of the overall ISTOF framework, and we would therefore recommend that these elements, too, be taken into account when using the instrument.

ISTOF is also a relatively complex instrument to use. The 45 high inference items require a good knowledge of the protocol by observers, and training, practice and the establishment of inter-rater reliability are important, even more so as the stakes rise where ISTOF is incorporated into evaluation or accountability systems. We would not recommend using ISTOF for such purposes without establishing reliability with the observers using the instrument, and without multiple observations. For the reasons discussed above, one lesson will never be sufficient to judge the effectiveness of a teacher as opposed to the effectiveness of teaching in a particular lesson, and we would recommend observing multiple lessons to establish sufficient reliability, since reliability has been found to increase asymptotically with the number of observations. Exact numbers are not easy to establish, with one well designed study showing that reliability increases rapidly up to around 5 or 6 observations with smaller increases for subsequent observations (Sterbinsky and Ross 2003), while a large-scale Dutch study suggests that if high reliability is to be achieved ($Ep^2 \geq 0.90$), up to 10 observations may be required (van der Lans et al. 2016). However, if observations are taken over a prolonged period, lack of trait invariance may occur, for example due to genuine changes in teacher behaviours (Meyer et al. 2011).

More generally, there are dangers to relying solely on observation for accountability purposes, so again the recommendation would be to use the instrument as part of a framework that could include student attainment

(value added), student views, and factors such as professionalism, collegiality and subject knowledge that are not part of the instrument (Coe et al. 2014). This study then provides some further insight into the value, but also the limitations, of using classroom observation instruments, as many of the strengths and weaknesses mentioned relate to issues with observation in general rather than ISTOF in particular. That said, the study of course has clear limitations, not least the scope. Three-fourth grade mathematics lessons form an insufficient sample to fully evaluate the instrument, especially when it is supposed to be generic and used across subjects. The extent to which this is possible, or even whether it would be equally valid in other areas of mathematics or other year groups must remain tentative. Nevertheless, the findings are promising with respect to the utility of ISTOF in terms of giving a broad view of the conduct of lessons which may inform the professional development of teachers, and therefore ultimately the development of mathematics learners.

References

- Antoniou, P., Kyriakides, L., & Creemers, B. P. M. (2015). The Dynamic Integrated Approach to teacher professional development: Rationale and main characteristics. *Teacher Development*, 19(4), 535–552. <https://doi.org/10.1080/13664530.2015.1079550>.
- Black, P., & William, D. (1998). Assessment and classroom learning. *Assessment in Education*, 5(1), 7–74. <https://doi.org/10.1080/0969595980050102>.
- Brophy, J. (1986). Teacher influences on student achievement. *American Psychologist*, 41(10), 1069–1077. <https://doi.org/10.1037/0003-066X.41.10.1069>.
- Chapman, C., Muijs, D., Reynolds, D., Sammons, P., & Teddlie, C. (2015). *Routledge international handbook of educational effectiveness and improvement research*. London: Taylor and Francis.
- Coe, R., Aloisi, C., Higgins, S., & Elliott Major, L. (2014). *What makes great teaching? A review of the underpinning research*. London: The Sutton Trust.
- Devine, D., Fahie, D., & McGillicuddy, D. (2013). What is ‘good’ teaching? Teacher beliefs and practices about their teaching. *Irish Educational Studies*, 32(1), 83–108. <https://doi.org/10.1080/03323315.2013.773228>.
- Devine, D., Fahie, E., McGillicuddy, D., MacRuairc, G. & Harford, J. (2010). *Report on the use of the ISTOF (International System of Teacher Observation and Feedback) protocol in Irish schools; Challenges, issues and teacher effect*. Dublin: School of Education, University College Dublin.
- Dignath, C., & Buttner, G. (2008). Components of fostering selfregulated learning among students. A meta-analysis on intervention studies at primary and secondary school level. *Metacognition and Learning*, 3(2), 231–264. <https://doi.org/10.1007/s11409-008-9029-x>.
- Flanders, N. (1970). *Analyzing teacher behavior*. Reading, MA: Addison Wesley.
- Glaser, B. G., & Strauss, A. L. (1967). *The discovery of grounded theory: strategies for qualitative research*. New York: Aldine DeGruyter.
- Gordon, T. J., & Helmer, O. (1964). *Report on a long range forecasting study*. Rand Paper P-2982. Santa Monica, CA: Rand Corporation.
- Hall, J., Lindorff, A., & Sammons, P. (2016). *Evaluation of the impact and implementation of Inspire Maths in Year 1 classrooms in England; Findings from a mixed-method randomised control trial*. Oxford: University of Oxford.
- Halpin, P., & Kieffer, M. (2015). Describing profiles of instructional practice: a new approach to analyzing classroom observation data. *Educational Researcher*, 44(5), 263–277. <https://doi.org/10.3102/0013189X15590804>.
- Heylighen, F. (2003). *Web dictionary of cybernetics and systems*. Retrieved May 5, 2017 from <http://pespmc1.vub.ac.be/ASC>.
- Kington, A., Day, C., Sammons, P., Regan, E., & Brown, E. (2009). *Effective classroom practise: A mixed-method study of influences and outcomes*. Symposium paper presented at the British Educational Research Association Annual Conference, University of Manchester, 2–5 September 2009.
- Klieme, E., Pauli, C., & Reusser, K. (2009). The Pythagoras Study: Investigating effects of teaching and learning in Swiss and German

mathematics classrooms. In T. Janik & T. Seider (Eds.), *The power of video studies in investigating teaching and learning in the classroom* (pp. 137–160). Munster: Waxmann.

Ko, J. (2010). *Consistency and variation in classroom practice: A mixed-method investigation based on case studies of four EFL teachers of a disadvantaged secondary school in Hong Kong*. Doctoral thesis. Nottingham, UK: University of Nottingham. Retrieved from: http://ethes.es.nottingham.ac.uk/1363/1/CVCP_SUBMISSION_FINAL_PB3.pdf [Accessed May 2017].

Ko, J., & Sammons, P. (2008). *Variations in effective classroom practices: Confirmatory factor analysis results from analysis of measures from the International System for Teacher Observation and Feedback (ISTOF) Scale and Quality and Teaching Lesson Observation Indicator (GRIFT) Scale*. Nottingham: University of Nottingham, School of Education.

Kyriakides, L. (2008). *Results from the ISTOF instrument*. Paper presented at the Annual Meeting of the American Educational Research Association, New York, 17 April 2008.

Marciniak, J., & Janssen, R. (2012). *The International System for Teacher Observation and Feedback questionnaire in the biology assessment—theory, evaluation, utility*. Paper presented at Biennial Meeting of the Special Interest Group Educational Effectiveness (SIG 18) of the European Association for Research on Learning and Instruction (EARLI), Zurich, Switzerland 29–31 August 2012.

Meyer, J. P., Cash, A. H., & Mashburn, A. (2011). Occasions and the reliability of classroom observations: alternative conceptualizations and methods of analysis. *Educational Assessment*, 16(4), 227–243. <https://doi.org/10.1080/10627197.2011.638884>.

Miao, Z., Reynolds, D., Harris, A., & Jones, M. (2015). Comparing performance: A cross-national investigation into the teaching of mathematics in primary classrooms in England and China. *Asia Pacific Journal of Education*, 35(3), 392–403. <https://doi.org/10.1080/02188791.2015.1056593>.

Mortimore, P., Sammons, P., Stoll, L., Lewis, D., & Russell, E. (1988). *School matters: The junior years*. Wells: Open Books.

Muijs, D., Chapman, C., Collins, A., & Armstrong, P. (2010). *Maximum impact evaluation. The impact of Teach First teachers in schools*. Manchester: University of Manchester, School of Education.

Muijs, D., Kyriakides, L., van der Werf, G., Creemers, B., Timperley, H., & Earl, L. (2014). State of the art—teacher effectiveness and professional learning. *School Effectiveness and School Improvement*, 25(2), 231–256. <https://doi.org/10.1080/09243453.20192011.4.885451>.

Muijs, D., & Reynolds, D. (2000). School effectiveness and teacher effectiveness: some preliminary findings from the evaluation of the Mathematics Enhancement Programme. *School Effectiveness and School Improvement*, 11(3), 247–263. [https://doi.org/10.1076/0924-3453\(200009\)11:3;1-G;FT273](https://doi.org/10.1076/0924-3453(200009)11:3;1-G;FT273).

Reynolds, D. (2006). World class schools: Some methodological and substantive findings and implications of the International School Effectiveness Research Project (ISERP). *Educational Research and Evaluation*, 12(6), 535–560. <https://doi.org/10.1080/13803610600874026>.

Reynolds, D., Creemers, B., Stringfield, S., Teddlie, C., & Schaffer, G. (Eds.). (2002). *World class schools: International perspectives on school effectiveness*. London: Routledge Falmer.

Reynolds, D., Salom, K., Delaiglesia, B., & Ramon, R. Mallorca (2012). *The ISTOF project-A preliminary report*. Spain: IAQSE (Institut d' Avaluació del Sistema Education), Ministry of Education of the Balearic Isles.

Sammons, P., Day, C., Kington, A., Gu, Q., Stobart, G., & Smees, R. (2007). Exploring variations in teachers' work, lives and their effects on pupils: Key findings and implications from a longitudinal mixed methods study. *British Educational Research Journal*, 33(5), 681–701. <https://doi.org/10.1080/01411920701582264>.

Sammons, P., Kington, A., Lindorff-Vijayendran, A., & Ortega, L. (2014). *Inspiring teaching: What we can learn from exemplary practitioners*. Reading: CfBT.

Sammons, P., Lindorff, A., Ortega, L., & Kington, A. (2016). *Inspiring*

teaching: Learning from exemplary practitioners. *Journal of Professional Capital and Community*, 1(2), 124–144. <https://doi.org/10.1108/JPCC-09-2015-0005>.

Soderlund, G., Sorlie, K., & Syse, I. (2015). *Mestringsforventninger i matematikk*. Paper presented at Finnut, Lilyhammer, 29-01-2015.

Sterbinsky, A., & Ross, S. (2003). *School observation measure reliability study*. Washington, DC: Centre for Research in Education Policy.

Teddlie, C., Creemers, B., Kyriakides, L., Muijs, D., & Yu, F. (2006). The International System for Teacher Observation and Feedback: Evolution of an international study of teacher effectiveness constructs. *Educational Research and Evaluation*, 12(6), 561–582. <https://doi.org/10.1080/13803610600874067>.

Teddlie, C., Kirby, P. C., & Stringfield, S. (1989). Effective versus ineffective schools: Observable differences in the classroom. *American Journal of Education*, 97(3), 221–236.

Teddlie, C., Kyriakides, L., & Yu, F. (2004). *A proposal to develop an internationally valid teacher observation system: The International System for Teacher Observation and Feedback (ISTOF)*. Paper presented at the Annual Meeting of the International Congress for School Effectiveness and Improvement, Rotterdam, The Netherlands, 5 January 2004.

Tsai, C. C. (2001). Relationships between student scientific epistemological beliefs and perceptions of constructivist learning environments. *Educational Research*, 42(2), 193–205. <https://doi.org/10.1080/001318800363836>.

van de Grift, W. (2007). Quality of teaching in four European countries: A review of the literature and application of an assessment instrument. *Educational Research*, 49(2), 127–152. <https://doi.org/10.1080/00131880701369651>.

van der Lans, R., van de Grift, W., van Veen, K., & Fokkens-Bruinsma, M. (2016). Once is not enough: Establishing reliability criteria for feedback and evaluation decisions based on classroom observations. *Studies in Educational Evaluation*, 50(1), 88–95. <https://doi.org/10.1016/j.stueduc.2016.08.001>.