ZDM (2018) 50:521–534
https://doi.org/10.1007/s11858-018-0915-7

ORIGINAL ARTICLE

Going beyond structured observations: looking at classroom practice through a mixed method lens

Ariel Lindorff[1] · **Pam Sammons**[1]

Accepted: 13 January 2018 / Published online: 23 January 2018

Abstract
In this paper, we extend a mixed method (MM) approach to lesson observation and analysis used in previous research in England, combining multiple structured observation instruments and qualitative field notes, to provide a framework for studying three videotaped lessons from 3rd-grade US mathematics classrooms. Two structured observation schedules are employed, one subject-specific and research-oriented and the other generic and inspection-oriented. Both instruments were previously developed based on evidence from the teacher effectiveness research (TER) knowledge base. Qualitative field notes, in addition to structured observation schedules, provide detailed narratives for each lesson video. Separate findings from each instrument and approach are presented, followed by an integrated analysis and synthesis of results. Although previous studies used similar methods to analyze teaching practice within broader research designs incorporating additional methods and perspectives (e.g. teacher interviews, pupil assessments, pupil questionnaires), this paper explicitly examines the strengths and limitations of the multi-instrument, mixed method approach to lesson observation. Using multiple observation instruments allows for triangulation as well as consideration of complementary foci (i.e. a content-specific instrument measures fine-grained aspects of practice not emphasized in a more generic instrument, and vice versa). Field notes facilitate rich descriptions and more thorough contextualization and illumination of teaching practice than structured observation ratings alone. Further, the MM approach allows for consideration of lesson features beyond those established in TER literature as sufficient to characterize 'effective' practice.

Keywords Classroom practice · Mixed methods · Structured observation · Video lesson · Mathematics instruction

1 Introduction

The use of standardized observation instruments is common practice in studies of teacher effectiveness and classroom practice. While such instruments are useful in large-scale studies, their utility for in-depth studies of smaller samples or for exploratory purposes is more limited. Any individual instrument is inherently bounded by the context(s) in which it was developed, the theoretical framework underpinning it, and its intended purpose (e.g. to study relationships between teaching practice and pupil outcomes; to compare teaching practice in different countries; to evaluate practice in a specific subject area; to inform professional development). Previous studies have used multiple observation instruments, as well as qualitative field notes, to mitigate the limitations of any single instrument while providing thorough descriptions of teaching practice (e.g. Hall et al. 2016; Kington et al. 2014; Sammons et al. 2014). The emphasis in these previous studies, however, was not on investigating strengths and weaknesses of the overall mixed methods (MM) approach to lesson observation, nor of its components. The novel contribution of this paper, therefore, is to illustrate and discuss the benefits and challenges of using a MM approach to lesson observation and analysis, as well as the strengths and weaknesses of each aspect of this approach (two quantitative observation schedules, and qualitative field notes), based on our analysis of three videotaped lessons in 3rd-grade US mathematics classrooms.

Electronic supplementary material The online version of this article (https://doi.org/10.1007/s11858-018-0915-7) contains supplementary material, which is available to authorized users.

✉ Ariel Lindorff
ariel.lindorff@education.ox.ac.uk

[1] Department of Education, University of Oxford, 15 Norham Gardens, Oxfordshire, Oxford OX2 6PY, UK

Below, the theoretical perspective framing our approach to observation and analysis is introduced, followed by an overview of the instruments and approaches employed and empirical support for the use of these to observe and evaluate lessons. Next we present quantitative, qualitative and integrated findings based on three videotaped lessons, with particular emphasis on how qualitative findings elaborate, extend and diverge from quantitative findings. Finally, we discuss strengths and weaknesses of the structured observation schedules and field notes, as well as of the combined MM approach as a whole, and draw implications for research, professional development, and teacher evaluation. We focus our analysis on features and practices in the three focal lessons rather than global judgments of teachers' effectiveness, as one lesson is not necessarily a typical or full representation of a teacher's classroom practice.

Overall, the purpose of this paper is illustrate this multi-instrument mixed methods approach to lesson observation and analysis, and to address the following questions:

1. What does each aspect of the approach tell us about the focal lessons, and what do we learn from integrating and synthesising findings across all aspects?
2. What are the strengths and weaknesses of each aspect of the approach, and of the approach as a whole?

2 Theoretical perspective

The overarching theoretical perspective framing our MM approach to lesson observation and analysis is grounded in the teacher effectiveness research (TER) knowledge base. More specifically, we place an emphasis on those practices which empirical evidence from previous research has shown to promote pupil attainment (see Teddlie et al. 2006; Muijs et al. 2014). Theoretical frameworks for studying teacher effectiveness can be traced back to early studies that took "student learning in classrooms as schools as a point of departure" (Creemers, 1994, p24), developing models including instructional factors such as Carroll's (1963) ratio of time spent to time needed for learning. Models for effective instruction have been further refined since via empirical studies to establish features of teaching that predict improved students' outcomes (academic and socio-emotional), and variously involve aspects including lesson structure, content delivery, behaviour management, interaction, focus, questioning, pupil involvement, and emotive/cognitive feedback (Ko, Sammons, & Bakkum, 2013).

Previous research has shown, however, that an effectiveness perspective may not fully characterise teacher practice. An exploratory MM study of "inspirational" teaching in England found that what constituted "inspirational" practice—across several themes emerging across the teacher perspectives, pupil responses, and lesson observations, relevant core features including: positive relationships, good classroom management, positive and supportive classroom climate, formative feedback, enjoyment, and a high quality learning experience overall—overlapped with but also extended understanding beyond those constituting effective practice (Sammons et al., 2014, 2016). For example, although secondary teachers in the sample were rarely observed using formal methods of differentiation (by providing different tasks or using distinctly different teaching approaches for different pupils), they met individual needs through informal approaches to personalizing learning experiences and by capitalizing on strong rapport and personal relationships with individual pupils. A similar approach to observation used in another study evaluating a mathematics textbook and mastery-oriented teaching approach in England also demonstrated that qualitative evidence illuminated features of teaching practice not covered in systematic observation schedules based only on TER literature (Hall et al. 2016). For example, findings illustrated how teachers were consistently using mixed-ability grouping, a key feature of the specific teaching approach being evaluated, but were adopting a variety of approaches to doing so according to their perceptions of the needs of their pupils. This ability grouping was not an aspect of classroom practice covered in the pre-existing quantitative observation schedules used, so that qualitative data played an important role in informing understandings of classroom practice in the context of a mastery-oriented teaching approach.

In keeping with a theoretical perspective framed by TER, the approach to lesson observation and analysis presented below employs two existing systematic observation schedules both explicitly designed based on evidence from the TER literature. The mathematics enhancement classroom observation recording system (MECORS; Schaffer et al. 1998), is specific to mathematics and was originally used to evaluate a mathematics intervention programme in the UK. The other is the quality of teaching (QoT) lesson observation form (van de Grift et al. 2004; 2007), also based on a review of TER and designed for use in primary schools with an orientation towards evaluation of quality based on school inspection frameworks from the Netherlands and the UK (but tested in several European countries). We combine these instruments with qualitative field notes for two reasons: first, to provide richer and deeper descriptions of practice, and second, to help contextualise findings because lessons were conducted in US classrooms while observation schedules were designed in non-US settings. The use of multiple quantitative instruments that were designed for different purposes avoids an overly narrow characterisation of teaching practice driven by the nature of a particular scale (e.g. based on frequency or high-judgment quality ratings). The use of quantitative instruments alongside qualitative

field notes helps to ensure that ratings of teaching practice are contextualised and underscored by vignettes, quotes and examples, and that the approach to observation and analysis is flexible enough to account for aspects of practice not emphasized in the quantitative instruments.

3 Framework, instruments and approaches

Below, we provide details of the measures and approaches for each of the three aspects of our design. While this paper is intended to illustrate the combined mixed methods approach and is based on only three lessons, it is important to note that implementation in a larger study would include multiple raters to allow for calculations of inter-rater reliability and to avoid possibilities of bias arising from individual raters' use of multiple instruments. It should be noted that both the systematic instruments have been used in previous studies and have published details of inter-rater reliability.

3.1 The mathematics enhancement classroom observation recording system (MECORS)

As described by the authors of the instrument, the MECORS was used in two stages. First, the observer recorded:

- type of activity (whole-class interactive teaching, lecture, group/pair work, individual practice, assessment, or management of resources/materials/physical space), coded each time this changed;
- detailed notes on the activity,
- number of pupils on-/off-task at 5-min intervals (time-sampling).

Because this instrument was originally designed for use by an observer physically present in the classroom, we modified the approach for video lesson observation. In particular, field notes and time-sampling were done in multiple passes (once per camera angle) to ensure the fullest possible information and alignment by time code across the notes from different camera angles in the same lesson.

The observer then filled out a rating sheet comprised of eight dimensions and a total of 57 items, each rated on a Likert scale from 1="behavior rarely observed" to 5="behavior consistently observed". Table 1 shows the eight domains in the order in which they appear on the observation schedule, and the number of items contributing to each.

Most of the categories above are relatively self-explanatory; "Classroom management" might be less so. Here, this refers to management of physical space and resources, distinct from behavior management. "Mathematics Enhancement Project (MEP) strategies" is a phrase from the original study for which the instrument was designed, and refers to teaching features relevant to: connecting new content to prior content, other areas of mathematics, and real-world context; use/promotion of correct mathematical language; teaching/encouragement of (a variety of) problem-solving strategies; and rapid-fire mental questioning. This category is the most specific to mathematics in the MECORS instrument (Schaffer et al. 1998).

Table 1 MECORS domain descriptions and item counts

Domain description	Number of items
Uses classroom management techniques	5
Maintains appropriate classroom behavior	5
Focuses and maintains attention on lesson	8
Provides pupils with review and practice	6
Demonstrates skills in questioning	14
Demonstrates Mathematics Enhancement Project (MEP) strategies	8
Demonstrates a variety of teaching methods	3
Establishes a positive classroom climate	8

MECORS items are framed in terms of teacher behaviors or practices, and rated on a frequency scale rather than judgments of the quality of a given practice. Items span a range of aspects of teachers' practice and provide a tool to identify general areas of strength/weakness (Muijs & Reynolds 2011). Appendix 1 contains the full list of items included in the MECORS observation schedule, with appropriate reference to the original authors of the instrument.

The rationale for using the MECORS to illustrate our approach to lesson observation and analysis is threefold. First, the inclusion of eight mathematics-specific items (such as 'The teacher uses correct mathematical language' and 'The teacher allows pupils to use their own problem-solving strategies' under the dimension of 'Demonstrates MEP strategies') and items related to established features of effective teaching specifically in mathematics (such as 'The teacher asks pupils to explain how they reached their solution' and 'Pupils are asked for more than one solution' under the dimension of 'Demonstrated skills in questioning') made the MECORS well-suited for the observation of the three mathematics lesson videos on which this paper focuses. Second, the research team was familiar with the instrument and experienced in its use from a previous research project in England. Third, the MECORS instrument also covered more general features of effective practice allowing for triangulation and comparison with the QoT instrument described below.

3.2 The quality of teaching (QoT) lesson observation form

The QoT instrument included nine dimensions of classroom practice, with each dimension measured with 2–4 indicators, and each indicator rated from 1="predominantly weak" to 4="predominantly strong" and supported by 1–5 "Good practice examples" rated either 0 ("no, I didn't observe this") or 1 ("yes, I have observed this"). We included an additional option of "not applicable" for these, as we found this to be appropriate in some instances. Table 2 shows the nine dimensions of the QoT and the number of items contributing to each.

A few of the below descriptors merit clarification. "Stimulating learning climate" with items such as, "The teacher stimulates the independence of pupils", relates to the teacher's facilitation of cohesion, cooperation, independence, and individual involvement on the part of pupils, whereas "Safe and orderly climate" items (such as, "The teacher promotes mutual respect") focus on fostering a relaxed atmosphere, mutual respect between pupils, pupils' self-confidence, and demonstration of respect for pupils in the teacher's language and behavior. "Effective classroom organisation" includes items relating to lesson structure, organisation, and orderly progression, as well as effective management of time and lesson materials and resources, rather than physical space. "Effective classroom layout," on the other hand, refers to physical space/décor. The "final judgement" is an overall rating of lesson quality, so is not fundamentally separate from ratings across the other categories. Appendix 2 contains the full list of items included in the QoT observation schedule, with appropriate reference to the original authors of the instrument.

In contrast to the MECORS, items on the QoT are rated first in terms of whether "good practice" examples (e.g. "The teacher allows pupils to finish speaking" for the item, "The teacher shows respect for pupils in behavior and language use" under the dimension, "Safe and orderly climate") are observed or not, then items in terms of the observers' perception of the extent to which a teacher displays strength or weakness for a particular indicator. While both the MECORS and QoT involve some degree of observer inference, the QoT levels of measurement may be seen as involving more subjective (high inference) judgement due to a focus on quality rather than frequency. Like the MECORS, QoT covers a broad range of classroom practice features and is intended to identify general areas of strength or weakness.

The rationale for using the QoT instrument to illustrate our approach was that it had been internationally validated and used in previous published research (more detail on this is given below in Sect. 4.2), and was developed from a combination of inspection- and research-based perspectives and evidence. Additionally, our research team was trained on this instrument and experienced in its implementation, having used it for multiple prior studies in England.

Using the MECORS and QoT in combination for observation of the same lessons affords opportunities for triangulation where overlap exists between the two instruments, and complementarity where items on one schedule cover aspects absent from the other. As noted elsewhere, we use these instruments to illustrate our approach here with ratings from a single researcher for three lesson videos; in a larger study we would either randomly assign different instruments to multiple raters for the same lessons, or randomly assign the order in which different instruments were used by raters if using only one rater per lesson.

Table 2 QoT dimension descriptions and item counts

Dimension description	Number of items
Safe and orderly climate	4
Stimulating learning climate	4
Clear objectives	2
Clear instruction	3
Activating pupils	2
Adaptation of teaching	2
Teaching learning strategies	3
Effective classroom organisation	4
Effective classroom layout	2
Final judgement	1

3.3 Qualitative field notes

Qualitative field notes are used to provide rich descriptions of classroom activities and tasks, teacher and pupil behaviors, classroom interactions, and classroom environment (including decorations and physical space) in these example lessons. Using a loose framework informed by previous research (e.g. Kington et al., 2014), notes focused on: Descriptions of tasks and activities, what pupils were doing, and what teachers were doing, with quotes recorded to illustrate specific interactions and teacher feedback, and descriptions of the physical environment (this last would normally be addressed before the lesson started, for a "live" rather than videotaped lesson). As such, these field notes were not strictly structured, but included time codes as a reference point to allow for alignment with video transcripts along with descriptions covering the aspects mentioned above as the observed lesson progressed. Researcher memos were also recorded alongside descriptions of what was happening in the classroom; these included any questions or comments that arose in relation to what was directly observed, for example, classroom routines not made explicit but apparent in teacher and student interactions or behaviors. A major

benefit of field notes as part of an overall MM approach is the potential to provide a greater degree of *specificity regarding particular features* of a teacher's behaviors and practices during an observed lesson, in contrast to the more general categories of behavior and practice offered by the structured observation schedules.

The field notes provide the opportunity both to give rich vignettes illustrating specific examples of classroom practice to support ratings on structured observation schedules, and to generate further understandings of specific teaching contexts and initiatives through the use of an inductive approach to analysis. The latter is characteristic of a grounded-theory approach to analysis, and allows themes to emerge from the qualitative data from a particular study. The qualitative analytical approach is described in more detail below in Sect. 5.

4 Empirical support for the framework and approach

There is precedent in previous empirical research literature for the use of each observation instrument described above, as well as for their integration with qualitative field notes in a mixed methods approach to lesson observation and analysis.

4.1 Empirical support for the MECORS instrument

The MECORS instrument was developed to evaluate a mathematics intervention in primary schools in England. It drew on instruments designed previously in the US and shown to be reliable, including the Special Strategies Observation System (SSOS; see Schaffer et al. 1994) and the Virgilio Teacher Observation Instrument (Teddlie et al. 1990). Findings from the study in England using the MECORS, including observations of 78 teachers, found that direct instruction was strongly and positively related to effective teaching scales. Internal consistency was above Cronbach's Alpha $\alpha = 0.8$ for all scales in the schedule (Muijs & Reynolds 2000).

The MECORS has also been used outside England. A study in Malta used the instrument but omitted some items in consultation with local teachers (Said 2013), suggesting that a larger-scale study of US lessons than this paper allows might adopt such consultation or consideration of whether items showing little variation should be candidates for exclusion. This study showed high reliability in terms of overall inter-rater agreement ($k = 0.89$, $p < 0.01$) based on 25 mathematics lessons rated by two observers, but some items had substantially lower inter-rater agreement (particularly, "The teacher uses a brisk pace", with $k = 0.67$, $p < 0.01$). Implications for scale validity and reliability would need to be tested, however, for a larger study than the present illustrative analysis of three lesson observations, if using the MECORS in a new setting.

4.2 Empirical support for the QoT instrument

The QoT instrument was piloted in England, Belgium, Germany, and the Netherlands; except in England, participating schools constituted within-country random samples, with 854 primary mathematics lessons observed in total (van de Grift 2007). Findings from the pilot demonstrated scales were internally consistent (based on Cronbach's Alpha) with values above $\alpha = 0.7$. For dual observations (two observers from two countries), overall inter-rater reliability was high (consensus above 83%) (Ibid.). Sufficient construct validity for international comparisons was assessed both by computing correlations to related concepts on a separate instrument (about 0.70 across "teaching" and "learning" categories on the alternative instrument) and correlations between QoT constructs and the overall judgment of teaching quality (between 0.59 and 0.72; Ibid.). A subsequent study investigated QoT measurement invariance across Flanders (Belgium), Lower Saxony (Germany), the Slovak Republic and The Netherlands, and the relationship of cultural differences to measurement differences; findings supported the reliability of QoT measures but indicated that some (particularly classroom management and adaptation of teaching) were differentially sensitive to student background and school/classroom characteristics across different countries (van de Grift 2014). In particular, measures related to clear instruction and activating pupils were reliable and fully scalar equivalent across the four countries, the measure of safe and stimulating climate was only partially scalar equivalent, and measures of classroom management and adaptation of teaching were only metrically equivalent and interacted differently across countries with variables related to student background and school and class features, suggesting caution in making multi-country comparisons using this instrument with regard to these latter measures (Ibid.).

4.3 Empirical support for the overall MM approach

In a MM framework, concepts of validity and reliability are extended to both individual methods and the combination of methods (Teddlie & Sammons 2010; Sammons & Davis 2016). For the quantitative instruments, these are addressed in the sections above. In qualitative inquiry, we consider trustworthiness and dependability as analogs of the more quantitatively-oriented terms "validity" and "reliability," respectively (Lincoln & Guba 1985). The use of both the qualitative approach to taking and analysing field notes, and the mixed methods approach overall, has precedent in previous studies (Hall et al. 2016; Kington et al. 2014; Sammons et al. 2014); participant practitioners' positive responses to

both the approaches and findings of those studies supports the trustworthiness of the overall approach, and the engagement of multiple researchers in the process of observation and analysis in at least one study (Sammons et al. 2014) has supported the dependability of the MM approach and findings arising out of it.

5 Methods and analytical approach

The approach to analyzing quantitative data were limited by the very small number of teachers, focus on only one videotaped lesson per teacher rated by a single researcher with training and experience using both observation schedules and qualitative field notes. This is sufficient for the purposes of this paper to illustrate the multiple-instrument, mixed methods approach. In a larger study, it would be ideal to have multiple researchers trained on different instruments to avoid potential bias; alternatively, this could be accomplished by having multiple researchers observe each lesson but randomly altering the order in which they complete their observation ratings.

In order to obtain results that could be discussed/compared across teachers and instruments, we calculated each teacher's mean scores for the items within each component of each instrument. We report means rather than sums because on both the QoT and the MECORS, the number of items per component varies considerably. Thus, a mean score is more straightforward to interpret in terms of meanings of original item scales.

Qualitative analysis was undertaken in several stages. First, field notes were coded using a grounded approach to allow themes to emerge from what was observed/recorded (Glaser 1992). This was followed by more fine-grained coding of data coded according to broader categories (e.g. "feedback" or "assessment") to more fully and specifically characterize patterns in the data (e.g. variations of practice within a particular category, or similarities across lessons) emulating an approach to analysis used in previous studies (Hall et al. 2016; Sammons et al. 2014).

6 Analysis of the three focal lessons

We begin with a brief overview of results from each structured observation schedule, followed by a summary of themes emerging from the thematic coding of qualitative field notes and how these are illustrated in each observed lesson.

6.1 Structured observation findings: mathematics enhancement classroom observation record

Table 3 shows the mean scores for the three teachers on each of the components of the MECORS instrument.

In Mr. Smith's lesson the highest-rated aspect was "focusing and maintaining attention on lesson". There was little off-task behavior at any of the time-sampling occasions, although exceptions to this are discussed below from qualitative findings. Items contributing to "providing pupils with review and practice" and "demonstrating skills in questioning" had average ratings of "often observed" (3.0). However, within these categories the ratings had a wide range across different items; for example, within the 14 items related to questioning skills, "pupils are asked for more than one solution" was rated 1 ("rarely observed"), while items specific to "high frequency of questions" and "academic questions" were rated 5 ("consistently observed"). Similarly, within the 6 items related to review and practice, the item specific to "clearly explains tasks" was rated 4 ("often observed") while another item specific to "offers effective assistance to individuals/groups" was rated 1 ("rarely observed").

In Ms. Young's lesson, the mean scores for components relevant to focusing and maintaining attention on the lesson, providing pupils with review and practice, and demonstrating questioning skills were all between "often" and "frequently" observed (i.e. slightly above the middle of the scale); however, like Mr. Smith's lesson, the ratings for individual items varied considerably. The lowest ratings for this lesson related to classroom and behavior management.

Table 3 MECORS mean scores for each teacher on each component

MECORS components	Mean		
	Mr. Smith	Ms. Young	Ms. Jones
Uses classroom management techniques	2.0	1.8	4.0
Maintains appropriate classroom behavior	1.5	2.0	3.8
Focuses and maintains attention on lesson	3.8	3.4	4.1
Provides pupils with review and practice	3.0	3.4	4.2
Demonstrates skills in questioning	3.0	3.2	3.7
Demonstrates MEP strategies	2.6	2.6	3.0
Demonstrates a variety of teaching methods	2.3	3.0	3.0
Establishes a positive classroom climate	2.1	2.5	3.6

The mean scores for Ms. Jones's lesson, compared to the two above, showed less variation across categories, with all mean scores above 3 ("often observed"). The highest ratings of this lesson corresponded to classroom management, focusing and maintaining attention on the lesson, review and practice, and behavior management. The lowest scores were for using MEP strategies and demonstrating a variety of teaching methods, but these were still "often observed".

Based on MECORS scales, the three lessons were most similar in terms of teachers' "use of MEP strategies", and least similar with regard to classroom and behavior management. Mr. Smith's and Ms. Young's lessons were most similar to one another in their mean scores across categories.

6.2 Structured observation results: quality of teaching

Table 4 presents the mean scores for the three teachers on each dimension of the QoT instrument.

Mr. Smith's lesson was rated roughly in the middle of the QoT scales according to mean scores for safe and orderly climate, clear objectives, clear instruction, and effective classroom layout. Adaptation of teaching and teaching learning strategies were the lowest-rated categories, as shown in Table 4; there was no apparent adaptation of materials, activities, or teaching approach to individual needs, little to no interaction between pupils, only one apparent solution or strategy for any of the problems, and little context given for problems and solutions. All other categories had mean scores reflecting "more weaknesses than strengths" (i.e. two on the QoT scale). According to the QoT's underlying framework for assessing quality of instruction, this lesson had more observed weaknesses than strengths overall, and two general areas of particular weakness that would potentially warrant further consideration (for research or professional development purposes).

Table 4 QoT mean scores for each teacher on each dimension

QoT dimensions	Mean		
	Mr. Smith	Ms. Young	Ms. Jones
Safe and orderly climate	2.5	1.8	3.5
Stimulating learning climate	2.0	2.8	3.3
Clear objectives	2.3	2.5	3.0
Clear instruction	2.7	3.7	3.7
Activating pupils	2.0	2.0	2.0
Adaptation of teaching	1.0	1.5	1.5
Teaching learning strategies	1.7	2.3	3.0
Effective classroom organisation	2.0	2.8	3.3
Effective classroom layout	2.5	2.5	3.5
Final judgement	2.0	2.0	3.0

From the QoT ratings, Ms. Young's lesson had wider-ranging mean scores across categories. "Clear instruction" was the highest-rated (close to 4, "predominantly strong"), while "adaptation of teaching" and "safe and orderly climate" were the lowest-rated categories. All of the other categories ranged from 2.0 to 2.8 (somewhere between 2, "more weaknesses than strengths," and 3, "more strengths than weaknesses"). Thus, the QoT clearly suggests a balance of strengths and weakness in this lesson.

The mean scores across categories for Ms. Jones's lesson were almost all high (between 3, "more strengths than weaknesses," and 4, "predominantly strong"). The only exceptions were "activating pupils" (rated 2, "more weaknesses than strengths") and "adaptation of teaching" (rated 1.5, between "predominantly weak" and "more weaknesses than strengths"). The teaching observed in this lesson was predominantly strong except for these two specific areas, so that a further focus on adapting instruction and assignments to individual learning needs and using teaching and questioning strategies that involve all pupils throughout a lesson might be areas for professional development or further inquiry.

According to the QoT scales, these three lessons were most similar across the (overall, generally weaker) areas of adapting for individual pupil differences and learning needs, and using methods/approaches that "activate" pupils. Ms. Jones's lesson was rated highest across the remaining categories, except for clear instruction, in which Ms. Young's lesson was rated equally to Ms. Jones's. For the most part, Ms. Jones's lesson had higher ratings across the other categories than Mr. Smith's, with the exception of "effective classroom layout" (for which the two lessons were rated equally) and "safe and orderly climate", for which Mr. Smith's lesson had a higher mean score.

6.3 Qualitative observation results: themes

Several themes emerged from the grounded initial approach to coding qualitative field notes. These fell into five broad categories: lesson structure and activities, teacher interaction and feedback, behavior management, pupil involvement and participation, and assessment.

6.3.1 Lesson structure and activities

Within this broad category, three thematic areas emerged from the coding for the three lessons. These included, in order of their frequency in the field notes: format of lesson activities (e.g. direct instruction, individual table work), timing/transitions, variety of activities, and differentiation/adaptation for individual needs. These are addressed in this section with respect to each lesson.

In Mr. Smith's lesson, almost the entire class period was spent on direct instruction. There were clear introductory and closure parts of the lesson; the teacher asked about prior knowledge during the introduction before proceeding with the lesson, and for closure pupils were given a problem set answer individually on paper. Transitions were quick and relatively smooth, perhaps in part because pupils remained in their seats throughout and few materials were distributed except protractors and printed closure exercises. Activities largely involved the teacher projecting angle images on the smart board and asking questions to the whole class. There was no evidence of differentiation or adaptation.Pupils were questioned collectively, an aspect of practice which would not have been obvious from the quantitative ratings alone.

In Ms. Young's lesson, the introduction involved the teacher asking about the previous mathematics lesson content, presenting pupils with two related multiplication problems, and using these to explicitly present the objectives of the lesson (i.e. to use doubling and halving strategies to solve problems). Transitions were not well organised; it took several attempts and approximately 4 min for children to assemble on the carpet when called, and the first few announcements of this transition (including ringing a bell) got limited pupil response. The lesson task involved one focal problem, but the teaching approach varied (direct instruction at the beginning and end and individual work with teacher circulation in between). Much of the lesson time involved pupils working at tables to justify the equivalence of two multiplication problems; this use of class time on a single problem is an aspect that emerged only from the qualitative analysis, but may well be an important consideration in analysing a mathematics lesson. Even during this time the teacher employed some direct instruction, repeating questions and key points to the whole class. Although there was little evidence of adaptation for individual needs, the teacher's questioning discussion with individuals and groups while circulating may have served as an informal approach to adapting instruction.

In Ms Jones's lesson, the introduction included explicit framing of the lesson objectives ("multiple ways to multiply a whole number times a fraction") and specific instructions for pupils to set up activity materials (writing the title; organising three sections on construction paper). There was a defined closure, but it was also clear that some planned lesson activities had not been completed (two of the planned three approaches had been addressed, with no time for pupils to record the second). Transitions were smooth, but there was little movement of pupils around the room. There was some variety in lesson activities, with the teacher telling stories to engage pupils with core concepts, asking them to write processes in a structured way, and finally working with manipulatives in small groups to demonstrate one solving strategy. With the exception of a few minutes spent on group work, the teaching approach mainly consisted of direct instruction. There was little formal adaptation for individual learning needs, but the teacher circulated regularly and occasionally spent time scaffolding task instructions for individual pupils, which suggests—as noted in Ms. Young's class—of an informal approach to differentiation/adaptation, an aspect for which having detailed field notes was useful to more fully characterise the teacher's practice.

Thematic patterns are apparent across these lessons. All three teachers emphasised direct instruction and used some formal lesson structure (i.e. introduction and conclusion). Cohesion and whole-class activities were prioritised, and there was little evidence of adaptation of tasks for individual learning needs (although two of the three teachers appeared to use informal approaches to adapt instruction for individuals/groups).

6.3.2 Teacher interaction and feedback

Four apparent aspects of teacher interaction and feedback stood out across the three lessons. These included the extent and nature of teachers' interactions with individual pupils, positive/negative feedback to pupils', and evidence of building/maintaining relationships with pupils.

In Mr. Smith's lesson, most teacher–pupil interaction was between the teacher and the whole class group, with the teacher asking frequent questions and allowing the class to respond simultaneously. Less frequently, the teacher called on volunteers to answer his questions, and when teaching how to properly use protractors, volunteers were invited to demonstrate on the smart board while the class observed. Occasionally, the teacher spoke to pupils individuals beyond the above. On one occasion he responded to a pupil's erroneous response, giving examples of when approximate angle measures were insufficient (e.g. "if you make your building 91° everybody's going to be walking slanted a little bit"). Feedback was also largely directed towards the whole class, an important distinction that would not have been apparent through the use of only the quantitative instruments, and was not expressed in strongly positive or negative terms. When the responses were correct (or mostly correct) he frequently said "Okay" and repeated correct answers, and used relatively neutral wording to correct wrong answers (e.g. "Actually, it's D"). Feedback focused mainly on whether answers were correct or incorrect, but sometimes the teacher strove to explain solution processes (e.g. how a protractor should be used and what mistakes to watch out for; explaining that pupils may have read the wrong line of numbers). There was some evidence of the teacher's ongoing attempts to build and maintain relationships, however, with little individual pupil-teacher interaction, it was difficult to gauge the extent to which Mr. Smith had established relationships with pupils.

Ms. Young interacted frequently with pupils in her classroom, speaking to individuals while circulating, using their names, and spending more time with some pupils to support their work. Ms. Young frequently used both positive and negative language (e.g. "Oh, that's nice!"; "You made the mistake, you fix it") in giving feedback to pupils. There was evidence of the teacher's relationships with individuals, but in the observed lesson this was largely negative, with the teacher making some global comments to individuals about their behavior or participation before this lesson, and commenting sarcastically at the end to two pupils, "Thank you, Student A and Student B, for disrupting the lesson throughout the day…" While quantitative instruments included categories relevant to many of the above behaviors, having specific examples from the field notes illustrates reasons *why* some of the quantitative ratings indicating less effective practice in this lesson.

Two main forms of teacher–pupil interactions were visible in Ms. Jones's lesson. The first involved Ms. Jones's feedback to answers during direct instruction, frequently confirming whether a response was correct and repeating the response to the class. The second involved Ms. Jones having quiet conversations with individuals when circulating, sometimes prompting pupils to reflect if they had not followed or understood instructions (e.g. "Look at your fraction. Do you have a denominator of four?"). The teacher gave succinct and frequent positive feedback (e.g. "Exactly", "Love it!"), and occasional, brief negations to incorrect answers ("No, it can't be zero 'cause it won't work"). There was positive feedback to the whole class at the end of the lesson ("Very good job, guys, today. I'm very proud of you,") and to individual groups at other times ("Thank you, yellow table. And thank you, pink table"). There was evidence that the teacher had previously built relationships with pupils, as reflected in her use of (and their excited engagement with) personal stories about her daughter, and the relaxed affect of both pupils and teacher. For example, the teacher's calling pupils "friends," and her laughter with them when one belched intentionally in response to a story reinforcing the meaning of "times," suggested a relaxed learning environment supported by a positive teacher–pupil dynamic. Here again, the qualitative field notes help to illustrate some of the reasons behind quantitative ratings confirming and extending findings on certain features of effective practice in this lesson.

To summarise, there was considerable variation across lessons in terms of the nature and extent of teachers' feedback to and interactions with pupils, and in the extent to which positive teacher–pupil relationships were evident.

6.3.3 Behavior management

Qualitative thematic coding revealed several aspects relevant to behavior management: frequency and seriousness of disruptions, response and responsiveness of the teacher to pupil behavior, and evidence of established classroom routines and norms.

Mr. Smith's class appeared largely appropriately-behaved and quiet throughout the lesson. There were few disruptions, and those that were evident were fairly minor. One boy appeared to look at something under his table for much of the lesson, made faces at the camera, and engaged a seatmate in off-task conversation. The teacher did not seem to see or respond to this. Classroom routines and norms were not explicit, but it appeared that these had been previously established, given the majority of pupils' ready engagement in whole class question-and-response processes and quiet waiting when the teacher was speaking or a classmate was demonstrating at the board.

In Ms. Young's class, there were numerous times when children were out of their seats and walking around the class, particularly when they were supposed to be working at tables. In a few instances, it was clear that this behavior was task-oriented; several children approached Ms. Young to check their work, and a few others went to get graph paper or other supplies. Some, however, appeared off-task. One boy approached another table to chat, another ran across the back of the room out of the camera's field of view. One of these pupils also appeared to spend the beginning of the lesson reading an unrelated book. The teacher sometimes responded quickly to redirect disruptive or off- task behaviour (e.g. "Sit down. You need to be part of this discussion or I'll put you in that room, too," after sending a child out of the classroom). On the other hand, some behaviors noted above got no apparent teacher response. There was some evidence of established rules and norms (e.g. "You're not supposed to leave your seat during the class discussion"), but this was less apparent with regard to when, how and why pupils could move around the room. Getting the class to quiet down, pay attention and come to the carpet appeared to be a challenge. Overall, behavior management was inconsistent, which may have been partly due to the high level of activity when the lesson required pupils to work at tables and use a variety of materials to support their work.

Pupils in Ms. Jones's lesson were mostly seated quietly when the teacher or another pupil was speaking, and there was little overt disruption. What disruptions did occur were minor and did not disrupt the flow of the lesson or continue for long (for example, one boy was seen trying to touch his seatmate's hair with a pencil, but stopped when his seatmate looked at him and both raised their hands to participate). The teacher overtly redirect pupil behavior often, but divided her time circulating around the room and stood in different locations during direct instruction so that she did appear to be monitoring the whole class. The teacher sometimes referenced classroom norms explicitly, usually framed in positive statements (e.g. "We raise our hands"). Evidence

of established routines and norms was suggested by generally consistent pupil behavior, raising hands to answer questions, and visible shared work habits (e.g. pupils having notes handy at the start of the lesson or turning to a partner to share notes).

Although each lesson had a distinct pattern of observed behavior and behavior management, both pupils' behaviors and teachers' responses to disruption were more consistent (though distinct) in the lessons that involved less pupil movement around the room. This suggests that the nature of lesson activities can affect teachers' behavior management, pupils' behavior, and an observer's ability to see clear patterns in both (within the limits of a single lesson).

6.3.4 Pupil involvement/participation

Notes on pupil involvement and participation fell into two main categories. These pertained to the extent to which pupils showed individual engagement in activities, and the extent to which they participated in discussions.

In Mr. Smith's lesson, most pupils were involved in calling out whole-class responses. One or two did not respond throughout the lesson, but even those who remained silent appeared to look and listen, with the exception of the one child involved in disruptive/off-task behavior described above. Only a few children volunteered to answer/demonstrate individually, however, and no apparent strategy or approach was used by the teacher to encourage broader participation. The only opportunity for individual work occurred at the end of the lesson, and it was impossible to gauge individual involvement as the video ended with instructions for this activity.

In Ms. Young's lesson, there were a few eager volunteers during class discussion. There was no evidence of specific strategies to elicit responses from non-volunteers. However, much time was spent working at tables, and during this time most pupils worked on their own or discussed in their groups. Only a few (seated at the back of the room) were apparently off-task. At the end of the lesson, most pupils listened quietly as classmates presented.

In Ms. Jones's lesson, most pupils in view of the camera raised hands to volunteer answers, so there was a high level of individual involvement in class discussion. The teacher did not use strategies to elicit answers from non-volunteers, however, she did ask for nonverbal signals when pupils completed a task, such as different gestures to indicate readiness to proceed at different times (e.g. making "bull horns" with their fingers, putting their thumbs up on the table, putting their hands on their heads), and children seemed almost universally to engage with these instructions (with some turning to neighbors to remind them of instructions). Every pupil seen on screen followed the instructions for the written and individual components of the lesson (e.g. cutting circles into fourths), and many called the teacher over to confirm that they were on track or debated with peers about how to complete a task, suggesting a high degree of individual involvement.

In short, despite variations in behavior and behavior management, there were common patterns of pupil involvement and participation. The three lessons were characterised by high levels of apparent attention and engagement with tasks, but participation in discussions was mostly limited to pupils who volunteered.

6.3.5 Assessment

Two aspects of assessment were identified in the field notes: informal and formal assessment of progress towards learning objectives. Here, we define formal assessment as a task completed individually by pupils and marked by the teacher to measure pupils' knowledge/skills in relation to learning objectives. We define informal assessment as any process/task/interaction providing the teacher with information about pupils' knowledge and skills.

At the end of Mr. Smith's lesson, pupils were asked to individually written questions; this may have constituted formal assessment, if papers were marked subsequently. There was little informal assessment during the lesson, at least on an individual level, as questions were posed to the whole class and most were answered in chorus. Mr. Smith appeared to get information informally about what the whole class knew, but not about individual learning, based on pupils' responses in chorus.

Ms. Young implemented no formal assessment, but engaged in some informal assessment. While pupils worked at their tables, she asked questions about what they were doing or corrected their work. In some instances she instructed pupils to share resources, and they were allowed to speak to one another while working, so there was no empirical evidence that Ms. Young was be able to gauge an individual child's knowledge and skills during the lesson. Additionally, the teacher sometimes asked questions but interrupted a pupil's response to offer her own phrasing, so that evidence of the individual pupil's understanding was limited.

Ms. Jones's lesson similarly included no formal assessment but some informal assessment. The teacher circulated frequently when pupils were completing tasks, looking at and responding to their work. However, there was little evidence that the teacher had any means to accurately assess individual knowledge and skills, because tasks were undertaken as a group and guided by the teacher rather than completed independently.

It should be apparent from these descriptions that there were strong similarities across the three lessons with regard to assessment, with overarching themes consisting of a lack

of both formal and informal assessment of individual pupil progress.

7 Integration and synthesis using the MM lens

Here we discuss the integrated findings across the various instruments and approaches. Emphasis is placed on the ways in which findings from the observation instruments and field notes elaborate on, extend, or contradict one another. This is followed by comments on the strengths and weaknesses of each of the structured observation instruments and the qualitative approach, and on the benefits and challenges of the MM approach as a whole.

7.1 Integration of results from observation schedules and field notes

Categories of the QoT and MECORS instruments do not have a one-to-one correspondence. Because these instruments are focused somewhat differently, and items are phrased differently and rated on different scales, we do not attempt a category-by-category comparison. Taking the two (or three, given a tie) highest rated categories for a particular lesson as areas of greatest strength, and the lowest two (or three, given a tie) as the areas of greatest weakness, we can establish for each observation instrument the main messages about the effectiveness of the teaching observed in each lesson. We examine how particular strengths and weaknesses in the three lessons compare as measured by these two instruments, then integrate these with themes from the qualitative findings.

7.1.1 Triangulation

First, we examine of how results from the different instruments and the field notes converge or diverge.

Some weaknesses and strengths highlighted by the two observation schedules appear to confirm one another. Ratings on both instruments suggest relative weaknesses in classroom climate and behavior management in Mr. Smith's and Ms. Young's lessons, and more effective practice in these areas in Ms. Jones's lesson. On the other hand, the two schedules lead to divergent conclusions in some cases. All three lessons have relatively high ratings on the MECORS scale for "Focuses and maintains attention on the lesson", and are likewise rated as having "More weaknesses than strengths" for "Activating pupils" (which may be seen as overlapping with though not having a one-to-one correspondence to the MECORS category mentioned) based on scores from the QoT. This may arise from the fact that the MECORS ratings are frequency-based and QoT scales are based on judgements of relative strength/weakness, or from the different phrasing of specific items. We suggest that the use of both schedules is more robust, as it is apparent that using one or the other might mask important information about observed teaching practice.

The qualitative findings enable further triangulation with the quantitative findings to see where they confirm or diverge. With regard to behavior management, the qualitative thematic analysis informs similar conclusions to the scores from the MECORS and QoT instruments, indicating that in Mr. Smith's and Ms. Young's lessons there were some noticeable off-task or disruptive behaviors and little or inconsistent responses from teachers, while in Ms. Jones's class there was very little observed disruption and some positive redirection from the teacher. In such instances where findings from both quantitative instruments and the qualitative field notes converge, conclusions about relative effectiveness of teaching practices are well supported. On the other hand, the MECORS instrument suggested that all lessons were strong in "Focusing and maintaining attention on the lesson," but qualitative findings highlighted considerable variety across lessons. Although all teachers stated objectives and checked for prior knowledge, the pacing of lessons differed greatly with many problems completed in Mr. Smith's class and few in Ms. Young's and Ms. Jones's lessons. Where the QoT indicated equal weakness in "Activating pupils" across all three lessons, the qualitative coding again revealed variation across lessons. In Ms. Jones's lesson, pupils were engaged and involved, but not specifically according to QoT items. Moreover, there was evidence of engagement in Ms. Young's lesson when pupils worked at tables, and less individual involvement in Mr. Smith's class but a use of technology that raised his lesson's mean QoT score for "Activating pupils." This echoes findings from previous studies using a similar mixed methods approach to lesson observation; in one study, qualitative field notes suggested variety in the use of mixed ability grouping (Hall et al., 2016), while in another study, it was noted that despite low ratings on quantitative scales relevant to adapting instruction and tasks, teachers were in fact engaging in informal but observable forms of differentiation that did not fit with the phrasing of the structured observation items (Sammons et al. 2014).

From this, we would suggest that the examination of results from multiple observation instruments alongside detailed qualitative notes is an important part of observing and analysing lessons. The scheduled observation approach provides a distillation of strengths and weaknesses within a particular framework, and the comparison of these across different frameworks and alongside detailed narrative field notes allows for more robust conclusions about the effectiveness of teaching practice that are less driven by the specific wording of a particular instrument or the nature of its

item scales (e.g. frequency-based ratings on the MECORS and strength-to-weakness ratings on the QoT), and that are reinforced by contextualised accounts of specific practices from the qualitative field notes. This also allows the observer (or, for the sake of teacher professional learning, an evaluator, supervisor or colleague) to engage in careful reflection before highlighting possible areas of strength/weakness or goals for professional development.

7.1.2 Elaboration and extension

The qualitative field notes not only allow for triangulation with the quantitative findings from two scheduled observation instruments, but also elaborate upon and extend those findings through detailed description of lesson features and teacher behaviors. We highlight here some key aspects of lessons for which qualitative notes provide information beyond that provided by MECORS and QoT ratings.

The MECORS and QoT instruments both provide indications of the quality or effectiveness of a teacher's feedback to pupils. In the MECORS instrument, this falls within the broader category of "Demonstrates skills in questioning"; in the QoT, under "Clear instruction". However, qualitative field notes provide greater detail regarding teachers' feedback on aspects that are missed by the quantitative instruments alone. Neither observation instrument clarifies whether feedback includes teacher responses to groups, individuals, or the whole class, whereas qualitative field notes reflect variety across the three observed lessons in this regard (Mr. Smith typically gave whole class; Ms. Young gave feedback to individuals, some behavior-focused and some including prompting before pupils had finished speaking; Ms. Jones used a range of whole-class, individual and group feedback). Depending on the purpose of the analysis, the quantitative data from observation schedules may not be specific enough to explore, evaluate, or differentiate between teachers' practice with regard to using feedback to enhance pupil learning.

Similarly, while both of the quantitative observation schedules include items relevant to behaviour management, and this is sufficient to provide an indication of general strength or weakness in this regard, qualitative field notes provide more detail to explain why ratings are high/low for a given lesson. Results from both quantitative instruments suggest a more positive climate for learning and stronger behavior management in Ms. Jones's lesson than in the other two, but findings from field notes bolster these quantitative findings by illustrating specific practices in this lesson that differed from the others: Ms. Jones tended to positively and consistently reinforce routines norms such as hand-raising, used more praise, and notes suggested she had developed positive relationships with pupils, all of which may have contributed to pupils' engagement and also proactively minimised disruptive behavior by creating a positive environment and classroom community.

A third key aspect for qualitative elaboration extending of quantitative findings concerns assessment. Although there are items on both quantitative instruments relating to formal and informal assessment, these are very general (both instruments have items phrased in terms of "checking for understanding," under "Provides pupils with review and practice" on the MECORS schedule and "Clear instruction" on the QoT; the QoT also has an item more closely linked to what we define as formal assessment, phrased as "checks the pupils' achievements", under the heading "Clear objectives"). Teachers' mean scores for the three lessons on relevant categories are middling to high across both instruments, but the qualitative findings revealed little individual assessment, informal or formal, in any of the three lessons. The field notes instead reveal descriptions of the sorts of informal assessment that took place (e.g. Mr. Smith's listening for majority responses and correcting when these were inaccurate, or Ms. Young's and Ms. Jones's circulation to look at pupils' work but not necessarily strictly independent work).

In short, the qualitative notes can add nuance to our understandings of teachers' practices and behaviors in the observed lessons above and beyond the general categories covered in structured observation schedules, and they can also help to explain the reasons behind quantitative ratings by providing contextualised descriptions of the activities, behaviors and interactions in observed lessons. This has been similarly demonstrated by prior attempts aimed at exploring features of teaching beyond those defined as effective (Sammons et al., 2014), in which it was useful to describe aspects not covered in quantitative instruments, and evaluating a particular approach to teaching mathematics (Hall et al., 2016), in which it was essential to understand how different teachers were using the approach in their classrooms and what this looked like (e.g. ways of structuring mixed-ability groups).

7.2 Strengths and weaknesses of each instrument/approach

Benefits of the MECORS instrument include its subject-specific mathematics orientation, detailed notes as part of the rating process so that ratings are more likely to be driven by evidence than if they relied on the observer's memory, and guidelines for structured coding (types of teaching and time-sampling of off-task or on-task pupils) that help to organise the note-taking process. Challenges include the use of a frequency scale (so that ratings are driven by whether and how often a practice is observed rather than the quality of the practice), and some ambiguity about scale steps [i.e.

the distinction between "often" (3) and "frequently" (4) is not obvious].

Strengths of the QoT include the use of "good practice" examples to illustrate items, relatively straightforward scales emphasising strength/weakness on each item, and minimal time to complete. Challenges of using the QoT include some poorly-defined items that may not translate well between contexts (e.g. "Effective classroom layout" for multi-subject or shared classrooms), and difficulty rating practices as observed or not observed without gradations of quality for "good practice" examples.

Strengths of the qualitative field notes include rich detail, and contextualised accounts of teaching practice, and the provision of a record to support quantitative ratings. Challenges include observer subjectivity, as including every detail of all aspects of a lesson is not feasible, so there are inherent selection processes that might involve bias when taking such field notes.

Across all of these instruments and methods, challenges are mitigated somewhat by observer training and experience, as well as careful attention to reflexivity both during and after a lesson observation. A trained mathematician and/or mathematics teacher may pick up details that a differently qualified observer might not.

7.3 Benefits and challenges of the overall mixed methods approach

A benefit of combining multiple instruments and qualitative and qualitative strands in lesson observation and analysis are that this capitalises on the strengths of each, while minimising the weaknesses of each. Quantitative observation ratings (on the MECORS and QoT) provide limited information with regard to assessment and teacher–pupil relationships, while field notes provided rich description of these features of teaching. The more content-specific MECORS informs a more detailed assessment of the use of questioning and instructional strategies, while the more generic QoT informs broader judgments of lesson quality. Drawing on multiple sources of evidence allows for triangulation and elaboration; although the MECORS and QoT emphasise some different features of a lesson they also overlap in many aspects, and qualitative field notes provide a narrative and descriptive vignettes to confirm, challenge and extend the information gleaned using quantitative instruments. As found in previous research on "inspirational" teaching (Sammons et al. 2014), qualitative field notes suggested that some of the features differentiating Ms. Jones's lesson from the other two, such as telling personal stories and laughing with pupils, went beyond descriptors of effective practice covered in quantitative instruments. For example, the use of a variety of strategies for eliciting pupil responses, both verbal and nonverbal, allowed pupils in Ms. Jones's class to participate and engage with the lesson in different ways, even during direct instruction; while quantitative findings suggested that teaching practice in this lesson showed many effective features, the qualitative field notes allowed for a more detailed account of the specific practices in context and *how* these worked for the pupils. Further, while in Ms. Jones's lesson there was little formal adaptation of teaching or activities for individual needs, but pupils were engaged and on task. Quantitative instruments alone provide little insight into informal strategies to adapting for individual learning needs, so that the description provided by qualitative field notes adds an important dimension to our understanding of a lesson, and makes a case for extending this informal adaptation aspect of the current knowledge base on effective teaching.

Ultimately, this mixing of different qualitative and quantitative evidence can support more complete explanation and understanding of effective and high quality practice.

8 Discussion and conclusions

Findings in this paper have implications for research, teacher evaluation, and professional development. In all of these applications, quantitative observation schedules provide an organised and tightly structured way of highlighting strengths and weaknesses of teaching practice, and qualitative field notes provide rich evidence of teacher behaviors, classroom climate and lesson flow. However, field notes are less useful for generalisation or comparisons with norms generated from past research based on large samples, a strength of using systematic schedules. The use of multiple observation schedules and qualitative field notes together can provide a robust and well-rounded analysis of teaching practice, whether this is intended to drive teacher evaluation, frame professional development goals, or extend research findings on effective practice to enhance the quality of teaching and learning in primary mathematics lessons.

Acknowledgements We thank the authors of the MECORS and QoT instruments for the permission to use their observation schedules in the approach we present in this paper. This work was not supported by any external funding body.

References

Carroll, J. B. (1963). A model of school learning. *Teachers College Record, 64*, 723–733.

Creemers, B. P. M. (1994). *The effective classroom*. London: Cassell.

Glaser, B. G. (1992). *Basics of grounded theory analysis: Emergence vs. forcing*. Mill Valley: Sociology Press.

Hall, J., Lindorff, A., & Sammons, P. (2016). *Evaluation of the impact and implementation of inspire mathematics in year 1 classrooms in England* (PhD Thesis). Oxford: University of Oxford.

Kington, A., Sammons, P., Brown, E., Regan, E., Ko, J., & Buckler, S. (2014). *Effective classroom practice*. Maidenhead: Open University Press.

Ko, J., Sammons, P., & Bakkum, L. (2013). *Effective teaching: A review of research and evidence*. London: CfBT.

Lincoln, Y. S., & Guba, E. G. (1985). *Naturalistic inquiry*. Beverly Hills: Sage Publications.

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.

Muijs, D., & Reynolds, D. (2000). School effectiveness and teacher effectiveness in mathematics: Some preliminary findings from the evaluation of the mathematics enhancement programme (primary). *School Effectiveness and School Improvement, 11*(3), 273–303.

Muijs, D., & Reynolds, D. (2011). *Effective teaching: Evidence and practice* (3rd edn.). London: SAGE Publications.

Said, L. (2013). *An examination of the pupil, classroom and school characteristics influencing the progress outcomes of young Maltese pupils for mathematics*. London: University of London.

Sammons, P., & Davis, S. (2016). Mixed methods approaches and their application in educational research. In D. Wyse (Ed.), *The BERA/SAGE handbook of educational research*. London: BERA/SAGE Publications.

Sammons, P., Kington, A., Lindorff-Vijayendran, A., & Ortega, L. (2014). *Inspiring teachers: Perspectives and practices*. Reading: CfBT Education Trust.

Sammons, P., Lindorff, A. M., Ortega, L., & Kington, A. (2016). Inspiring teaching: Learning from exemplary practitioners. *Journal of Professional Capital and Community, 1*(2), 124–144.

Schaffer, E. C., Muijs, D., Kitson, C., & Reynolds, D. (1998). *Mathematics enhancement classroom observation record*. Educational Effectiveness and Improvement Centre: Newcastle upon Tyne.

Schaffer, E. C., Nesselrodt, P. S., & Stringfield, S. (1994). The contributions of classroom observation to school effectiveness research. In D. Reynolds, B. P. M. Creemers, P. S. Nesselrodt, E. C. Schaffer, S. Stringfield & C. Teddlie (Eds.), *Advances in school effectiveness research and practice* (pp. 133–150). Amsterdam: Pergamon.

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.

Teddlie, C., & Sammons, P. (2010). Applications of mixed methods to the field of educational effectiveness research. In B. Creemers, L. Kyriakides & P. Sammons (Eds.), *Methodological advances in educational effectiveness research* (pp. 115–152). Milton Park: Routledge.

Teddlie, C., Virgilio, I., & Oescher, J. (1990). Development and validation of the virgilio teacher behavior instrument. *Educational and Psychological Measurement, 50*(2), 421–430.

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.

van de Grift, W. (2014). Measuring teaching quality in several European countries. *School Effectiveness and School Improvement, 25*(3), 295–311.

van de Grift, W., Matthews, P., Tabak, L., & de Rijcke, F. (2004). *Preliminary lesson observation form for evaluating the quality of teaching*. Utrecht; London: Inspectie van het Onderwijs; Office for Standards in Education.

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