Studies in Educational Evaluation

Hybrid content-specific and generic approaches to lesson observation: Possibilities and practicalities

--Manuscript Draft--

Manuscript Number:	JSEE_2020_35R2
Article Type:	VSI: Merits and Limitations of Researching
Keywords:	classroom observation; mathematics instruction; instructional quality; generic and content-specific dimensions; observational instrument(s); observation instruments
Corresponding Author:	Ariel Lindorff University of Oxford Oxford, UNITED KINGDOM
First Author:	Ariel Lindorff
Order of Authors:	Ariel Lindorff
	Armin Jentsch
	Gabriele Kaiser
	Candace Walkington
	Pamela Sammons
Abstract:	This paper considers various approaches to classroom observation that combine generic and mathematics content-specific dimensions of instructional quality. Using results from previous research in which three research teams each analysed the same three mathematics lessons (from fourth-grade mathematics classrooms in the USA) using different frameworks, we compare features of the frameworks and assess the extent to which these lead to convergent, divergent or complementary assessments of instructional quality. These findings inform reflections on how a synthesis of existing conceptualisations of instructional quality captures shared aspects of different frameworks but may be differentially useful than individual frameworks depending upon the purposes of observations. Specifically, single frameworks may be particularly useful within specific contexts and for professional development and accountability, but a synthesis of frameworks can foster more coherent cross-cultural understandings of instructional quality. We argue that establishing international networks of scholars can facilitate collaborations aiming to investigate and understand instructional quality.
Suggested Reviewers:	Sigrun Karin Ertesvag Professor, University of Stavanger sigrun.ertesvag@uis.no Expertise in classroom observation and mixed methods
	Alison Kington Professor, University of Worcester a.kington@worc.ac.uk Expertise in lesson observation and mixed methods
Response to Reviewers:	

Dear Editor-in-Chief Van Petegem,

We would like to again thank you and the reviewers for their comments and suggestions on our revised manuscript, titled “Hybrid content-specific and generic approaches to lesson observation: Possibilities and practicalities”.

We have addressed the helpful suggestions provided by Reviewer 1, with details provided below.

To facilitate review of our revisions by you and/or the reviewers, we have highlighted changes in the revised version of the manuscript. We believe that these small changes have further improved upon the most recent version of the manuscript.

We thank you for your consideration of this revised manuscript.

Detailed responses to reviewers’ comments and suggestions:

Reviewer 1:

On page 2, the authors use the phrase teacher valued-added. I think this should be value-added scores or value-added rankings.

We have added the word “scores” to “value-added” on p2 and in the corresponding footnote on the same page in the new manuscript.

The section describing the focal frameworks is much condensed and I appreciate the response to the feedback from reviewers regarding this section. One suggestion I have is to add brief descriptions (e.g., one or two sentence) of the main constructs. For example, the TEDS-Instruct dimensions include classroom management, student support, cognitive activation, but the authors could give a brief definition for each (they do this for the other two dimensions and this is quite helpful). I would suggest doing this for each framework. Similarly, for UTEP, what is meant by teacher communication of STEM content, etc. For MECORS, the authors state that there are 57 indicators, but they could give a brief overview of the categories they measure. I don't mean for the authors to greatly expand this section, merely that I think some small amount of grounding in the operational

We have added descriptions accordingly for each of the frameworks. These are highlighted in the revised submission for the reviewer’s and editor’s convenience, and are located on pages 4, 6, 7 and 8 of the revised manuscript.

Reviewer 2:

No changes were requested, and we thank the reviewer for confirming that the first round of revisions much improved the manuscript.

Title: Hybrid content-specific and generic approaches to lesson observation: Possibilities and practicalities

Highlights

- Substantial overlap in content coverage across three “hybrid” frameworks
- Marked differences, however, in conceptualisations of some instructional constructs
- Ratings of three lessons show how instructional quality may be captured differently
- Analysis across frameworks can fostering cross-cultural understandings
- Single frameworks may be more efficient for professional development/accountability

Title: Hybrid content-specific and generic approaches to lesson observation: Possibilities and practicalities

Ariel Lindorff[a] (corresponding author), e-mail: ariel.lindorff@education.ox.ac.uk

Armin Jensch[b]

Gabriele Kaiser[b,d]

Candace Walkington[c]

Pamela Sammons[a]

[a]Department of Education, University of Oxford, 15 Norham Gardens, Oxford OX2 6PY, UK.

[b]Faculty of Education, University of Hamburg, Von-Melle-Park 8, 20146 Hamburg, Germany

[c]Department of Teaching and Learning, Southern Methodist University, 6401 Airline Rd., Dallas, TX 75205, USA

[d]Institute for Learning Sciences and Teacher Education, Australian Catholic University, 229 Elizabeth Street, Brisbane CBD, QLD, 4000

Declaration of interest

None.

Acknowledgments

We would like to thank Harvard University for allowing us to use videos from the NCTE video library of lessons in classrooms in the USA for this work.

Author contributions

Ariel Lindorff: Conceptualisation, formal analysis, writing – original draft. Armin Jensch: Conceptualisation, writing – review & editing. Gabriele Kaiser: Conceptualisation, writing – review & editing. Candace Walkington: Conceptualisation, writing – review & editing. Pamela Sammons: Conceptualisation, writing – review & editing.

Title: Hybrid content-specific and generic approaches to lesson observation: Possibilities and practicalities

Abstract

This paper considers various approaches to classroom observation that combine generic and mathematics content-specific dimensions of instructional quality. Using results from previous research in which three research teams each analysed the same three mathematics lessons (from fourth-grade mathematics classrooms in the USA) using different frameworks, we compare features of the frameworks and assess the extent to which these lead to convergent, divergent or complementary assessments of instructional quality. These findings inform reflections on how a synthesis of existing conceptualisations of instructional quality captures shared aspects of different frameworks but may be differentially useful than individual frameworks depending upon the purposes of observations. Specifically, single frameworks may be particularly useful within specific contexts and for professional development and accountability, but a synthesis of frameworks can foster more coherent cross-cultural understandings of instructional quality. We argue that establishing international networks of scholars can facilitate collaborations aiming to investigate and understand instructional quality.

Keywords

instructional quality, classroom observation, observational instrument(s), mathematics instruction, generic dimensions, content-specific dimensions, observation instruments

Introduction[1]

Understanding and defining features of instructional quality in mathematics education has long been a focus of educational research and policy. Past research has demonstrated links between the quality of instruction ---- not only in mathematics lessons, but also in other subject areas -- and student achievement (see Hattie, 2008; Hill et al., 2005; Seidel & Shavelson, 2007). Although relatively early such efforts as well as the teacher effectiveness knowledge base within which they are located were criticised for being under-theorised and placing too much emphasis on easily measured behaviours (Ornstein, 1991), over time conceptualisations of instructional quality have become diverse, sophisticated and complex (Muijs et al., 2014; Schlesinger & Jentsch, 2016), and the approaches used to measure aspects of instructional quality have become a focus of educational research in their own right (see e.g. Charalambous & Praetorius, 2018b).

The majority of studies involving the observation of instruction in mathematics education do not use any formal framework (Bostic et al., 2019). However, such frameworks have been increasingly used and emphasized in the observation of instruction since the 1970s (Hilberg, Waxman & Tharp, 2004). A diverse range of frameworks exists, but debate and wide variation remain regarding what is important

[1] List of abbreviations used in this article:
ICALT: International Comparative Analysis of Learning and Teaching; MAIN-TEACH: Multi-layered and integrated quality of teaching; MECORS: Mathematics Enhancement Classroom Observation Recording System; MEP: Mathematics Enhancement Programme; MET: Measures of Effective Teaching; NCTE: National Center for Teacher Effectiveness; QoT: Quality of Teaching; STEM: Science, Technology and Mathematics; TEDS: Teacher Education and Development Study; UTOP: UTeach Observation Protocol;

to measure about instructional quality, the terminology used to describe it, and effective approaches to measuring it.

One of the key ways in which relevant frameworks differ regards whether the frameworks focus on either generic (i.e., not specifically associated with teaching in a specific subject area) or content-specific (i.e., specifically associated with the teaching and learning of a specific content area like mathematics) aspects of instruction, or in some cases, both combined. Generic frameworks such as the Classroom Assessment Scoring System (CLASS; Pianta & Hamre, 2009), the Dynamic Model of Educational Effectiveness (Creemers & Kyriakides, 2008), the Framework for Teaching (FfT; Danielson, 2013), the International Comparison of Learning and Teaching instrument (ICALT; van de Grift, 2014), the International System for Observation and Feedback (ISTOF; Teddlie et al., 2006), the Rapid Assessment of Teacher Effectiveness (RATE; Strong, 2011) and the three basic dimensions framework authored by Klieme and colleagues (2009), all have the advantage of facilitating investigations of instructional quality across subject areas, and in some cases (e.g., ISTOF and ICALT) comparisons across cultural or country contexts.

In mathematics education, however, previous theoretical and empirical work suggests that there are content-specific aspects of teaching and learning that are important (i.e., linked to student outcomes; see e.g. Hiebert & Grouws, 2007; Seidel & Shavelson, 2007) and that are not addressed in generic conceptualisations of instructional quality (Klieme & Rakoczy, 2008; Praetorius et al., 2014). Content specific frameworks are designed to capture aspects such as the use of mathematical language (Hill et al., 2008), mathematical accuracy (Learning Mathematics for Teaching, 2011), using multiple approaches to solving problems (Smith & Stein, 2011), using and connecting multiple representations of mathematical concepts (Mitchell et al., 2014), and connecting the mathematics to other topics, subjects and/or everyday life (Kersting et al., 2012).

Previous research has also compared generic and content-specific frameworks for observing instruction to gain insight into what these different lenses may capture in common or differently (e.g., Blazar et al., 2017). The Measures of Effective Teaching (MET) study (Kane & Staiger, 2012) used five frameworks to rate the same lessons on a large scale, in order to investigate correlations between ratings of instruction on different instruments and teacher value-added scores[2]; later re-analysis of the MET study data found that eight of ten dimensions of instruction were measured on all five included instruments, and classroom management was the dimension most strongly related to student outcomes based on teacher value-added scores (Gill et al., 2016). Boston and colleagues (2015) compared three observation protocols, each with different "foci and methods of use" (p. 155) within the context of mathematics instruction, and proposed that each tool was especially appropriate for one of several different emphases in investigating instruction (reform-oriented instruction, selection and implementation of cognitively challenging tasks and discussion, or mathematical rigour and richness of a lesson). Another review of six observation instruments in the context of mathematics instruction, three generic and three content-specific, emphasised the importance of taking context into account as different characteristics

[2] "Teacher value-added scores" refers to a measure of student attainment/progress attributable to the teacher, i.e. controlling for other factors such as students' prior attainment and/or background characteristics.

of instruction may be more or less relevant in different settings, and additionally highlighted how combining the use of multiple instruments and including other sources of evidence can improve reliability and robustness when making judgments about the quality of instruction (Ingram, Sammons & Lindorff, 2018).

A larger-scale effort by Charalambous and Praetorius (2018a) brought together scholars from a wide range of different contexts to consider the use of 11 different frameworks (some content-specific, some generic, and some hybrid) to investigate the quality of mathematics instruction via the empirical analysis of a common set of three video-recorded mathematics lessons in a special issue of *ZDM Mathematics Education*. Through that collaborative endeavour, Praetorius and Charalambous (2018) developed a common structure for comparing frameworks in order to describe and understand the different analyses of the same lessons arising from the use of different frameworks. This identified a list of 21 aspects of instructional quality nested in seven broader dimensions. Subsequently, Charalambous and Praetorius (In this issue) further developed this list into what they call the MAIN-Teach model of instructional quality -- so named for its multi-layered, integrated (with content-specific and generic items within dimensions where appropriate) nature, focussed on the quality of teaching -- that took into account theorised effects (both direct and indirect) on student outcomes, as well as the "layering" of dimensions in terms of how some (e.g. adaptation/differentiation) may act as precursors for others (see Figure 1). The MAIN-Teach model also rephrased potentially misleading or restrictive terminology from the list of dimensions that preceded it (Ibid).

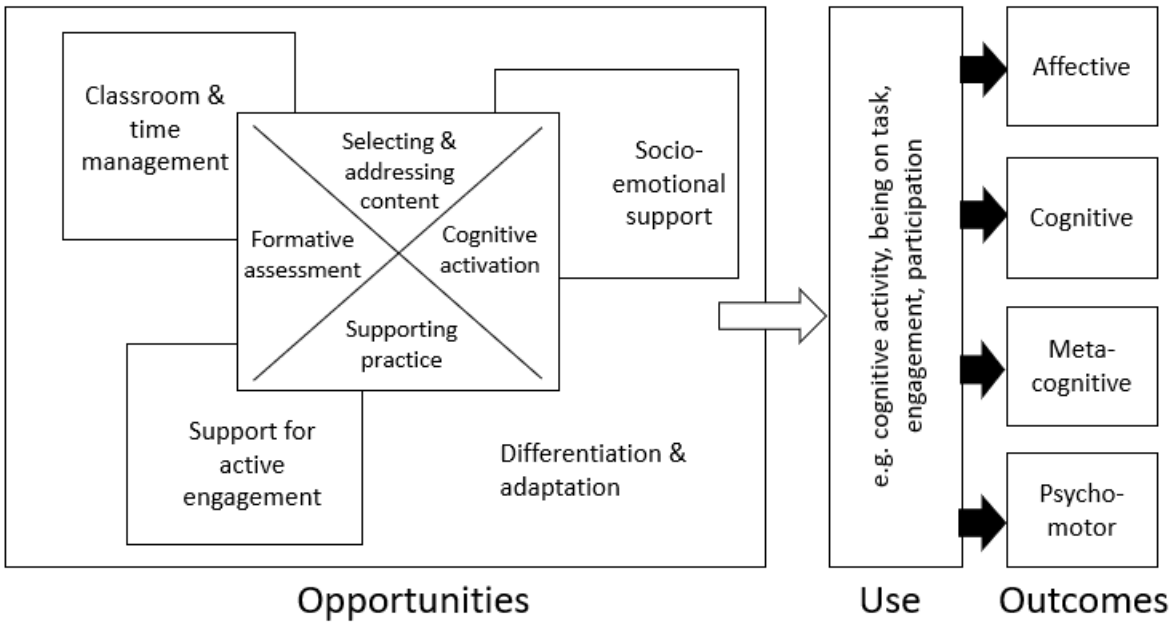

Figure 1: MAIN-Teach model (adapted from Charalambous & Praetorius, in this issue)

To our knowledge, however, previous comparisons and syntheses across frameworks have not had a specific focus on comparing hybrid frameworks via an analysis of their content as well as based on empirical results from analysing a common set of lessons. In the above-mentioned special issue, for example, articles remained relatively insulated from each other in order to maintain independence in

their application of individual frameworks to the same three lessons, and few attempts were made by the authors of the frameworks to build coherence across their findings. In this paper, we endeavour to address this gap by comparatively analysing the content of, and ratings based on, three selected hybrid frameworks. Our selection of frameworks is based on past participation of joint coding activities using hybrid frameworks, which allows us to use empirical data for the collaborative endeavour reported in this paper by comparing ratings for the same set of selected video-recorded lessons. We believe this kind of cross-team collaboration is fruitful for advancing our understanding of observation frameworks in mathematics education, as it allows for direct interrogation, comparison and synthesis of different ways of conceptualising and measuring instructional quality.

This collaborative work aims to address the following questions:

1) In what ways are the three frameworks similar or different, and what implications does this have for the ways in which they conceptualise and capture instructional quality?
    a. How are the approaches to rating lessons similar or different across the three frameworks?
    b. What are the similarities or differences across frameworks in the way in which instructional quality is conceptualised?
    c. What are the similarities or differences across frameworks in the way in which instructional quality is captured in practice?
2) What are the merits and the limitations of drawing on multiple frameworks of instructional quality as opposed to using single frameworks?
3) What are the merits and the limitations of comparing several individual frameworks as opposed to a larger-scale synthesis of existing conceptualisations of instructional quality (e.g. the MAIN-Teach model)?

Focal frameworks for this article

Observational instrument from the TEDS-Instruct project

The TEDS-Instruct observation instrument was originally developed for a project in which researchers investigated the link between teachers' mathematical competencies and student achievement, mediated by instructional quality (Jentsch et al., 2020). The generic framework used as a basis for developing the TEDS-Instruct observation instrument (see Klieme & Rakoczy, 2008; Lipowsky et al., 2009) had been commonly accepted in German-speaking countries and consisted of three dimensions: Classroom management, student support and cognitive activation. *Classroom management* concerns the practices that minimise disruption to learning time, including the use of rules and routines, effectively dealing with disruptions and delivering organised and structured lessons (Kounin, 1970). *Student support* concerns the help provided by teachers to individual students, including intervening as needed, differentiating learning opportunities, providing constructive feedback, developing and sustaining positive relationships, and fostering a positive and respectful climate for learning. *Cognitive activation* is defined in terms of the level of the learning opportunities provided to students by the teacher, and the extent to which this provides challenge and stimulates high-level learning processes. Based on a systematic review of relevant literature (Schlesinger & Jentsch, 2016), the TEDS-Instruct

authors identified two further dimensions of instructional quality particular to mathematics: *Subject-related quality* and *teaching-related quality*. The former is defined in terms of the accuracy and depth of mathematical content, while the latter concerns pedagogical aspects relevant to students' understanding and making sense of the mathematical content being taught (Authors & Others, 2018).

The full list of items and indicators for each dimension of the TEDS-Instruct instrument is given in Appendix A. Each item is rated on a scale from 1 ("Does not apply at all") to 4 ("Does fully apply"), and indicators for each item provide typical examples for each item. Items are considered to be high-inference, but six further low-inference categories were also developed for use with the instrument and rated on a scale of either 0 ("not supported"), 1 ("supported slightly"), or 2 ("mathematical competence focused in the lesson"). These low-inference categories include: Use of mathematical language; promotion of mathematical modelling; promotion of problem-solving; reasoning and proof; adequate use of calculations; and adequate use of mathematical tools. The focus of TEDS-Instruct items is primarily on the teacher rather than the students, although as the authors of the instrument note, "it is often not possible to assess the quality without considering students' interactions and reactions to the teacher's behaviour" (Authors & Others, 2018, p. 486). The original TEDS-Instruct study, in which the observational instrument for instructional quality was developed, focused on the analysis of the impact of teachers' competencies on students' learning gains with instructional quality serving as mediator (Kaiser et al., 2017) and included explicit assessment of the validity and reliability of the instrument (Authors & Others, 2018).

The authors of the instrument propose its utility for both research and practitioner-oriented purposes. The instrument was designed to be used in-vivo (i.e., in the classroom while a lesson is being taught) in order to avoid distortions. For the purpose of the analyses reported here, however, it was applied to the analysis of video-recorded lessons, which may be relevant in considering how ratings were arrived at as what is observed inevitably depends upon the angle and focus of the camera(s).

UTeach observation protocol

The UTeach observation protocol (UTOP) was designed in the context of teacher education and continuing professional development in the USA, in order to address challenges relevant to evaluating teaching and informing continuing professional development in Science, Technology, Engineering and Mathematics (STEM) teaching. The UTeach model for teacher preparation incorporates the notion that deep content knowledge is necessary but not sufficient to facilitate excellent teaching in STEM subjects (UTeach Institute, 2014a), and involves courses on how students learn STEM subjects, supervision and evaluation of student teachers by former teachers, early and continuing field experience, and compact degree plans (Author & Other, 2018). The UTOP was originally developed to evaluate the practice of in-service teachers who graduated from the UTeach programme and was designed to be applicable across the full span of a teaching career (from first-year to master-teacher status). The theory that underpins the UTeach model, and therefore the UTOP, builds on a vast body of knowledge on STEM teaching centred on six foci (Ibid):

1. *Teacher communication of STEM content:* This focuses on pedagogical content knowledge (PCK), including how teachers "integrate their subject knowledge across domains, place it meaningfully within real world contexts, and connect learning to historical and current events" (Author & Other, 2018, p508).
2. *Problem-based/discovery learning:* This concerns approaches that allow students to learn through investigation, inquiry and/or exploration.
3. *Facilitating intellectual engagement:* This focuses on strategies that facilitate deep student engagement with content, such as higher-order questioning, fostering student ownership and freedom, and providing appropriate scaffolding.
4. *Classroom management:* This involves the management of student behavioural issues, setting rules and norms, setting clear expectations, giving clear instructions, and organising space and resources efficiently, as well as facilitating high levels of time on task (Author & Other, 2018).
5. *Lesson structure and assessment:* This focuses on aspects considered to be common across different STEM lesson formats, including a clear sequence to build conceptual understanding of key ideas, activating prior knowledge, opportunities to engage with key concepts, formalisation of concepts through exploration and/or discussion, and appropriate formative and/or summative assessment.
6. *Equity, diversity and access:* This concerns how teachers account for issues of equity, diversity and access, including with regard to bias and discrimination, social justice issues, culturally relevant and inclusive pedagogy, and multicultural education.

The full set of items for each dimension of the UTOP observation instrument is given in Appendix B. Each dimension is rated on specific contributing items as well as on a synthesis rating. Each item is rated on a Likert scale from 1 ("Not observed at all/Not demonstrated at all") to 5 ("Observed to a great extent/Demonstrated to a great extent"), and synthesis ratings are also on Likert scales from 1 (lowest quality or consistency) to 5 (highest quality or consistency), with the specific scales defined differently according to the nature of a given dimension. Synthesis ratings are intended to be holistic rather than calculated from the other items within a dimension. Walkington, Marder, and Adams (in press) report five case studies of how the UTOP has been used to inform STEM teaching in different contexts, and studies validating the UTOP observation protocol have been discussed in depth elsewhere (Author & Other, 2018).

As the explicit purpose of the UTeach observation framework was to inform professional development and teacher training, it was important to include a way of illustrating and documenting evidence to support ratings. To accomplish this, observers are expected to take field notes which can be used as support for ratings, consider student work samples (if available), write one or two paragraphs describing the lesson as context for the ratings, and write summary comments addressing the evaluation of the quality of the lesson overall (UTeach Institute, 2014b). Generally, the framework is intended to be accompanied by an interview with the observed teacher after the lesson. While that component could not realistically be included in the analysis of video-recorded lessons drawn upon in this paper, it is relevant to mention in understanding the broader purpose and orientation of the framework.

A multi-instrument, mixed-method approach to lesson observation

The third framework considered in this paper is not associated with one specific observation instrument but involves combining multiple observation schedules alongside detailed qualitative field notes, on the principle that this allows for a more thorough, triangulated and detailed understanding of instruction using a mixed methods 'lens'. We shall refer to this framework hereafter as the "multi-instrument framework". The multi-instrument framework has been used in the context of both large- and small-scale studies (e.g., Kington et al., 2014; Lindorff, Hall & Sammons, 2019; Sammons et al., 2016), and is a general approach with the potential to incorporate different conceptualisations of instructional quality. The selection of structured observation instruments is chosen according to the focus of each study; where the focus is specifically on mathematics, for example in an evaluation of a specific approach to mathematics teaching in Year 1 in England (Hall, Lindorff & Sammons, 2016; Lindorff, Hall & Sammons, 2019), a combination of generic and content-specific instruments was employed to address the subject-specific focus of investigation as well as a conceptualisation of teaching quality that cuts across different subject areas. In this paper, the instruments used within the multi-instrument framework include the Mathematics Enhancement Classroom Observation Record (MECORS; Schaffer et al., 1998) and the Quality of Teaching (QoT) lesson observation form (van de Grift, 2007) which was later adapted and updated to become the International Comparison of Learning and Teaching (ICALT) instrument (van de Grift, 2014).

The MECORS instrument was developed to evaluate a particular programme in England, the Gatsby Mathematics Enhancement Programme (Muijs & Reynolds, 2000). The instrument contains a total of 57 items, each rated on a scale from 1 ("behaviour rarely observed") to 5 ("behaviour frequently observed"). Ratings on this instrument concern frequency rather than explicit judgments of quality. Items measure eight dimensions of instruction:

1. Uses classroom management techniques: Rules and consequences are clearly understood, time is used efficiently, disruptions are limited and materials are managed efficiently.
2. Maintains appropriate classroom behaviour: Behaviour is corrected accurately, immediately and constructively, and the teacher monitors the whole classroom.
3. Focuses and maintains attention on the lesson: Objectives of the lesson are clearly stated, the teacher checks for prior knowledge, teaching is clear and accurate, directions/explanations are detailed, key points are emphasised and a brisk pace and academic focus are evident.
4. Provides pupils with review and practice: Assistance to individuals and groups is effective, the teacher responds to student needs (e.g. reteaching if the error rate is high) and checks for understanding, tasks are clearly explained and the lesson is summarised.
5. Demonstrates skills in questioning: This dimension concerns a wide range of questioning and feedback techniques, e.g. asking open-ended questions, elaborating on answers, asking for more than one solution and giving immediate, accurate and positive feedback.
6. Demonstrates "Mathematics Enhancement Programme" (MEP) strategies: This involves using realistic problems/examples, encouraging and using correct mathematical language, encouraging a range of solution strategies, and connecting material to prior learning and other areas of mathematics.

7. Demonstrates a variety of teaching methods: This focuses on explanations that vary in complexity, variety of instructional approaches, and use of instructional aides such as manipulatives.
8. Establishes a positive classroom climate: This dimension concerns the teachers' enthusiasm, high expectations, positive tone, relationships with pupils, encouragement of pupil interaction and communication, and preparation of a positive and educational physical classroom environment.

By contrast, the QoT instrument was designed based on a review of the teacher effectiveness literature and consultation with school inspectors from the Netherlands and the UK and employs ratings of the quality of aspects of instruction. Items on the QoT are rated on a scale from 1 ("predominantly weak") to 4 ("predominantly strong"), with indicators of good practice associated with each item rated on a binary scale (1="yes, I have observed this" or 0="no, I didn't observe this") to support ratings. Dimensions of the QoT include:

1. Safe and orderly school climate: Promoting a relaxed atmosphere and mutual respect, and supporting pupils' self-confidence.
2. Stimulating learning climate: Stimulating independence and individual involvement, ensuring cohesion and promoting cooperation.
3. Clear objectives: Clarifying lesson objectives when the lesson begins, and assessing whether the objectives have been achieved by the time it ends.
4. Clear instruction: Giving clear instructions and explanations as well as appropriate feedback to pupils.
5. Activating pupils: Involving all pupils, and using methods or strategies that "activate" pupils such as ICT, scaffolded tasks or discussion forms.
6. Adaptation of teaching: This includes adapting instruction as well as assignments and approaches (e.g. time to complete assignments) to individual pupil needs.
7. Teaching learning strategies: Facilitating interaction in the classroom, teaching pupils ways to organise their thinking and learning, connecting to other areas of learning or to real world contexts, and allowing/stimulating alternative solutions.
8. Effective classroom organisation: This dimension concerns the extent to which the lesson is well-structured, learning time is used effectively and the lesson progresses in an orderly way, and materials are appropriate and well-organised.
9. Effective classroom layout: This dimension focuses on the physical space in the classroom, including the décor, accessibility of materials and ability for the space/furniture to be rearranged as needed.

Appendices C (MECORS) and D (QoT) provide the dimensions of these instruments along with their contributing items.

The MECORS and QoT instruments have both been validated in previous studies (for the MECORS, e.g. Muijs & Reynolds, 2000; Said, 2013; for the QoT instrument, e.g. van de Grift, 2007). The validation of the mixed methods, multi-instrument approach is less straightforward; for this purpose, it is essential to

consider the already-established validity and reliability of the chosen quantitative instruments, but also the qualitative analogues of validity and reliability (trustworthiness and credibility; Lincoln & Guba, 1985) for the qualitative field notes as well as the validity of mixing these methods (as argued by Sammons & Davis, 2017). In past studies using the multi-instrument, mixed methods framework, participant teachers' responses to the methods and findings supported the credibility of the use of the qualitative field notes as well as the overall combination of methods (e.g. Kington et al., 2014; Lindorff, Hall & Sammons, 2019; Sammons et al., 2016).

The multi-instrument framework is primarily research-oriented, although research findings based on its use are intended to inform the quality and improvement of instruction and have the potential to support professional development through teacher use and reflection. Observation instruments and field notes used within this framework are typically employed alongside other instruments such as teacher interviews and student questionnaires, in order to investigate classroom instruction from multiple stakeholder perspectives. Here, we focus only on the observation framework itself, as it was not possible to engage in other forms of data collection in the context of video-recorded lessons.

Materials and methods

Our empirical approach to addressing the research questions draws upon a previous project, in which each of several teams of researchers analysed the same three video-recorded lessons from fourth-grade mathematics classes in the USA and contributed separate papers reporting on those analyses in a special issue of *ZDM Mathematics Education* (Charalambous & Praetorius, 2018a). In this paper, we reanalyse the results from that endeavour across the three different hybrid frameworks included in that project (UTOP, TEDS-Instruct, and multi-instrument).

Focal lessons

The three video-recorded lessons were of fourth-grade mathematics lessons in classrooms in the USA. Teachers' identities have been anonymised and lessons are referred to by common teacher pseudonyms selected so that different research teams could use these to refer to specific lessons.

The first lesson, taught by "Mr Smith", had a duration of approximately 40 minutes and focused on teaching students about angles. The students were seated at tables, with the teacher at the smart board leading a whole-class discussion for the majority of the lesson, with students receiving a worksheet to practice estimating and measuring angles independently near the end of the lesson. The second lesson, taught by "Ms Young", lasted approximately 70 minutes and involved students modelling strategies to multiply whole numbers, with an emphasis on how doubling and halving affects a product. The lesson began by reviewing a homework problem and using this as a prompt for students to create a word problem in whole-class discussion, then students worked in groups to demonstrate the equivalence of two multiplication statements (15 times 8 and 30 times 4) while the teacher circulated amongst tables, before finally coming together to discuss solutions as a class. The third lesson, taught by "Ms Jones", had a duration of 56 minutes and involved teaching students several ways to multiply a whole number by a fraction. The teacher gave detailed instructions to guide students in preparing displays including an example of each method. Some time was spent with students working on their examples of the first –

previously introduced – method (writing a whole number as an improper fraction, multiplying numerators and then denominators) while the teacher circulated, and then the teacher introduced the second method (repeated addition) and led students through a practical demonstration of this method using paper circles cut into fourths before summarising new learning at the end of the lesson. More detailed descriptions of the lessons are given in the introductory paper to this special issue (Charalambous & Praetorius, in this issue), as well as in the special issue of *ZDM Mathematics Education* (see Charalambous & Praetorius, 2018a).

Framework-specific documents

Although the starting point for the collaborative endeavour resulting in this article was the three papers (Authors, 2018; Authors & Others, 2018; Author & Other, 2018) in the *ZDM Mathematics Education* special issue (Charalambous & Praetorius, 2018a) mentioned above, we reviewed a selection of relevant documents as background for our analysis to more fully account for the theoretical foundations, design, implementation and intended purpose of each framework. Depending on availability, these aspects included a combination of the following:

- Observation schedule(s) used by raters
- Any published additional guidance for the observation and rating process
- Previous publications addressing empirical support for the use of each framework
- Papers from the *ZDM Mathematics Education special* issue in which the three focal lessons were analysed using each framework

Analytical approach

Our re-analysis was undertaken in two clearly defined stages. In a first pass, we examined similarities and differences both in the frameworks themselves and in the results and conclusions that were drawn from them. This involved three stages, taking into consideration the approaches to observation, framework content and ratings of the three focal lessons, respectively.

Comparison of observation/rating approaches

First, we considered the documentation for each framework that specified how observations were undertaken. This included the extent to and ways in which quantitative and qualitative approaches (e.g. field notes) were used in each framework, the number of observers, and the intended mode of observation (e.g. video-recorded or in-vivo lesson rating).

Comparison of content

Second, using a content analysis of the dimensions, items and indicators in each framework, we created a mapping of inductively derived codes for instructional aspects measured across frameworks. Although different language was used in different frameworks, this approach allowed for as-close-as-possible

comparisons across different rating instruments to consider overlap in what common aspects were measured and the similarities and differences in their conceptualisations.

Comparison of ratings

Third, we compared the ratings of the same three lessons (noted above) from the *ZDM Mathematics Education* special issue project in order to investigate whether and how judgments of instructional quality on similar aspects of each lesson were similar or different based on the framework used. We classified ratings as high, moderate or low to facilitate comparisons of selected aspects (including some generic and some content-specific aspects of instruction), as different instruments measured aspects of instruction on different scales (making direct numerical comparisons inappropriate). Based on the comparisons of ratings on selected aspects, we then considered possible reasons for convergences and divergences in judgments across frameworks.

Finally, in order to consider the relative merits and limitations of analysing a bounded selection of frameworks as described above versus a broader synthesis of many existing frameworks, we then used the MAIN-Teach model (Charalambous & Praetorius, in this issue) as a reference point. Although the MAIN-Teach model is not unique in its conceptualisation of instructional aspects in connection to student outcomes, it is a useful example of a model arising from the analysis of a substantial number of different frameworks arising from varied settings from around the globe, making is a useful basis for comparison to our analysis.

Findings

Comparison of observation/rating approaches

In order to answer our first research question regarding similarities and differences across the three focal hybrid frameworks, we first consider the approaches taken to rating lessons using each.

There was some variation across the three frameworks concerning how generic and content-specific aspects were combined. The TEDS-Instruct and UTOP instrument designs specifically combined considerations of generic and subject-specific aspects of instruction, and subject specific items largely constituted a clearly defined sub-section of each instrument (also making up a minority of items compared to generic aspects). The multi-instrument framework, however, used one generic and one subject-specific instrument separately to rate the same lesson. Both strategies (i.e., both combined and separate instruments) would allow for correlations to be investigated across subject-specific and generic dimensions in a study with a larger sample of teachers and/or lessons, but it is possible that using combined versus separate instruments to evaluate subject-specific and generic aspects may have differentially affected the way in which observers approached their ratings for these two categories of instructional aspects (Popping, 2010).

All three frameworks involved quantitative ratings of items, but the nature and use of qualitative data varied. For the TEDS-Instruct framework, lessons were transcribed and coded through a Qualitative

Content Analysis approach using deductive codes; this simultaneously placed an emphasis on classroom discourse (rather than nonverbal cues or interactions) and drew upon relevant theory to define a-priori coding categories. For the UTOP framework, qualitative data consisted of observer field notes, which were used to provide evidence to support ratings. The multi-instrument framework also employed researcher field notes, but these were coded inductively later in the analysis process in order to triangulate with and elaborate upon numeric ratings (completed immediately after lesson observations) as well as to illuminate aspects of the lessons not covered in structured observation instruments.

For the TEDS-Instruct and UTOP teams, two observers rated each lesson. The multi-instrument team used one observer in this instance but noted that multiple observers would be preferable (resources allowing) in a larger study. Interestingly, the purpose for having more than one observer for a given lesson was not always as simple as establishing rating reliability or validity. The TEDS-Instruct team, for example, called particular attention to the possibility that disagreement between observers facilitate substantively useful intellectual discussions and, in turn, deeper understandings of teaching and learning.

None of the three frameworks was specifically designed for use with video-recorded lessons. The TEDS-Instruct instrument was designed for in-vivo rating, while the UTOP instrument included some items designed to be accompanied by a post-observation teacher interview. The multi-instrument framework was designed for use alongside additional data collection methods intended to draw on multiple perspectives, particularly teacher interviews and student questionnaires. On one hand, the comparative analysis across these three frameworks based on video lessons is missing some of the context of an in-vivo lesson observation. On the other hand, all observers watched lessons from the same perspective (forced by the focus of video recordings), which helped to standardise what they observed in each lesson more than would have been possible with multiple observers in a live classroom.

The number of items included in each framework suggests some trade-offs between specificity and rating efficiency. The MECORS used in the multi-instrument framework contains the largest number of items, and a larger proportion of subject-specific items, than any of the other observation instruments used in the three frameworks, which naturally entails a larger time investment to complete ratings for a given lesson. This is somewhat mitigated by the lower level of inference required for the MECORS compared to the other observation instruments, while providing a finer-grained description (based on quantitative ratings alone) of subject-specific aspects. The trade-off between detail and efficiency is an important consideration in assessing the fit-for-purpose of using a particular framework or combination of frameworks, though this consideration is not unique to hybrid frameworks. If the purpose for evaluating instruction is formative and intended to support individual teachers' professional development, the decision of whether to opt for detail or efficiency depends on whether an evaluation will be used to define goals (in which case it might be sensible to gain an efficient and global overview of instructional strengths and weaknesses) or to assess progress towards existing targets (in which case more detailed and specific information might be more useful). If the purpose of observation is to contribute to a summative evaluation of instructional quality (e.g. as part of a high-stakes promotion or retention protocol for evaluating individual teachers), efficiency is likely to be a key consideration as

best practice would entail the observation of multiple lessons rather than a snapshot of one lesson. In general, however, we would not recommend the use of observation for high-stakes, summative purposes, especially in isolation (i.e. as the sole source of information about instructional quality), as these are likely to bias teacher behaviour during observations and therefore provide a less-than-authentic representation of a particular teacher's instructional quality.

Comparison of framework dimensions and items

Still focusing on our first research question regarding similarities and differences across frameworks, we now turn to the ways in which instructional quality is conceptualised within each. Although the language used to describe the different structural levels of observation instruments varies by framework, here we use a common set of terms to facilitate comparisons across frameworks. "Dimensions" henceforth refers to broad constructs within a particular framework, "items" refers to the more specific aspects nested within a dimension which are rated by the observer, and "indicators" refers to even more finely-grained descriptors nested within each item (as applicable; not all frameworks contain three levels of description in their structured observation instruments).

In comparing the conceptualisations of instructional quality in the three hybrid frameworks, we took into account whether a given aspect of instruction was covered at all in each framework, how it was categorised (as subject-specific or generic), its grain size (i.e. as a dimension of instruction or as a specific item or number of items nested within a dimension or dimensions), and its apparent relationship to other commonly-covered aspects (nested within or overarching). Table 1 presents the reader with the mapping of aspects of instruction and their operationalisation across frameworks.

An important point learned from the process of mapping constructs and conceptualisations across frameworks and instruments is that the distinction between content-specific and generic aspects is not always clear or consistent, and individual aspects of instruction are conceptualised as having very different grain sizes and requiring different levels of detail within different frameworks. However, overall there is a great deal of overlap in the concepts covered across frameworks, despite their being ordered and nested in different ways. Variations in the level of detail within different frameworks are worthy of further consideration in terms of the ways in which instructional quality is captured in each and may reflect differences in the priority accorded to different dimensions of instruction or views about their ease of measurement. Rather than a superficial overview of the full range of aspects of instruction addressed by the various frameworks, we focus in depth below on a selection of three aspects conceptualised as consistently generic, variously content-specific and generic, and consistently content-specific across frameworks, in order to illustrate these main findings.

Table 1: Mapping of aspects of instruction and their operationalisation across frameworks

		TEDS-Instruct	UTOP	Multi-instrument	
				MECORS	QOT
Classroom management	*Grain size**	1 dimension; 6 items	5 items	2 dimensions; 10 items	3 items

	Nested within dimension(s)?	n/a	- Classroom environment - Implementation	- Uses appropriate classroom management techniques - Maintains appropriate classroom behaviour	Effective classroom organisation
Classroom environment	*Grain size*	4 items	1 dimension; 7 items	1 dimension; 11 items	3 dimensions; 12 items
	Nested within dimension(s)?	- Classroom management - Student support	n/a	- Establishes a positive classroom climate - Focuses and maintains attention on lesson - Provides students with review and practice	- Safe and orderly climate - Stimulating learning climate - Activating students - Teaching learning strategies - Effective classroom layout
Lesson structure	*Grain size*	1 item	1 dimension; 6 items	1 item	1 item
	Nested within dimension(s)?	Classroom management	n/a	Focuses and maintains attention on lesson	Effective classroom organisation
Objectives	*Grain size*	1 indicator	Part of 1 item	1 item	1 dimension; 2 items
	Nested within dimension(s)?	Classroom management	Lesson structure	Focuses and maintains attention on lesson	Clear objectives
Questioning	*Grain size*	3 items	2 items	12 items	3 indicators
	Nested within dimension(s)?	Cognitive activation	Implementation	- Demonstrates skills in questioning - Demonstrates MEP strategies*	- Clear instruction - Activating students
Adaptation/ differentiation	*Grain size*	2 items; 1 indicator	2 items	1 item	1 dimension; 2 items
	Nested within dimension(s)?	- Student support - Teaching-related quality	- Classroom environment - Implementation	Provides students with review & practice	Adaptation of teaching
Problem solving	*Grain size*	1 qualitative aspect	1 item	2 items	1 item
	Nested within dimension(s)?	n/a	Lesson structure	Demonstrates MEP strategies	Teaching learning strategies
Connections	*Grain size*	2 items; 2 indicators; 1 qualitative aspect	3 items	3 items	1 item; 1 indicator
	Nested within dimension(s)?	- Cognitive activation - Subject-related quality - Teaching-related quality	- Implementation - Mathematics content	Demonstrates MEP strategies	- Clear instruction - Teaching learning strategies
Feedback/ Formative assessment	*Grain size*	2 items	3 items	6 items	1 item
	Nested within dimension(s)?	- Student support - Subject-related quality	- Lesson structure - Implementation	- Focuses and maintains attention on lesson - Provides students with review & practice - Demonstrates skills in questioning	Clear instruction
Practice/Fluency/ Summary	*Grain size*	2 items	1 item	1 item	1 indicator
	Nested within dimension(s)?	- Cognitive activation - Teaching-related quality	Mathematics content	Provides students with review & practice	Clear instruction
Explanations	*Grain size*	1 item	2 items	3 items	2 items
	Nested within dimension(s)?	Subject-related quality	- Mathematics Content	- Focuses and maintains attention on lesson - Provides students with review & practice - Demonstrates a variety of teaching methods	Clear instruction

Quality of content	*Grain size*	2 items	3 items	n/a	n/a
	Nested within dimension(s)?	- Subject-related quality '- Teaching-related quality	Mathematics content	n/a	n/a
Correctness/ Accuracy	*Grain size*	1 item	2 items	1 item	n/a
	Nested within dimension(s)?	Subject-related quality	Mathematics content	Focuses and maintains attention on lesson	n/a
Multiple representations	*Grain size*	1 item	1 item	1 item	n/a
	Nested within dimension(s)?	Teaching-related quality	Mathematics content	Demonstrates a variety of teaching methods	n/a
Mathematical language	*Grain size*	1 indicator; 1 qualitative aspect	Part of 2 items	2 items	1 indicator
	Nested within dimension(s)?	Subject-related quality	Mathematics content	Demonstrates MEP strategies	Effective classroom layout

*"Demonstrates MEP strategies" in the MECORS instrument refers to strategies from the Mathematics Enhancement Programme based on a review of the relevant literature (Muijs & Reynolds, 2000).

All frameworks included some coverage of *Classroom management*, but the grain size of this aspect of instruction varied. In the TEDS-Instruct framework, classroom management consisted of a generic dimension measured via six items, including the use of lesson time, clear rules and routines, preventing disturbances, productive atmosphere, advance organisation and structuring of learning processes, and lesson structure. These last two items were not conceptualised as part of classroom management in other frameworks. In the UTOP framework, classroom management was measured via five items nested within the generic *Classroom environment* and *Implementation* dimensions; the descriptors for these items covered management strategies enhancing the classroom environment, students on task, classroom organization and setup, lesson timing, and lab safety. In the MECORS instrument used in the multi-instrument framework, classroom management was measured via two generic dimensions with five items each. The first dimension, *Uses classroom management techniques*, included items concerning rules and consequences being understood, starting the lesson on time, using transition time effectively, having materials ready and collecting/distributing them effectively, and limited disruptions. The second dimension, *Maintains appropriate classroom behaviour,* included items concerning using a reward system to manage behaviour, correcting behaviour immediately, accurately and constructively, and monitoring the entire classroom. In the generic QoT instrument also used in the multi-instrument framework, the term "classroom management" was literally interpreted to mean the management of materials during the lesson, but 3 of the items within the *Effective classroom organisation* dimension could be conceived of as measuring classroom management with a broader interpretation. These items addressed orderly progression of the lesson (e.g. entering and leaving the classroom in an orderly manner, intervening in a timely and appropriate way to any disruptions, and acting as a watchdog for agreed codes of behaviour and rules), efficient use of learning time, and ensuring classroom management in terms of having lesson materials ready to use and making clear which materials to use as well as lesson materials being adapted to the level and experience of students. These different operationalisations indicate that conceptualisations of classroom management vary considerably, and where this aspect falls in terms of its importance for instruction and relationship to other aspects of instruction can also be quite different in different frameworks. One key area of difference concerned what was being "managed". All frameworks included some consideration of behaviour and time

management, but in the TEDS-Instruct instrument alone, what was being "managed" included the structure of the lesson itself. In the UTOP and QoT instruments, classroom management included physical classroom organisation and setup. The level of specificity with which frameworks took a particular stance on what constituted effective classroom management also differed; the MECORS took the most specific and arguably prescriptive approach, for example, with an item concerning the use of a reward system. A similarity across frameworks, however, was that classroom management was consistently conceived of as a generic aspect of instruction[3].

Questioning is considered a generic aspect in most of the frameworks, but not all. In the TEDS-Instruct instrument, three items within *Cognitive activation* relate implicitly if not explicitly to questioning, including challenging questions and tasks, supporting metacognition, and activating prior knowledge and co-construction of knowledge in the lesson. In the UTOP instrument, questioning is addressed by 2 items focussed on questioning strategies that encourage participation, check skill development, and develop conceptual understanding, within the *Implementation* dimension. In the MECORS instrument, *Demonstrates skills in questioning* constitutes a dimension that is at least somewhat content-specific, with 11 items pertaining to the frequency, academic nature, and open-endedness of questions as well as probing when responses are incorrect, elaborating on answers, asking students how they reached their solutions, asking for more than one solution, using appropriate wait time, noting students' mistakes, guiding students through errors, and clearing up misconceptions. While some of these items could apply across subject areas, some are more specific to mathematics (e.g. those focusing on how students reached their solutions and asking for multiple solutions). One further item in the MECORS instrument (in the *Demonstrates MEP strategies*[4] dimension) specifically concerns the use of a quick-fire mental questioning strategy. In the generic QoT instrument, no items exclusively focus on questioning, but 2 indicators for the item concerning involving all students in the lesson (within the *Activating students* dimension) and 1 indicator for the item concerning clear instructions and explanations (within the *Clear instruction* dimension) relate to questioning. The main points of difference in the ways in which questioning is conceptualised across instruments concern the level of specificity with which questioning strategies are explicitly addressed, and the extent to which the evidence of instructional quality on this aspect considers the teacher's questioning or its outcomes. For example, whereas the MECORS explicitly addresses a long and specific list of different questioning strategies with the emphasis on what the teacher asks, other frameworks appear to focus more on what happens as a result of the teacher's questioning (e.g. student metacognition, co-construction of knowledge, participation, conceptual understanding).

A few constructs are consistently placed within content-specific dimensions of instruction. For example, *Multiple representations* are covered in each framework, though to a varying extent; in the TEDS-Instruct framework, this constitutes an explicit item within the *Teaching-related quality* dimension,

[3] The UTOP considered lab safety as part of classrooms management, and this would not necessarily apply to non-STEM disciplines, but is less relevant to mathematics than to lab sciences.

[4] "Demonstrates MEP strategies" in the MECORS instrument refers to strategies from the Mathematics Enhancement Programme based on a review of the relevant literature (Muijs & Reynolds, 2000).

while in the UTOP instrument multiple representations are given as an example of the use of elements of abstraction within the *Mathematics content* dimension, and the MECORS instrument contains an item concerning the use of manipulatives/learning aids within the *Demonstrates a variety of teaching methods* dimension but does not explicitly mention multiple representations. The QoT instrument does not address this aspect at all, which is reasonable given its generic focus. Although different terminology is used for this aspect of instructional quality across frameworks, there appears to be considerable agreement regarding its importance in a content-specific context.

Comparison of ratings

In order to understand similarities and differences across our focal frameworks, it is important to consider how these frameworks capture instructional quality in practice. As above, we focus our analysis of ratings across the three frameworks for the three focal lessons on the same three commonly measured aspects of instruction in order to illustrate main findings. As the rating scales used for the three hybrid frameworks differ considerably, both in the scales and meanings of ratings (i.e. operationalisation in terms of frequency versus judgements of quality, or both), it would be inappropriate to compare scores mathematically. Instead, in this section we consider whether aspects measured across all frameworks are rated similarly or differently in terms of high, moderate or low ratings, and how interpretations of those ratings converge or diverge across frameworks. We further explore examples in the qualitative notes from the same lessons used to justify ratings, in order to provide insight into the ways in which observers interpreted the same set of behaviours and discourses in the three focal lessons.

From this empirical comparison of ratings across frameworks we learn that where the same aspects of instruction were differently conceptualised across frameworks (as identified in the comparisons across dimensions and items, above), this could lead to quite different judgments of instructional quality. Further, some of the differences observed in rating results across frameworks reveal possible consequences of the influences of different cultures, experience and orientations across frameworks, as evidenced by both the behaviours and interactions on which observers focused (based on qualitative comments/notes) and their judgments (of high/middle/low frequency and/or quality).

Classroom management

Table 2: Ratings relevant to classroom management across frameworks and lessons

		Lesson		
Framework/Instrument		*Mr Smith*	*Ms Young*	*Ms Jones*
	TEDS-Instruct	Moderate-high	Moderate	Moderate-high
	UTOP	High	Low	Moderate-high

Multi-instrument framework	*MECORS*	Low	Low	High
	QoT	Low	Low	High

Mr Smith's lesson was rated moderate-to-high for most relevant items on the TEDS-Instruct instrument, but lower for advance organisation and lesson structure, while UTOP ratings were high and MECORS ratings were generally low for classroom management-related items. On the QoT, ratings for the relevant items were low except for a higher rating with regard to the management of materials. Thus, the TEDS-Instruct and UTOP frameworks paint a broadly positive picture of classroom management in this lesson, while the ratings on the multi-instrument framework are less favourable. The UTOP results include a description of the class as generally engaged, at least on the surface, and students being on-task and following the lesson. Qualitative findings from field notes used within the multi-instrument framework included descriptions of several low-level disruptions (e.g. off-task conversations between students, a student making faces at the camera), but also a sense that routines and norms had been established prior to the lesson despite not being made explicit.

For Ms Young's lesson, TEDS-Instruct ratings were moderate, UTOP ratings were low, and MECORS ratings were low with the exception of accurate correction of behaviour. On the QoT instrument, ratings were fairly low with regard to the efficient use of time and orderly progression of the lesson, but higher on the management of lesson materials. These results suggest that the researchers using the different frameworks had, to some extent, different interpretations of the same behaviours in terms of classroom management overall. Qualitative comments from the TEDS-Instruct research team indicate that rules and routines had to be repeated in the lesson, and descriptions from the UTOP and multi-instrument research teams drew attention to difficulties and inconsistencies in Ms Young's behaviour management within the lesson (e.g., one student was removed from the room, while some disruptive behaviours went unaddressed or were inefficiently addressed, and transitions – for example, between the carpet and tables -- were poorly organised). While there was a consensus that classroom management in this lesson revealed some problems, the judgments made as a result clearly varied in their severity across frameworks, and different features of the teacher's practice with regard to this aspect seemed to stand out to different observers using the different frameworks (or, in fact, different observers using the same framework, as was noted by the UTOP research team; Author & Other, 2018, p. 514).

Ms Jones's lesson received moderate-to-high scores on the TEDS-Instruct instrument for most classroom management-related items, but low scores on advance organisation and structuring of learning processes/lesson structure. On the UTOP instrument, this lesson was rated highly on most classroom management-related items but lower on timing. MECORS ratings for this lesson on items relevant to classroom management were mostly high, with the exception of a moderate score for correcting behaviour accurately and effective use of transition time, and QoT ratings were generally high. The varied judgments across frameworks and instruments relating to specific aspects of classroom management (notably with regard to the use of time) reflects different conceptualisations of aspects contributing to the broader construct of classroom management. Qualitative accounts collectively note Ms Jones's demonstration of clear classroom routines and rules as well as prevention of and responses

to disruptions, but the description in the UTOP findings called attention to the tight level of control exerted by the teacher and wasted time on less consequential portions of the lesson as negative aspects of classroom management in this lesson.

Classroom management provides a useful example of how observers' and frameworks' different cultures, experience and orientations may have played a role in differences in ratings across frameworks. For example, structure and timing seem to be similarly rated across frameworks, possibly because these are less likely aspects of classroom management to be perceived differently across Germany (TEDS-Instruct), the USA (UTOP) and the UK (multi-instrument). However, perceptions of the severity of problems (e.g. with regard to what constituted a disruption) were more varied, which may reflect cultural differences in the interpretations of the same observed behaviours in a given lesson. Differences may also be at least to some extent a result of the different underlying scales and their meaning. The MECORS, in contrast to the other three instruments, focuses primarily on *frequency*, the QoT on strengths vs. weaknesses, the TEDS-Instruct on the extent to which an aspect of teaching applied to the lesson, and the UTOP on frequency and/or quality depending upon the specific item being rated.

Questioning

Table 3: Ratings relevant to questioning across frameworks and lessons

		Lesson		
Framework/Instrument		*Mr Smith*	*Ms Young*	*Ms Jones*
	TEDS-Instruct	Low	Low-moderate	Low-moderate
	UTOP	Mix of high/low	Low-moderate	Low
Multi-instrument framework	*MECORS*	Mix of high/low	Mix of high/low	Mix of high/low
	QoT	Mix of high/low	Mix of high/low	Mix of high/low

Mr Smith's lesson received low scores for questioning-related items on the TEDS-Instruct instrument. On the UTOP instrument, Mr Smith scored highly on using questions to facilitate interaction, but lower on higher-order questioning to develop understanding. MECORS ratings were high on most questioning items, but low on higher-order questioning and use of appropriate wait time. QoT ratings were high for asking questions that were understood by students but low for posing questions that initiate reflection and using sufficient wait time. Qualitative descriptions from the different frameworks similarly pointed to the emphasis on students responding in chorus and the lack of wait time between questions and answers. With regard to questioning, then, there seems to be some degree of agreement across frameworks. Many questions were asked in the lesson, but these were largely lower-order questions with little time given for students to consider their responses or to respond individually.

For Ms Young's lesson, ratings were moderate for questioning-related items on the TEDS-Instruct instrument, except for a low rating for supporting metacognition. UTOP ratings were moderate for using questioning to develop skills but low for higher-order questioning. On the MECORS instrument, this lesson received a mixture of high, moderate and low ratings (e.g. high for academic questions, moderate for open-ended questions and probing incorrect responses, and low for asking for more than one solution and using appropriate wait time). QoT ratings were high for asking questions that were understood by the students, but low for posing questions that initiate reflection and giving sufficient wait time. There was some consensus across frameworks and instruments concerning the limited provision of wait time between questions and answers, but qualitative descriptions called attention to some of the finer nuances of questioning in this lesson. For example, the UTOP research team noted that some higher-order questions were asked, but that the teacher spent a great deal of time monologuing without asking questions, while findings from qualitative field notes in the multi-instrument framework noted that there was no evidence of the use of strategies to involve non-volunteers.

Ms Jones's lesson was rated low-to-moderate on the TEDS-Instruct framework for questioning-related items, and notes accompanying the ratings suggested prior knowledge was activated somewhat but knowledge was not co-constructed. UTOP ratings were low on the relevant scales. MECORS ratings were a mixture of high (e.g. frequent and academic questions, elaborating on answers, noting students' mistakes and guiding them through errors), moderate (e.g. open-ended questions, probing incorrect responses, and asking for explanations), and low (e.g. asking for more than one solution, appropriate wait time, clearing up misconceptions, and use of a rapid-fire mental questioning strategy). On the QoT, ratings were high for asking questions that were understood by the students, but low for posing questions that initiate reflection, use of appropriate wait time and giving opportunities to non-volunteers. Qualitative comments on Ms Jones's lesson elaborate on the focus on students copying down strategies (TEDS-Instruct), lack of higher-order questions (UTOP), and instances of individual probing while the teacher circulated (multi-instrument framework).

The above results suggest some specific aspects of questioning for which ratings diverged and converged across frameworks based on the conceptualisation of questioning, the level of inference required, and the grain size of the items used to measure it. Areas of consensus seemed to relate to aspects of questioning that required lower inference (e.g. frequency of questions, or whether open-ended questions were asked). Where ratings differed across frameworks, this seems at least in part to be due to the different ways in which questioning was conceptualised. For example, the TEDS-Instruct framework included explicit consideration of facilitating metacognition, which was conspicuously absent from the descriptors of items in other instruments. Meanwhile, the MECORS instrument broke questioning down into a large number of finely-grained items concerning highly-specific aspects of questioning that may have been implicitly considered in other frameworks, but did not constitute distinct items (e.g. probing incorrect responses or asking for multiple solutions).

Multiple representations

Table 4: Ratings relevant to multiple representation across frameworks and lessons

		Lesson		
Framework/Instrument		*Mr Smith*	*Ms Young*	*Ms Jones*
	TEDS-Instruct	Low	Moderate	Low
	UTOP	Moderate	Moderate	Moderate
Multi-instrument framework	*MECORS*	High	High	Moderate
	QoT	Not applicable	Not applicable	Not applicable

Mr Smith's lesson was rated low for use of multiple representations on the TEDS-Instruct instrument, moderate on the UTOP instrument, and fairly high on the MECORS instrument. Qualitative notes within the TEDS-Instruct findings mentioned that angles were represented figuratively and symbolically, but that these representations were not always clearly connected to one another, which helps to explain the apparent discrepancy across frameworks depending upon whether multiple representations were rated based on their existence or their explicit connection.

On the TEDS-Instruct instrument, Ms Young's lesson was rated moderately for using multiple representations, with qualitative comments mentioning the use of verbal, symbolic and figurative representations of whole numbers and linking these representations multiple times. The relevant rating on the UTOP instrument was moderate, but no explicit mention of multiple representations was made in the qualitative comments, which instead focused on the algorithm used. The MECORS rating was high for use of manipulatives, which aligns with the mention of certain representations (cubes) in the TEDS-Instruct qualitative notes.

Ms Jones's lesson was rated low on the TEDS-Instruct instrument, but moderate on UTOP and MECORS instruments. Qualitative comments from the TEDS-Instruct findings help to explain this discrepancy, noting that although a combination of verbal, figurative and symbolic representations were used, these were not well-connected.

Overall, discrepancies in ratings across frameworks for items relevant to the use of multiple representations are very likely to be attributable to explicitly different ways of conceptualising and measuring this aspect of content-specific instruction. While the TEDS-Instruct framework accounts for the explicit use of and connections between multiple representations, the most relevant item within the UTOP framework pertains to elements of abstraction *including* multiple representations, and the MECORS instrument only accounts for the use of resources and manipulatives. Given the differing observable activities that would be rated based on each of these conceptualisations, it is unsurprising that ratings differed across these frameworks for this aspect of instruction.

Merits and limitations of drawing on multiple frameworks rather than single approaches

To answer our second research question, we assess the merits and limitations of considering multiple rather than single frameworks as we have done above.

Some useful insights arise from these comparisons. The analyses of content and ratings show that the different frameworks cover many of the same fundamental aspects of instruction, and there is often considerable alignment across at least some frameworks in the judgments made about the same set of lessons. However, there is also evidence of different interpretations of the same observed behaviours and interactions as well as core differences in the ways in which seemingly similar concepts are applied across frameworks. Analysing where the findings based on different frameworks with respect to these three lessons converge and diverge helps to shed light on a more nuanced and detailed picture of instruction in each lesson, and facilitates understandings of *why* ratings differed, where they differed, with the help of the qualitative descriptions recorded by each research team to accompany their ratings.

This type of multiple-framework comparison also poses challenges. It involves considerable complexity and detail, and reveals that terminology can become ambiguous when different frameworks have emerged from parallel literatures in different cultures and contexts. The time required to account for multiple frameworks and comparisons across them is substantial, so that for some purposes the efficiency of using a single framework selected for appropriateness for a particular context could be preferable.

Merits and limitations of the above approach using MAIN-Teach as a reference point

Our inductive approach to comparing the three hybrid frameworks is productive as it avoids an over-emphasis on a single (potentially context-bound or culturally biased) theoretical conceptualisation of teaching quality. There are, however, important limitations to an inductive, specific approach to comparing different observation frameworks like that undertaken in the preceding sections.

In order to answer our third research question, we consider the benefits and limitations of our analytical approach above by using the MAIN-Teach model as a reference point to consider what might have been missed as well as what additional insights our specific comparisons provide. It is worth noting that we use this model because it arose from a synthesis of many different frameworks for observing instructional quality, not because it is unique in conceptualising the links between instruction and student outcomes; many such models have preceded the MAIN-Teach model in the education literature in this respect (e.g. Hill & Chin, 2018; Kaiser et al., 2017, Kersting et al., 2012).

Although both the analysis of our three specific frameworks and the MAIN-Teach model include a hybrid of both content-specific and generic aspects of instruction, there are some clear points of contrast between our findings and the MAIN-Teach model's synthesis of frameworks:

Prioritisation/ordering of dimensions of instructional quality

The explicitly multi-layered structure of the MAIN-Teach model attempts to theorise relationships between different aspects of instruction and, in a sense, to order them (i.e., in terms of which aspect might be a precondition for or facilitate another). For example, differentiation and adaptation forms a layer underlying other aspects of instruction in the model (see Figure 1). This feature is absent from the TEDS-Instruct, UTOP and multi-instrument frameworks, in which items or dimensions relating to differentiation and adaptation are part of a list of instructional aspects, rather than a prioritised organisation conceptualising one aspect as a precondition for another. Our analysis therefore starts at the more basic level of comparing across lists of instructional aspects, while the MAIN-Teach model draws on a larger number of frameworks and builds upon previous comparative work.

There are some advantages to drawing upon a model of the relationships between different aspects of instruction, rather than a list-based comparison. Depending on the purpose for observing, a theoretically driven model (such as MAIN-Teach) can help to determine priorities for improvement (e.g. for earlier-career teachers, it might be sensible to focus on the aspects that are preconditions for others), or can serve as a sense-making device to construct a coherent narrative to describe instruction in a particular lesson and to search for patterns across lessons/classrooms.

Conceptualisation of dimensions of instruction as generic/content-specific/both

The MAIN-Teach model explicitly addresses how some dimensions of instruction may be treated as both generic and content-specific. For example, different aspects of quality feedback may be generic (e.g., concrete) or content-specific (e.g., mathematically accurate) in the model. In each of our three hybrid frameworks, there was less specific and explicit treatment of the dual (content-specific and generic) nature of certain dimensions. This helps to explain why there were some observed differences in ratings across the TEDS-Instruct, UTOP and multi-instrument frameworks when similarly worded items are conceptualised differently across individual instruments. For example, the operationalisation and judgments of questioning in ratings of the three focal lessons above picked up generic aspects in some frameworks (e.g. co-construction of knowledge) and content-specific aspects in others (e.g. asking for more than one solution).

Use of consistent terminology

The MAIN-Teach model further offers a common language to describe and discuss aspects of instruction, whereas it is apparent that the same terms are used differently, and different terms are used, across the three frameworks analysed above. For example, the MAIN-Teach model explicitly addresses three separate aspects of classroom management (behaviour, time and physical space/materials). The three hybrid frameworks, as noted above in our analysis, variously prioritise these aspects in their conceptualisations of classroom management. The MAIN-Teach model thus provides a structure for reflecting on the emphases and definitions used in different individual frameworks and for understanding the ways in which their conceptualisations of aspects of instructional quality may differ.

Nonetheless, there are also limitations of the MAIN-Teach model and similar syntheses of existing work on a larger scale in comparison to our bounded analysis of three frameworks.

Comprehensiveness instead of specificity

The level of detail in a large-scale synthesis risks being less rich as the result of an attempt to account for a large number of different frameworks, while our analysis of the content of our three frameworks and the ratings based on them is able to provide highly-specific information about each framework and the comparisons between them. For example, whereas behaviour management is conceptualised only in terms of prevention (rules, routines and monitoring) and intervention (dealing with disorder in a constructive and timely manner) in the MAIN-Teach model, individual frameworks vary in the specificity and detail with which they characterise the features of classroom management in a lesson. The differences that emerge from comparisons across the individual frameworks yield insight into how classroom management is conceptualised differently in each (e.g. in terms of grain size, emphasis on behaviour/time/physical space, and prescriptiveness). A large-scale synthesis can provide a summary of key themes across frameworks but loses some of these nuanced differences in conceptualisation.

Loss of context-specificity

The three focal hybrid frameworks each emerged in specific contexts (Germany, the UK and the USA) and for particular purposes, as discussed in the "Focal frameworks for this article" section above. In synthesising frameworks across contexts, the MAIN-Teach model has the potential to provide shared understandings but risks losing context-bound priorities that reflect local emphases in policy and practice. For applications that are context-specific such as teacher evaluation for accountability purposes and observation-driven teacher professional development, comparisons across individual frameworks can facilitate explicit discussion of contextual differences and emphases whereas the MAIN-Teach model's synthesis loses these details.

In short, the MAIN-Teach model and similar syntheses of existing work on a large scale are likely to be useful for facilitating discussion and debate between different researchers across cultures and conceptualisations of instructional quality using a common language. Individual frameworks and analyses across them may, however, be more useful for culturally- and context-bound applications to professional development and accountability when specificity and fine-grained detail are essential to inform improvement-oriented goal-setting, professional development and formative evaluation.

Conclusions

Our analysis of the TEDS-Instruct, UTOP and multi-instrument frameworks shows that the different instruments used to rate instructional quality had considerable overlap in their coverage of aspects of instruction, but the grain size at which these aspects were captured and the dimensions within they were conceptually nested differed, as did (in many cases) their characterisation as content-specific or generic. The extent to which aspects that were measured across frameworks were rated similarly varied somewhat, seemingly due to a combination of different interpretations of the same observed behaviours and different conceptualisations of aspects of instruction despite (in some cases) apparently similarly worded items. Qualitative descriptions from each research team help to illustrate and elaborate on differences in ratings where these occurred, and make it possible to deduce where

researchers had noticed different features of lessons relevant to the same aspect of instruction, or how specific details concerning a particular aspect led to divergent ratings within individual frameworks.

Drawing on our analysis of the three frameworks in terms of their content and ratings on three specific lessons, we conclude that the process of analysis helped to provide more detailed and nuanced understandings of instructional quality. However, the time required to undertake such analysis is considerable, and the opportunity to observe the same lessons using different frameworks is rare. Further, the complexity introduced in trying to compare different frameworks, as well as the differences in the use of terminology across frameworks, constitute a limitation of this type of endeavour.

Following our analysis of the three frameworks considered in this article, we interrogated the relative benefits and challenges of this approach versus work that synthesises existing frameworks on a larger scale, using the MAIN-Teach model as a reference point. We found that, on one hand, a broader synthesis such as the MAIN-Teach model has the potential to provide a common language for discussing instructional quality and may facilitate cross-cultural understandings of how different instructional aspects relate to one another, and allows for considerations of complexity with regard to aspects of instruction that are sometimes conceptualised as generic and sometimes as content-specific. On the other hand, syntheses such as the MAIN-Teach model are less able to provide fine-grained detail compared to individual frameworks, within their context- and culturally-bound conceptualisations of instructional quality. As a result, we suggest that individual frameworks and comparisons of frameworks may still be more useful for some research purposes and applications of observation to professional development and accountability when detailed characterisation is a primary consideration.

Reflecting upon the analysis of our three frameworks, as well as the synthesis of existing conceptualisations of instructional quality that produced the MAIN-Teach model (Charalambous & Praetorius, in this issue), it is worth remarking upon how productive collaboration among scholars aimed towards understanding and capturing instructional quality can be fostered in the future. The MAIN-Teach model that we used as a basis for comparison to our more bounded analysis of three frameworks provides one example of the potential of cross-cultural collaborations between scholars to facilitate shared understandings of instructional quality. Our work in this paper focussing on the TEDS-Instruct, UTOP and multi-instrument frameworks illustrates another more bounded collaborative endeavour by scholars from the USA, England and Germany, and demonstrates how detailed analysis of frameworks from these different countries and cultures can shed light on similarities and differences in the conceptualisation of instructional quality and the various aspects of it.

To extend this work, we propose that establishing international networks of scholars who can work together in a sustained way, building upon one-off instances of collaboration and focusing on differences as well as similarities in the conceptualisation of aspects of instruction and their relationships to one another, is a promising way to further the investigation and understanding of instructional quality on a global scale. Although in this paper the focus has been on mathematics instruction, we suggest that such collaborations are likely to be useful beyond this specific subject area. Moreover, we suggest that such future collaborative work could further explore (using larger, cross national samples) how different frameworks and theoretical models can be used to investigate relationships between aspects of instructional quality and student outcomes, both socio-emotional and academic.

Funding

This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

References

Authors (2018). [Peer-reviewed journal article details omitted for blind review]

Authors & Others (2018). [Peer-reviewed journal article details omitted for blind review]

Author & Other (2018). [Peer-reviewed journal article details omitted for blind review]

Blazar, D., Braslow, D., Charalambous, C. & Hill, H. (2017). Attending to general and mathematics-specific dimensions of teaching: Exploring factors across two observation instruments. *Educational assessment, 22*(2), 71-94. DOI: 10.1080/10627197.2017.1309274

Boston, M., Bostic, J., Lesseig, K., & Sherman, M. (2015). A comparison of mathematics classroom observation protocols. *Mathematics teacher educator, 3*(2), 154-175. DOI: 10.5951/mathteaceduc.3.2.0154

Bostic, J., Lesseig, K., Sherman, M. & Boston, M. (2019). Classroom observation and mathematics education research. *Journal of mathematics teacher education* (Online first). DOI: 10.1007/s10857-019-09445-0

Charalambous, C. & Praetorius, A-K. (Under review). Creating a forum for researching teaching and its quality more synergistically. *Studies in Educational Evaluation.*

Charalambous, C. & Praetorius, A-K. (2018a). Studying instructional quality in mathematics through different lenses: in search of common ground (Special Issue). *ZDM Mathematics Education, 50*(3).

Charalambous, C. & Praetorius, A-K. (2018b). Studying mathematics instruction through different lenses: setting the ground for understanding instructional quality more comprehensively. *ZDM Mathematics Education, 50*(3), 355-366. DOI: 10.1007/s11858-018-0914-8

Creemers, B. P. M., & Kyriakides, L. (2008). *The dynamics of educational effectiveness: A contribution to policy, practice, and theory in contemporary schools.* London & New York: Routledge.

Danielson, C. (2013). The Framework for Teaching evaluation instrument, 2013 Edition. Retrieved December 1st, 2019 from https://danielsongroup.org/downloads/2013-framework-teaching-evaluation-instrument

Gill, B., Shoji, M., Coen, T. & Place, K. (2016). *The content, predictive power, and potential bias in five widely used teacher observation instruments* (REL 2017-191). Washington, D.C.: U.S. Department

of Education, Institute of Educational Sciences, National Center for Educational Evaluation and Regional Assistance, Regional Education Laboratory Mid-Atlantic.

Hall, J., Lindorff, A. & Sammons, P. (2016). *Evaluation of the impact and implementation of Inspire Maths in Year 1 classrooms in England.* Oxford: Oxford University Department of Education & Oxford University Press. DOI: 10.13140/RG.2.2.23042.07369

Hattie, J. (2008). *Visible learning: A synthesis of over 800 meta-analyses relating to achievement.* Abingdon: Routledge.

Hiebert, J., & Grouws, D. A. (2007). The effects of classroom mathematics teaching on students' learning. In Lester, F.K. (Ed.), *Second handbook of research on mathematics teaching and learning* (pp. 371–404). Charlotte: Information Age.

Hilberg, R.S., Waxman, H.C. & Tharp, R.G. (2004). Introduction: Purposes and perspectives on classroom observation research. In Waxman, H.C., Tharp, R.G. & Hilberg, R.S. (Eds.), *Observational research in US classrooms: New approaches for understanding cultural and linguistic diversity*. Cambridge: Cambridge University Press. DOI: 10.1017 /CBO9780511616419.001

Hill, H. C., Blunk, M. L., Charalambous, C. Y., Lewis, J. M., Phelps, G. C., Sleep, L., & Ball, D. L. (2008). Mathematical knowledge for teaching and the mathematical quality of instruction: An exploratory study. *Cognition and Instruction, 26*(4), 430–511. DOI: 10.1080/07370000802177235

Hill, H. C. & Chin, M. (2018). Connections between teachers' knowledge of students, instruction, and achievement outcomes. *American Journal of Education*, 55(5), 1076-1112.

Hill, H. C., Rowan, B., & Ball, D. L. (2005). Effects of teachers' mathematical knowledge for teaching on student achievement. *American Educational Research Journal, 42*(2), 371–406. DOI: 10.3102/00028312042002371

Ingram, J., Sammons, P. & Lindorff, A. (2018). *Observing effective mathematics teaching: A review of the literature*. Reading: Education Development Trust. Retrieved October 1st, 2019 from https://www.educationdevelopmenttrust.com/EducationDevelopmentTrust/files/84/843fd2fb-3a64-4045-83c1-6ad18a746430.pdf

Jentsch, A., Schlesinger, L., Heinrichs, H., Kaiser, G., König, J., & Blömeke, S. (2020). Erfassung der fachspezifischen Qualität von Mathematikunterricht: Faktorenstruktur und Zusammenhänge zur professionellen Kompetenz von Mathematiklehrpersonen. *Journal für Mathematik-Didaktik*. DOI: 10.1007/s13138-020-00168-x

Kaiser, G., Blömeke, S., König, J., Busse, A., Döhrmann, M. & Hoth, J. (2017). Professional competencies of (prospective) mathematics teachers – cognitive versus situated approaches. *Educational Studies in Mathematics*, 94(2), 161-182, 183-184

Kane, T. & Staiger, D. (2012). *Gathering feedback for teaching: Combining high-quality observations with student surveys and achievement gains.* Seattle: Bill & Melinda Gates Foundation. Retrieved October 3rd, 2019 from https://files.eric.ed.gov/fulltext/ED540960.pdf

Kersting, N. B., Givvin, K. B., Thompson, B. J., Santagata, R., & Stigler, J. W. (2012). Measuring usable knowledge: Teachers' analyses of mathematics classroom videos predict teaching quality and student learning. *American Educational Research Journal*, 49(3), 568–589. doi:10.3102/0002831212437853.

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

Klieme, E., Pauli, C., & Reusser, K. (2009). The Pythagoras study. In T. Janik & T. Seidel (Eds.), *The power of video studies in investigating teaching and learning in the classroom* (pp. 137–160). Münster: Waxmann.

Klieme, E., & Rakoczy, K. (2008). Empirische Unterrichtsforschung und Fachdidaktik. Outcome-orientierte Messung und Prozessqualität des Unterrichts. Zeitschrift für Pädagogik, 54, 222–237. urn:nbn:de:0111-opus-43488

Kounin, J. S. (1970). *Discipline and group management in classrooms.* New York: Holt, Rinehart and Winston.

Learning Mathematics for Teaching (2011). Measuring the mathematical quality of instruction. *Journal of Mathematics Teacher Education*, 14, 25–47. DOI: 10.1007/s10857-010-9140-1

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

Lindorff, A., Hall, J. & Sammons, P. (2019). Investigating a Singapore-based mathematics textbook and teaching approach in classrooms in England. *Frontiers in education, 4*(37), 1-21. DOI: 10.3389/feduc.2019.00037

Lipowsky, F., Rakoczy, K., Pauli, C., Drollinger-Vetter, B., Klieme, E. & Reusser, K. (2009). Quality of geometry instruction and its short-term impact on students' understanding of the Pythagorean Theorem. *Learning and instruction, 19*(6), 527-537. DOI: 10.1016/j.learninstruc.2008.11.001

Marder, M., Walkington, C., Abraham, L., Allen, K., Arora, P., Daniels, M., Dickinson, G., Ekberg, D., Gordon, J., Ihorn, S. & Walker, M. (2010). *The UTeach Observation Protocol (UTOP) Training Guide.* UTeach Natural Sciences, University of Texas Austin.

Mitchell, R., Charalambous, C. Y., & Hill, C. H. (2014). Examining the task and knowledge demands needed to teach with representaions. *Journal of Mathematics Teacher Education, 17*(1), 37–60. DOI: 10.1007/s10857-013-9253-4

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. DOI: 10.1076/0924-3453(200009)11:3;1-G;FT273

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. DOI: 10.1080/09243453.2014.885451

Ornstein, A.C. (1991). Teacher effectiveness research: Theoretical considerations. In Waxman, H.C. & Walberg, H.J. (Eds.), *Effective teaching (*pp. 63-80). Berkeley, CA: McCutchan.

Pianta, R., & Hamre, B. K. (2009). Conceptualization, measurement, and improvement of classroom processes: Standardized observation can leverage capacity. *Educational Researcher, 38*(2), 109–119. DOI: 10.3102/0013189X09332374

Popping, R. (2010). *Introduction to interrater agreement for nominal data.* Cham, Switzerland: Springer.

Praetorius, A-K., Charalambous, C. (2018). Classroom observation frameworks for studying instructional quality: looking back and looking forward. *ZDM Mathematics Education, 50*(3), 535-553. DOI: 10.1007/s11858-018-0946-0

Praetorius, A.-K., Pauli, C., Reusser, K., Rakoczy, K., & Klieme, E. (2014). One lesson is all you need? Stability of instructional quality across lessons. *Learning and Instruction*, 31, 2–12. DOI: 10.1016/j.learninstruc.2013.12.002

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

Sammons, P. & Davis, S. (2017). Mixed methods approaches and their application in educational research. In Wyse, D., Selwyn, N. & Smith, E., *The BERA/SAGE Handbook of educational research* (Vol. 2, pp. 477-504). London: SAGE. DOI: 10.4135/9781473983953.n24

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. DOI: 10.1108/JPCC-09-2015-0005

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

Schlesinger, L. & Jentsch, A. (2016). Theoretical and methodological challenges in measuring instructional quality in mathematics education using classroom observations. *ZDM Mathematics Education, 48*(1-2), 29-40. DOI: 10.1007/s11858-016-0765-0

Seidel, T., & Shavelson, R. J. (2007). Teaching effectiveness research in the past decade: The role of theory and research design in disentangling meta-analysis results. *Review of Educational Research, 77*(4), 454–499. DOI: 10.3102/00346 54307 310317

Smith, M. S., & Stein, M. K. (2011). *5 practices for orchestrating productive mathematics discussions.* Reston: NCTM.

Strong, M. (2011). *The highly qualified teacher: What is teacher quality and how do we measure it?* New York: Teachers College Press.

Teddlie, C., Creemers, B. M. P., Kyriakides, L., Muijs, D., & Fen, Y. (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. DOI: 10.1080/13803610600874067

UTeach Institute (2014a). *UTeach elements of success.* Austin, TX: University of Texas. Retrieved on December 2nd, 2019 from https://institute.uteach.utexas.edu/sites/default/files/UTeach-Elements-of-Success.pdf

UTeach Institute (2014b). *UTeach observation protocol.* Austin, TX: University of Texas. Retrieved on December 2nd, 2019 from https://utop.uteach.utexas.edu/node/3

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. DOI: 10.1080/00131880701369651

van de Grift, W. (2014). Measuring teaching quality in several European countries. *School Effectiveness and School Improvement, 25*(3), 295–311. DOI: 10.1080/09243453.2013.794845

Walkington, C., Marder, M., & Adams, B. (in press). Developing the UTOP: A flexible STEM observation instrument based on UTeach principles. To appear in Goodell, J.E. & Koç, S. (Eds.), *Preparing STEM Teachers: A Replication Model*. Charlotte: Information Age Publishing.

Supplementary Files