SURVEY PAPER

Studying mathematics instruction through different lenses: setting the ground for understanding instructional quality more comprehensively

Charalambos Y. Charalambous[1] · Anna-Katharina Praetorius[2]

Abstract
Researchers from different fields have developed different observational instruments to capture instructional quality with a focus on generic versus content-specific dimensions or a combination of both. As this work is fast accumulating, the need to explore synergies and complementarities among existing work on instruction and its quality becomes imperative, given the complexity of instruction and the increasing realization that different frameworks illuminate certain instructional aspects but leave others less visible. This special issue makes a step toward exploring such synergies and complementarities, drawing on the analysis of the same 3 elementary-school lessons by 11 groups using 12 different frameworks. The purpose of the current paper is to provide an up-to-date overview of prior attempts made to work at the intersection of different observational frameworks. The paper also serves as the reference point for the other papers included in the special issue, by defining the goals and research questions driving the explorations presented in each paper, outlining the criteria for selecting the frameworks included in the special issue, describing the sampling approaches for the selected lessons, presenting the content of these lessons, and providing an overview of the structure of each paper.

1 Introduction

Several scholars (e.g., Cohen 2011; Doyle 1986; Lampert 2001; Leinhardt 1993; Wragg 2012) have repeatedly emphasized that instruction is a particularly complex phenomenon involving innumerable teacher–student and student–student interactions around the content. It is not surprising then that over the past two decades several frameworks and associated observational instruments with different foci have been developed to study and analyze instruction (see for example Schlesinger and Jentsch 2016, for a compilation of frameworks on mathematics instruction). These, however, have

only rarely been connected to each other, although this seems highly promising for obtaining a comprehensive picture of teaching.

In the current special issue, our general goal is to bring together researchers from different fields who have conceptualized instructional quality in different ways and have developed or used different observational instruments to capture it. For enabling such a comparison, all research groups analyze the same three lessons from their specific perspective(s). Building on these analyses, in the concluding paper of the special issue we aim at building a more comprehensive picture of the complex phenomenon of instruction by identifying complementarities and building synergies among different frameworks. Through that comparison we aim to start developing what Grossman and McDonald (2008) called a common grammar and lexicon of teaching.

In the following, we first give an initial overview of existing frameworks, organizing them along a continuum from more generic to more content-specific. Second, we elaborate on the advantages of bringing together different frameworks occupying different junctures of this continuum. Third, we

✉ Charalambos Y. Charalambous
cycharal@ucy.ac.cy

[1] Department of Education, University of Cyprus, Theophanides Building, Rm 508, 11-13 Dramas Str., 1077 Nicosia, Cyprus

[2] Department of Education, University of Zurich, Freiestrasse 36, 8032 Zurich, Switzerland

Generic frameworks
- Developed and used to study instructional quality across different subject matter
- Attend to general demands of teaching

Content-specific frameworks
- Developed and used to study instructional quality in a single subject matter (e.g., mathematics)
- Informed by subject-specific demands of teaching within a particular discipline
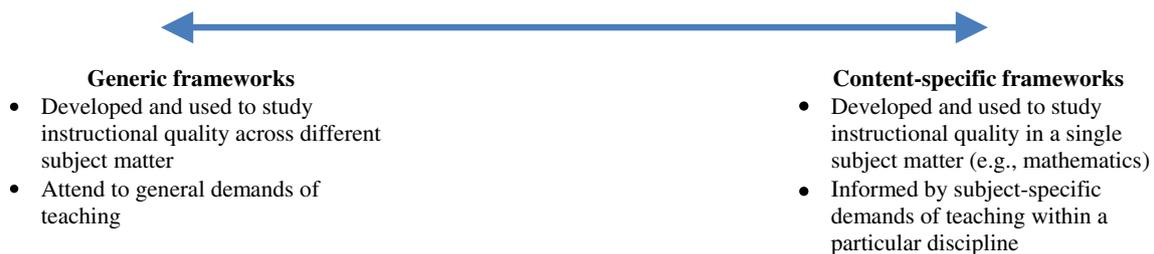
Fig. 1 A continuum of classifying frameworks and observational instruments

summarize prior attempts that concurrently built on different frameworks. Next, after explaining how this special issue builds on and extends prior work aimed at integrating different frameworks in studying instructional quality, we present its aim based on the desiderata identified. We then outline the criteria for selecting the frameworks included in the special issue, describe our approach for sampling lessons, and present the content of the lessons analyzed in each paper. We conclude with an overview of the structure shared by all individual papers.

2 Existing frameworks and their different foci

Existing frameworks and observational instruments on instructional quality have different foci with respect to their subject-specificity. These differences can be considered to form a continuum (see Fig. 1). Situated towards one end are frameworks that have been developed and used to capture instructional quality in different subject matter. They attend to general instructional aspects such as appropriately managing the instructional time, giving timely and relevant feedback to students, and structuring information (cf., Bell et al. 2012; Kane et al. 2011; Muijs et al. 2014). These frameworks as a rule do not consider specific manifestations of instruction that pertain to teaching particular subject matter nor are they informed by the demands that teaching within a particular discipline imposes on teachers. For example, such frameworks might attend to the form of teacher-student interactions and communication, without attending to the content of such interactions. A typical example is the well-known IRE (initiate-response-evaluate) type of teacher-student chain of interactions which attends to who initiates an interaction and how this interaction unfolds, without attending to the correctness of the teacher's or students' contributions. Situated towards the other end are frameworks that have been developed and used to capture instructional quality in specific subject matter. For instance, in mathematics—which is the focus of this special issue—instead of capturing teacher-student communication in general, these frameworks attend to these interactions through a content-related lens,

examining aspects such as the precision and accuracy in communication and the appropriateness of the mathematical language and notations used. As such, these frameworks are explicitly or implicitly informed by the subject-specific demands that teaching within a given discipline imposes on teachers.

The past two decades have seen the development of several frameworks that can be considered to be geared toward the generic end. These frameworks include, for example, the Classroom Assessment Scoring System (CLASS, Pianta and Hamre 2009), the Dynamic Model of Educational Effectiveness (DMEE, Creemers and Kyriakides 2008), the Framework for Teaching (FfT, Danielson 2013), the Rapid Assessment of Teacher Effectiveness instrument (RATE, Strong 2011), and the Three Basic Dimensions framework (hereafter TBD, Klieme et al. 2009). Being convinced that there is value in comparing and contrasting instructional quality across countries, some scholars have also attempted to develop generic frameworks that capture instructional aspects assumed to 'travel' across and be applicable to different contexts and educational systems (e.g., the International System of Teacher Observation and Feedback, ISTOF, Teddlie et al. 2006). It is important to mention that some of these generic frameworks are not located at the very end of the continuum as they may also contain aspects that are informed by the content to be taught (e.g., the content-focused Instructional Dialogue in CLASS).

Other frameworks, largely developed during the past decade or so, are more geared toward the content-specific end, since they have been developed to capture instructional aspects that are thought to be unique to teaching specific content areas or have a particular functioning and specialized manifestations when occurring in the teaching of these areas. In mathematics, such aspects include, for example, using and connecting representations (e.g., Ball 1992; Cuoco and Curcio 2001; Mitchell et al. 2014), engaging students in offering or considering multiple solution approaches (Smith and Stein 2011; Tsamir et al. 2010), employing appropriate and precise mathematical language when presenting the content (Hill et al. 2008), and reasoning abstractly, by either working on or manipulating abstract symbols (decontextualizing), or giving

meaning to such symbols (contextualizing) (Common Core State Standards Initiative 2014). Over the past decade, several such frameworks and observational tools in mathematics have been developed. These include, for instance, the Elementary Mathematics Classroom Observation Form (Thompson and Davis 2014), the Instructional Quality Assessment (IQA, Matsumura et al. 2008), the Mathematical Quality of Instruction (MQI, Learning Mathematics for Teaching 2011), and the Mathematics-Scan (M-Scan, Walkowiak et al. 2014). Explicit international endeavors for capturing mathematics-specific aspects of instructional quality also exist, one of the most famous being the TIMSS video study (Hiebert et al. 2003). Some frameworks, such as the Reformed Teaching Observation Protocol (Sawada et al. 2002), and The Inside the Classroom Observation and Analytic Protocol (Horizon Research 2000) pertain to closely related subject matter (e.g., mathematics and science). As such, they can be considered to be situated close to the content-specific end of the continuum, without, however, occupying the extreme endpoint.

Other frameworks are situated in the middle of the continuum. Some of these frameworks have been developed by capitalizing on both generic and content-specific instructional aspects from already existing frameworks. For example, the TEDS-Instruct framework (Schlesinger et al. 2018) incorporates generic and content-specific instructional aspects by using the TBD of Klieme and colleagues' work (2009)—classroom management, student support, and cognitive activation—but adding more content-specific instructional aspects gleaned from reviewing mathematics-specific instructional frameworks (e.g., MQI and IQA). Similarly, the UTeach Observation Protocol (UTOP, Marder and Walkington 2014) attends to content-specific instructional aspects, but complements them with more generic instructional aspects taken from existing generic literature (cf. Muijs et al. 2014). Other frameworks—e.g., The Teaching for Robust Understanding (TRU, Schoenfeld 2013) framework—have been developed with the intention of capturing instructional quality across different subject matter, but later on, have incorporated content-specific dimensions and have included content-related manifestations of more generic teaching practices. Sammons and colleagues (Sammons et al. 2016) combine the content-specific Mathematics Enhancing Classroom Observation Recording System (MECORS, Schaffer et al. 1998) with the generic Quality of Teaching (QoT, Van de Grift 2007) framework and additionally enhance these frameworks with unstructured observations in an attempt to better understand and describe the complexities inherent in instruction. Although these frameworks differ in how they were generated and in the extent to and the manner in which they capture instructional quality with respect to more generic and more content-specific aspects, for ease of communication, we collectively call them hybrid models.

3 Advantages of bringing together different frameworks

Different frameworks capture different aspects of instruction; in fact, as is evident in this special issue, even frameworks belonging to the same category (generic or content-specific) do not necessarily capture the same instructional aspects—thus highlighting the importance of drawing upon different frameworks when attempting to understand teacher and student interactions around the content. One of the most telling pieces of evidence regarding this fact is based on the analysis of a short instructional episode of a famous US mathematics educator, Deborah Loewenberg Ball. Known as the "Shea numbers" piece, this snippet of instruction has been analyzed through different lenses (cf., Schoenfeld and Pateman 2008). Adopting a perspective that focuses on the interactions of the teacher and the students around the mathematical content, Ball et al. (2008) have positively delineated the work carried out in this segment by focusing on the mathematical argumentation and reasoning in which students engaged during the lesson. By attending more to issues of equity and participation, Posner (2008) pointed to certain ways in which instructional quality in this segment could have been improved. Furthermore, Schoenfeld (2008) analyzed the teacher's dilemmas and her decisions during different junctures of instruction, showing Ball to be extremely flexible and responsive both to her students and the content and to attend to multiple conflicting goals at the same time. Adopting a different lens and focusing on classroom discourse, Horn (2008) zoomed in on students' argumentation and documented how the classroom environment nurtured and supported mathematical disagreements that were both intellectually productive and socially acceptable.

Hence, it seems reasonable to argue that none of the existing frameworks alone has the capacity to capture instruction in its entirety (an argument we revisit in the concluding paper of this special issue). Therefore, combining different frameworks has several advantages. First, it enables us to identify overlaps, both in conceptualization and wording across frameworks, and thus find common grounds. Second, it allows us to see where frameworks complement each other, each uniquely contributing to the insight we can gain from analyzing instruction. Third, it helps us start building a common lexicon of instruction by identifying common terms used across different frameworks to describe similar or identical instructional aspects, dissimilar terms used to denote the same or similar instructional aspects, as well as similar or identical terms that denote different aspects of instruction. It can also help us start developing a common grammar of instruction, by identifying instructional aspects that are prioritized by

different frameworks (which could be thought to function as key components of instruction) and interrelations between and across these aspects (which could be taken to denote the structure of instruction).

4 Prior attempts to work at the nexus of different frameworks

Acknowledging the importance of working at the nexus of different frameworks, scholars over the past few years have started combining different frameworks in their work. Our review of the literature (for an overview of the search terms and search engines we used, see Appendix A) suggests that these attempts have followed different, often overlapping, directions: (a) focusing on the dimensionality of instructional quality, (b) comparing different instruments to inform selection decisions, (c) comparing the predictive validity of different instruments, and (d) combining different frameworks to better capture the complex phenomenon of instruction. Although this review is meant to be neither comprehensive nor exhaustive, it is indicative of the progress made in this area. In the following, we present an overview of the studies conducted following each of the four directions listed above.

4.1 Focusing on the dimensionality of instructional quality

The first direction was geared toward understanding the dimensionality of instructional quality. The large-scale study Measures of Effective Teaching (MET, Kane and Staiger 2012), which employed two frameworks situated toward the more generic end of the continuum discussed above (CLASS and FfT) and three frameworks focusing on content-specific aspects of instruction (MQI and UTOP for mathematics as well as the Protocol for Language Arts Teaching Observations, PLATO, for language arts) to analyze approximately 7500 video-recorded lessons from about 1330 teachers, provides some insights into this issue. The high but not perfect correlations of instructional quality between instruments attending more to generic ($\rho_{\text{CLASS–FfT}} = 0.88$) or content-specific aspects of instruction (for mathematics $\rho_{\text{MQI–UTOP}} = 0.85$) and the somewhat lower correlations between instruments falling into these two different categories (ranging from $\rho = 0.67$ to $\rho = 0.74$) suggested that, although sharing a considerable amount of common variance, each instrument also measures some unique instructional aspects. This result therefore supports the multidimensional nature of instruction and underscores the need to use different frameworks and observational instruments—both generic and content-specific—to fully capture instructional quality.

Drawing on smaller samples, other studies have also utilized different observational frameworks, aiming to more directly study the dimensionality of instructional quality. Using two instruments attending to generic instructional aspects (CLASS and FfT) and two instruments focusing on content-specific aspects (MQI and PLATO) to code data collected from 458 middle-school teachers who participated in the *Understanding Teacher Quality* study, a recent study (Lockwood et al. 2015) generated empirical evidence on two overarching latent constructs for both mathematics and language arts: the first was related to classroom management and the second pertained to what they called instructional quality and student support. These results were partly corroborated in another study that utilized one generic (CLASS) and one content-specific instrument (MQI) to analyze over 2000 videotaped mathematics lessons from approximately 400 upper elementary school teachers (Blazar et al. 2017). This latter study again yielded a classroom management factor and two factors pertaining to teachers' instructional practices (one related to "ambitious instruction" as described in Cohen 2011, and the other capturing emotional and instructional support); additionally, it pointed to a pure content-specific factor pertaining to mathematical errors and imprecision during instruction. Collectively then, these studies empirically corroborate the multidimensional nature of instruction. Because single frameworks obviously capture (slightly) different instructional aspects, taking them together has more potential to yield a more comprehensive delineation of the phenomenon at hand. As a consequence of this increasing realization, scholars have also invested in using advanced statistical models to determine the best ways of combining measures and assigning weights to them to better capture instructional quality (see, for example, Cantrell and Kane 2013; Mihaly et al. 2013). However, because most of these studies are quantitative in nature, they do not lend themselves to identifying the affordances and limitations of different frameworks in capturing instructional quality.

4.2 Comparing different frameworks as a means to inform selection decisions

A second direction pertains to a group of studies aimed at comparing different frameworks and observational instruments to select the most appropriate for certain aims. These works followed different approaches, ranging from theoretical analyses of the frameworks and instruments under consideration—with an emphasis on aspects such as their theoretical background, foci, and psychometric properties (e.g., Boston et al. 2015; Kilday and Kinzie 2009)—to eliciting the perceptions of their intended end-users (e.g., Henry et al. 2007; Martin-Raugh et al. 2016) on issues such as their ease of use, perceived accuracy, levels of rater agreement, and perceived advantages and disadvantages. A key intention of

these studies was to inform selection decisions for certain classroom observation purposes (e.g., to identify high-quality or reform instruction) or for certain grade levels. Hence, these studies largely juxtaposed the selected instruments, offering guidelines as to when each should be selected over the other(s). As such, despite their affordances in informing selection decisions, these studies mostly seemed rather to be considering the available frameworks and associated protocols in parallel, instead of treating them as different lenses that can help better explore instructional quality by functioning complementarily to each other.

4.3 Comparing the predictive validity of different frameworks and associated observational instruments

The third direction relates to studies in which different frameworks and observational tools have been used to better explain student learning. In addition to the MET study reviewed above, other studies have followed this line of thinking. For example, at an international level, a recent secondary analysis of TIMSS Grade-8 data (Charalambous & Kyriakides 2017) showed that more variance in student performance can be explained when using variables from both a generic (DMEE) and a content-specific (MQI) instrument, as opposed to employing just one of them. Interestingly, the study also showed that the contribution of generic and content-specific variables in explaining student performance varies across different countries, thus calling for more qualitative studies that unravel the mechanisms producing these differences. Smaller-scale studies have also documented the increased predictive validity of using multiple frameworks. For instance, a recent study used two instruments that are either geared toward the more generic (CLASS) or the more content-specific end (MQI) of the continuum discussed above, in order to study instructional quality in upper elementary grades (Blazar and Kraft 2017). This study showed that certain types of student learning are better explained by instructional aspects that are more proximal to these types of learning. Specifically, the dimensions of teacher emotional support and classroom organization (from CLASS) were more predictive of students' self-efficacy in mathematics and students' self-reports of their own behavior in class respectively, compared to other instructional aspects. In contrast, the dimension of mathematical errors in presenting content (from MQI) was negatively related to students' mathematics performance and self-efficacy in mathematics. This complementarity of frameworks in explaining student learning was also illustrated in another study (Doabler et al. 2015) that capitalized on three different frameworks (Classroom Observation of Student–Teacher Interactions-Mathematics, Quality of Classroom Instruction, and Ratings of Classroom Management and Instructional Support) as multiple lenses

for exploring instruction. Focusing on kindergarten, this study showed that the first two instruments complemented each other in better capturing quantitative and qualitative aspects of instructional quality as well as in explaining different aspects of student performance, as captured by different types of student tests. Because they spanned different grades (from kindergarten to middle school), collectively these studies empirically corroborate the importance of considering multiple frameworks for understanding and explaining student learning at different levels of schooling.

4.4 Combining different frameworks to better understand instruction

Whereas the studies of the previous group focus first and foremost on a comparison of frameworks with respect to their predictive validity, the studies of the fourth direction attempt to delve deeper into the degree to and the ways in which combinations of different frameworks can help in better capturing instructional complexity. These studies are thus more pertinent to the perspective and the main overarching goal of this special issue; we therefore describe them in more detail in the following.

Unlike the other studies in this direction, which have compared different frameworks using empirical data, the first study, conducted by Schlesinger and Jentsch (2016), aimed at finding common ground among 11 different mathematics-focused frameworks by comparing them in a theoretical sense. Acknowledging the difficulties inherent in identifying commonalities among them due to different wording and level of detail of the instructional aspects they covered, these scholars came up with two overarching groups, namely a group pertaining to the more mathematical aspects of instruction (e.g., mathematical language, mathematical errors, and mathematical concepts) and another group incorporating also pedagogical aspects (e.g., students' discussion, or connecting classroom practice to mathematics).

Aiming better to capture and understand the quality of teacher-student instructional interactions in the classroom, the second study (McGuire et al. 2016) analyzed teacher-student interactions across ten videotaped observations drawn from five different prekindergarten classrooms delivering the same mathematics curriculum. Two classroom observation instruments, one more generic (CLASS) and one more content-specific (the Classroom Observation of Early Mathematics–Environment and Teaching, COEMET) instrument, were used toward this end. The study showed that although the generic instrument could delineate instructional quality in broad strokes, the content-specific instrument provided finer grained details of teachers' practice.

Seeking better to understand instruction, and in particular to examine what makes high quality mathematics instruction distinct from high quality teaching in general, the third

study (Pinter 2013) also used a generic (CLASS) and a content-specific (M-Scan) instrument to investigate patterns in fourth-grade teachers' implementation of Standards-based mathematical practices. A purposive sample of ten teachers exhibiting strong instructional quality as measured by CLASS, but different levels in their implementation of mathematics-related practices, as measured by M-Scan, was selected. By analyzing 30 lessons (three from each teacher), the study showed notable differences across the two groups of teachers in several instructional aspects (e.g., lesson structure, focus on conceptual learning, purposeful error use, use of problems utilizing real world contexts, and employment of real-time formative assessment), suggesting that using just one of the instruments alone does not suffice for capturing instructional quality adequately.

The next two studies, both focusing on secondary teachers, were concerned with issues of teacher education. The first (Dubinski et al. 2016) utilized three observation instruments (a teacher-, a student-, and an overall classroom-observation instrument, adapted from existing instruments developed in the 1980s and 1990s) as a means of developing deeper insights into potential differences in the practice of pre-service and in-service secondary school teachers in Grades 8–12. The study pointed to differences both within and across these two groups. More importantly, however, it documented that each of the three instruments employed revealed a different perspective of the classroom procedures. Combined, these three perspectives provided a more comprehensive picture of instruction, which would not have been possible by using any single instrument alone. Therefore, the authors concluded by emphasizing the merits of this complementarity in lenses, suggesting that "they supplied a rich, multidimensional conceptualization of the student–teacher dynamics" which helped them to delve deeper into the practice of these two groups and "discern differences in pedagogy and classroom environment that would not have been evident via other data collection methods" (p. 104). Also focusing on secondary teachers, the second study of this pair (Booker 2014) followed 62 beginning middle-school teachers over 3 years. Using one instrument geared more toward the generic end (CLASS) and one content-specific instrument (IQA), this longitudinal study aimed at exploring whether the use of two different observational instruments would better capture any potential changes in these teachers' practice. Indeed, hierarchical linear growth modeling suggested that using only a single instrument would have yielded a partial picture in teachers' growth and learning over time: whereas these teachers exhibited notable improvements in generic aspects of their work (e.g., in their social/emotional practices and their cross-subject instructional practices), even at the end of the study, they did not progress in the mathematics-related aspects of their work (e.g., connecting mathematical ideas, and pressing students to support

their contributions with evidence and/or mathematical reasoning). Such partial delineations of instructional growth, as the author argued, would fail to provide teachers with the types of scaffold and feedback they need for continuous improvement.

Utilizing some aspects from MQI and the Reformed Teaching Observation Protocol (RTOP), as well as the TRU framework as multiple lenses for exploring the quality of the mathematics offered in College Algebra lessons, the last study (Gibbons 2016) examined the lessons taught by four instructors, aiming to evaluate these instruments for their ability to capture variation in instructional quality in mathematics. The author concluded that, whereas none of the instruments was entirely successful in capturing variation in mathematics in these lessons, different instruments exhibited certain strengths and limitations. For example, the RTOP prompted considerations about the qualities of the mathematics in the lessons observed, but did not help in attending to the variation present during instruction. Similarly, the MQI explicitly captured the errors that occurred during instruction, but did not attend to the quality of the presentation of procedures featured in these lessons. Likewise, the TRU lent itself to capturing the accuracy in the presentation of the content but, like the other two instruments, placed emphasis on conceptual explanations and sense-making at the expense of attending to the explicit presentation of procedures often featured in the lessons.

In summary, spanning different schooling levels from kindergarten to tertiary education, these six studies are telling of the merits of using different frameworks as multiple lenses for observing instruction—with the key underlying idea being that these frameworks complete rather than compete with each other. In this respect, the different classroom observation frameworks are viewed similarly to the affordance provided by a multi-lens microscope, which provides a more nuanced delineation of its object of observation. It is exactly in this sense that we see the frameworks presented in the papers of this special issue: not in the manner of competitive horse-racing, but rather as frameworks with specific strengths and limitations, which can therefore complement each other in helping us better understand instruction and its complexities.

5 Building on and extending prior work: the aim of the current special issue

The papers presented in this special issue build on and extend the line of research initiated by the six aforementioned studies, in three important ways. First, unlike the preceding studies that utilized at most three frameworks, in this special issue we examine instruction through a large number of frameworks (see Sect. 6). Employing many frameworks

and associated classroom observation protocols, was envisioned to generate richer opportunities for exploring synergies and complementarities across frameworks and observation instruments. Second, in the studies reviewed above, the frameworks and associated protocols were not necessarily utilized by their developers or members of the developing team. By addressing this limitation in 9 of the 11 papers that follow,[1] in this special issue we envisioned that the framework developers would be in a better position to discuss their respective frameworks. We acknowledge, however, that identifying and discussing limitations of these frameworks and outlining areas for improvement might be harder for the framework developers; yet, because the framework developers might not have systematically engaged in such reflective discussion in their previous work, we see this aspect as one of the novel features of each individual paper. Third, each individual paper describes the framework under consideration in depth with respect to underlying theoretical assumptions, operationalization, reliability, and validity. We are convinced that it is essential to illuminate the theoretical origins of different frameworks, and through that to understand commonalities and differences among them. At the same time, no substantive discussions can be undertaken about potential complementarities among these instruments, unless we have sufficient information on the reliability and validity of the scores yielded from such instruments—something that is also addressed in the individual papers. In this way, the special issue can be taken to provide an extensive and updated overview of a collection of current frameworks utilized to study instructional quality in mathematics. Finally, because each individual paper analyzes the same three lessons, the special issue provides a platform for searching for synergies and complementarities among these frameworks in the ways they capture instructional quality.

By bringing together researchers from different fields—mathematics education, educational effectiveness research, and psychology—who have conceptualized instructional quality in different ways and have developed or used different associated observational instruments to capture instructional quality, in this special issue we aim at exploring synergies and complementarities among the existing works on instruction and its quality. This juxtaposition is anticipated to help advance attempts better to capture and understand instructional quality in general, and instructional quality in the field of mathematics in particular—and through that to develop better insights into how instruction can contribute to student learning. Towards this end, we first consider each framework individually and ask:

1. What is the theoretical rationale behind each framework of instructional quality?
2. How does each framework operationalize the dimensions of instructional quality it focuses on?
3. What is the evidence for each framework with respect to its reliability and validity?

By using three lessons as reference points to enable an empirical comparison of the capacity of different frameworks to capture instructional quality (see Sect. 7), more critically we ask the following questions:

4. What are the strengths and limitations of the different frameworks in terms of capturing instructional quality?
5. What do these different conceptualizations/frameworks have in common and in what ways might they differ?
6. In what ways might they function complementarily to each other?

For answering our research questions, we selected a broad range of frameworks (see Sect. 6). Papers focusing on each single framework help address Research Questions 1–4. In the concluding paper of this special issue, all research questions are addressed for the purpose of comparing the frameworks and discussing their capacity in analyzing instruction.

6 Selection criteria for the frameworks

We address the research questions listed above by bringing together research groups of scholars who have developed and/or utilized various frameworks. The frameworks have been selected based on a set of criteria. Firstly, we included frameworks that are situated in different junctures of the continuum shown in Fig. 1: frameworks geared toward the more generic end, frameworks attending more to content-specific aspects, and frameworks that can be characterized as hybrids because they incorporate both generic and content-specific instructional aspects, or they have been developed by capitalizing on generic and content-specific frameworks. Secondly, to enable not only a comparison among these categories but also to attend to differences between frameworks within each category, we included four frameworks geared toward the more generic end (i.e., CLASS, DMEE, ISTOF, and the TBD), three frameworks situated closer to the content-specific end (i.e., IQA, MQI, and M-Scan), and four frameworks/approaches that can be considered hybrid, since they integrate generic and content-specific aspects (i.e., TRU, UTOP, TEDS-Instruct, and the combined work of MECORS with QoT). Thirdly, we avoided sampling

---

[1] Lindorff and Sammons (this issue) have not developed the frameworks utilized in their work but combine existing frameworks to better understand instructional quality. Similarly, Berlin and Cohen (this issue) are not amongst the original developers of the CLASS instrument.

frameworks that have been developed to capture specific topic areas in mathematics (e.g., Litke's 2015 framework on algebra) or specific instructional aspects (for example, frameworks focusing on teachers' work to enhance student self-regulated learning, such as that of Dignath-van Ewijk et al. 2013, or metacognition, such as that of Nowińska and Praetorius 2017), since our intention was to study instructional quality in a broader sense, without limiting our attention to particular topics. Fourthly, we sampled frameworks that have been in use for some years and have been shown to be reliable and valid in prior work (e.g., CLASS, DMEE, and MQI). Fifthly, to maximize variability and enhance the generalizability of our research, we sampled frameworks that have been developed in, and applied to, different educational contexts, and which are informed by different theoretical foundations (e.g., ISTOF). Finally, we included some of the most cited frameworks for measuring instructional quality (e.g., CLASS) but made sure that we also included some of the most recent framework developments (e.g., TRU and TEDS-Instruct).

7 Sampling and description of the three lessons

We decided to have all groups analyze the same three lessons to bring some coherence, and to enable a comparison across frameworks. Apparently, this decision limits the conclusions that can be drawn from the analyses presented in the individual papers of this special issue, as this small number of lessons constrains what can be said in terms of instructional quality. As we know from prior research, selecting one lesson per teacher is not sufficient for evaluating teachers in general (see e.g., Hill et al. 2012; Kane & Staiger 2012; Praetorius et al. 2014). Hence, the analyses that follow should by no means be taken to make generalizations about instructional quality in these teachers' lessons, since, unlike prior studies in which these frameworks have been used, the focus in this special issue is *not* on teachers and the instructional quality in their lessons per se, but rather *on the capacity of the frameworks to capture instructional quality*—and it is exactly in this respect that we see a key added value of both this special issue in general and of each individual paper in particular. By focusing on these three lessons, we envisioned the enabling of more alignment and comparability across the delineations of instruction yielded from each framework; in turn, this was envisaged better to support the search for synergies and complementarities across different frameworks and observational lenses. Thus, for an *initial* search for common ground—as is the aim of the current special issue—we are convinced that the advantages of this reduction outweigh the limitations. We emphasize that this is an initial search because analyzing more lessons and varying

the variables that could contribute to instructional quality (e.g., grades and topics taught) might provide additional opportunities for searching for synergies and complementarities among the frameworks sampled towards this end.

The selected fourth-grade lessons have been drawn from the National Center of Teacher Effectiveness (NCTE) video library at Harvard University, which contains fourth- and fifth-grade lessons taught during the 2010–2011 and 2011–2012 school years.[2] The lessons selected for this special issue have been drawn from a subset of this library that incorporates lessons for which teachers had given extended consent for use in different projects. To provide multiple entry points through which the different projects/groups representing the frameworks outlined above could access and analyze these lessons, two sampling criteria were set. First, the lessons should be taught by teachers differing in their levels of teaching effectiveness (low, mid, and high), as captured by teachers' value added scores on State tests and an NCTE test designed to measure students' performance in cognitively demanding tasks (see more in Blazar et al. 2016). To avoid coloring the analyses presented next, these levels of effectiveness were not communicated to the authors of the papers; and to secure teacher confidentiality and anonymity, this information is not presented in the concluding paper either. Second, the lessons should pertain to different topics from different content standards as defined by the National Council of Teachers of Mathematics (NCTM, 2000); hence, we selected a lesson from measurement and geometry (Mr. Smith), a lesson on pre-algebra (Ms. Young), and a lesson from numbers and operations (Ms. Jones, all pseudonyms). Topic and teaching effectiveness were combined randomly in selecting the three lessons. By purposefully varying the content of the lessons and teachers' level of effectiveness, we attempt to maximize the variation of the phenomenon under consideration. Doing so was envisioned to provide more opportunities for individual frameworks to illustrate their capacity in capturing and studying instruction; this, in turn, was anticipated to facilitate a better investigation of synergies and complementarities across the frameworks under consideration. In what follows, the three lessons are described briefly.

7.1 Mr. Smith's lesson on measuring angles

During the first five and a half minutes of this approximately 40-min lesson on measuring angles, the teacher, Mr. Smith, asks questions to remind students of concepts discussed in previous lessons (e.g., how angles are formed, what is

---

the opening of an acute, an obtuse, a right, and a straight angle). He then introduces the concept of reflex angles and, using Smartboard, projects different angles, asking students to name them, while also giving them some mnemonics to scaffold them in remembering their names.

At about minute 12, the teacher projects pictures of protractors on the board. Before students are asked to measure angles using their protractors, they discuss a multiple-choice problem on estimating the degrees of given angles, which, as he mentions, can be found on State tests. Mr. Smith suggests that they use scratch paper to identify two "easy" angles (the right and the straight angle) which can be handy when solving such problems. By folding the scratch paper into half so that the corner is equipartitioned, he then introduces a third "easy" angle—that of 45°.

At approximately minute 19, the class starts discussing how they can use protractors to measure angles. The teacher emphasizes that the vertex of an angle needs to be placed exactly on the midpoint/hole of the protractor. By asking why protractors typically have two sets of numbers, Mr. Smith draws students' attention to the importance of using the appropriate number scale; he also suggests that they cross-check their measurement by considering the type of the angle measured. He then projects a short video-clip on the Smartboard and narrates the steps for using the protractor to measure angles. Next, two students are invited to the Smartboard and use the protractor to measure an acute and an obtuse angle; another couple of students uses a protractor to make angles of given size.

The class then classifies given angles (presented on the SmartBoard), goes over a couple of multiple-choice items on angles and their classification, and solves a problem on complementary angles. Five minutes before the end of the videotaped lesson, the teacher hands out a worksheet to students and asks them first to work on estimating given angles (by using the "easy" angles discussed before) and then to use their protractors to measure different angles.

7.2 Ms. Young's lesson on doubling and halving factors

This 70-min lesson starts with the teacher, Ms. Young, stating the objectives of the lesson, which are written on a flip-chart paper: investigating how simultaneously doubling and halving factors affects the product and developing strategies for multiplying that involve breaking numbers apart. She then asks students to go over a problem assigned for homework the previous night (how 16×3 and 16×6 are related). A student observes that 6 is equal to the sum of 2 threes, and hence suggests that if you take the product of 16×3 twice, you will get the product of 16×6. The teacher then challenges the class to come up with a word problem or an array that proves what their classmate has suggested: when

doubling one of the factors, the product is also doubled. A student proposes drawing three boxes each containing 16 books and another row of three such boxes. Another student draws two arrays, each including 3×16 boxes. Prompted to propose a word problem, a third student talks about six shelves each containing 16 apples and then taking half of the shelves, which results in half of the apples. The teacher then encourages students to identify relationships among the given numbers.

At about minute 18, the class moves to a second group of multiplication sentences: 15×8 and 30×4. Students figure out that both multiplications have the same product, but the teacher encourages them to use pictures, arrays, cubes, or story problems to prove that the two multiplications are equivalent. Working in groups, students spend about half an hour on this task. The teacher circulates and offers assistance (e.g., doubling one dimension but keeping the area constant; cutting a given shape in half). At times, she also scolds students who seem to be disengaged; at other times, she encourages students to use the appropriate mathematical language (e.g., the products are equivalent, not the same).

At minute 55, the class is brought back together to share their work. Different students share their solution methods. For example, a boy shares a picture (a 30×4 area model) in which he first drew four boxes each including 30 objects and then cut the four boxes into half, thus creating eight boxes each including 15 objects. A girl proves the equivalence by first drawing a 15×8 array, then partitioning it into two 15×4 arrays, and finally moving the one of the resulting arrays beneath the other to get a 30×4 array. Using grid paper, another boy cuts his 15×8 rectangle into half and joins the two resulting 15×4 rectangles to get a longer 30×4 rectangle. In reiterating this student's way of thinking, Ms. Young emphasizes that although the shape of the rectangles has been changed, in both cases the area is kept the same.

Toward the end of the lesson, Ms. Young explains the rationale of doubling and halving the factors of a given multiplication by pointing out that the ultimate purpose of this activity is to get factors that are easier to multiply (e.g., 30×4 is easier to multiply than 15×8 because one of the factors is a multiple of 10). She extends this thought by pointing out that this strategy should be avoided when the factors are odd and cannot be cut into half. She wraps up the lesson by announcing that they will continue working on multiplication strategies.

7.3 Ms. Jones' lesson on multiplying a whole number by a fraction

At the beginning of this 56-min lesson, the teacher, Ms. Jones, tells the class that they are going to continue working on learning three ways to multiply a whole number by a fraction. The first method is to write the whole number as

an improper fraction, then multiply numerators, and finally multiply denominators before simplifying to a mixed number. To remember this method—which has been introduced in the previous lesson—the teacher gives each student a yellow sheet of construction paper and has them partition it into three parts. She then asks them to write in the first third a multiplication sentence including a whole number from 1 to 5 and a proper fraction in fourths. Students are expected to solve their multiplication using the first method modeled by the teacher on the whiteboard while solving $5 \times \frac{3}{4}$ as a sequence of four steps. Students are also given time to work on their multiplication and the teacher circulates in the classroom and provides help; at times, she also brings the class back together, discussing certain steps in this method (e.g., what students should do if the quotient of the numerator over the denominator of their final improper fraction does not yield a remainder).

Half way through the lesson—and once students have had an opportunity to apply the first method—the teacher introduces the second method (i.e., repeated addition), using the same multiplication problem worked on the board before. To help students understand repeated addition (and especially what the whole number represents), Ms. Jones first introduces a couple of real-life scenarios (e.g. patting a baby on the back five times; handing in blocks five times). She then talks about adding a group of three one-fourths to a bucket, and continuing to do so until five such groups have been added. Drawing a connection to the approach discussed before, she associates the 15 ¼'s in her bucket drawing with the numerator of the final improper fraction in the first method; she also reminds students that in previous grades, they discussed multiplication as repeated addition of whole numbers. Ms. Jones then has students act out this process by cutting circles into fourths and handing the appropriate number of fourths to another student, then counting pieces to determine their answer, next checking if their fourths correspond to the numerator of the final improper fraction they got when applying the first method, and finally putting the fourths back into full circles to form their mixed number. Students are given about five minutes to act out this process in pairs. In wrapping up the lesson, Ms. Jones emphasizes that multiplication can be thought of as repeated addition and recapitulates some connections between the two methods considered in the lesson.

8 Structure of the papers in the special issue

To enhance comparability, the papers that follow share the same structure. Firstly, the theoretical rationale of the framework/approach and theoretical conceptualizations informing each work are presented, so that the reader is provided with sufficient background and detail for each

framework (Research Question 1). Secondly, existing instruments and operationalizations of the frameworks are provided to enable the reader understand how each group/project goes about studying instruction and its quality (Research Question 2). This is followed by empirical support in terms of the framework's/approach's reliability and validity (Research Question 3). Finally, the three lessons described in Sect. 7 are analyzed, using the lens that each framework offers. In the context of this analysis, the strengths and weaknesses of the framework/approach are outlined (Research Question 4). The analysis of the three lessons and the discussion of the strengths and limitations of each framework presented in each paper allows for the exploration of synergies and complementarities, thus addressing Research Questions 5 and 6, an endeavor undertaken in the concluding paper of this special issue.

9 Conclusion

The past years have seen an accumulated body of studies aiming better to conceptualize, operationalize, and measure instructional quality through different lenses. The time seems ripe to capitalize on this accumulated and rapidly expanding work by bringing together several different approaches. We believe that only through better understanding instruction can we really improve it and consequently have an impact on student learning. We hope that this special issue makes a step toward this end by setting the groundwork for understanding instructional quality more comprehensively.

Appendix A

Our literature search was based on the search engines Scopus, Web of Science, and Google Scholar, using the following search terms:

- Classroom observation AND mathematics.
- Multiple observation instruments AND mathematics.
- Multiple observation instruments AND mathematics instruction.
- Multiple lenses for classroom observations.
- Multiple lenses for lesson observations.
- Multiple lenses for observations.
- Different lenses to capture teaching AND mathematics.
- Generic teaching practices AND mathematics.
- Content specific teaching practices AND mathematics.
- Integrat* generic and content-specific practices.
- Teacher observations AND mathematics.

References

Ball, D. L. (1992). Magical hopes: Manipulatives and the reform of math education. *American Educator, 16*(2), 14–18, 46–47.

Ball, D. L., Lewis, J., & Thames, M. H. (2008). Making mathematics work in school. In A. J. Schoenfeld & N. Pateman (Eds.), *A study of teaching: Multiple lenses, multiple views: Journal for research in mathematics education, monograph #14* (pp. 13–44). Reston: NCTM.

Bell, C. A., Gitomer, D. H., McCaffrey, D. F., Hamre, B. K., Pianta, R. C., & Qi, Y. (2012). An argument approach to observation protocol validity. *Educational Assessment, 17*(2–3), 62–87. https://doi.org/10.1080/10627197.2012.71501.

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

Blazar, D., & Kraft, M. A. (2017). Teacher and teaching effects on students' attitudes and behaviors. *Educational Evaluation and Policy Analysis, 39*(1), 146–170. https://doi.org/10.3102/0162373716670260.

Blazar, D., Litke, E., & Barmore, J. (2016). What does it mean to be ranked a "high" or "low" value-added teacher? Observing differences in instructional quality across districts. *American Educational Research Journal, 53*(2), 324–359. https://doi.org/10.3102/0002831216630407.

Booker, L. N. (2014). *Examining the development of beginning middle school math teachers' practices and their relationship with the teachers' effectiveness*. Unpublished doctoral dissertation, Vanderbilt University, Nashville, Tennessee, USA.

Boston, M., Bostic, J., Lesseig, K., & Sherman, M. (2015). A comparison of mathematics classroom observation protocols. *Mathematics Teacher Educator, 3*(2), 154–175. https://doi.org/10.5951/mathteacduc.3.2.0154.

Cantrell, S., & Kane, T. J. (2013). Ensuring fair and reliable measures of effective teaching. Seattle: Bill & Melinda Gates Foundation. http://www.metproject.org/reports.php. Accessed 19 Feb 2015.

Charalambous, C. Y., & Kyriakides, E. (2017). Working at the nexus of generic and content-specific teaching practices: An exploratory study based on TIMSS secondary analyses. *The Elementary School Journal, 117*(3), 423–454. https://doi.org/10.1086/690221.

Cohen, D. K. (2011). *Teaching and its predicaments*. Cambridge: Harvard University Press.

Common Core State Standards Initiative [CCSSI] (2014). *Mathematics Standards*. http://www.corestandards.org/Math/. Accessed 20 Feb 2014.

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.

Cuoco, A. A., & Curcio, F. R. (Eds.). (2001). *The roles of representation in school mathematics (NCTM 2001 Yearbook)*. Reston: NCTM.

Danielson, C. (2013). *Enhancing professional practice: A framework for teaching* (2nd edn.). Alexandria: Association for Supervision and Curriculum Development.

Dignath-van Ewijk, C., Dickhäuser, O., & Büttner, G. (2013). Assessing how teachers enhance self-regulated learning–A multi-perspective approach. *Journal of Cognitive Education and Psychology, Special Issue of Self-Regulated Learning, 12*(3), 338–358. https://doi.org/10.1891/1945-8959.12.3.338.

Doabler, C. T., Baker, S. K., Kosty, D. B., Clarke, B., Miller, S. J., & Fien, H. (2015). Examining the association between explicit mathematics instruction and student mathematics achievement. *Elementary School Journal, 115*(3), 303–333. https://doi.org/10.1086/679969.

Doyle, W. (1986). Classroom organization and management. In M. C. Wittrock (Ed.), *Handbook of research on teaching* (3rd edn., pp. 392–431). London: Macmillan.

Dubinski, W. N., Waxman, H. C., Brown, D. B., & Kelly, L. J. (2016). Informing teacher education through the use of multiple classroom observation instruments. *Teacher Education Quarterly, 43*(1), 91–106.

Gibbons, C. J. (2016). *Variations of mathematics in college algebra instruction: An investigation through the lenses of three observation protocols*. Unpublished Master's thesis, Oregon State University, Corvallis, Oregon, USA.

Grossman, P., & McDonald, M. (2008). Back to the future: Directions for research in teaching and teacher education. *American Educational Research Journal, 45*(1), 184–205. https://doi.org/10.3102/0002831207312906.

Henry, M. A., Murray, K. S., & Phillips, K. A. (2007). *Meeting the challenge of STEM classroom observation in evaluating teacher development projects: A comparison of two widely used instruments*. http://sirc.mspnet.org/index.cfm/19945. Accessed 12 Feb 2013.

Hiebert, J., Gallimore, R., Garnier, H., Givvin, K. B., Hollingsworth, H., Jacobs, J., … & Stigler, J. (2003). *Teaching mathematics in seven countries: Results from the TIMSS 1999 video study*. U.S. Department of Education. Washington DC: National Center for Education Statistics.

Hill, H. C., Blunk, M., Charalambous, C. Y., Lewis, J., 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. https://doi.org/10.1080/07370000802177235.

Hill, H. C., Charalambous, C. Y., & Kraft, M. (2012). When rater reliability is not enough: Teacher observation systems and a case for the G-study. *Educational Researcher, 41*(2), 56–64. https://doi.org/10.3102/0013189X12437203.

Horizon Research, I. (2000). *Inside the classroom observation and analytic protocol*. Chapel Hill: Horizon Research, Inc. http://www.horizon-research.com/instruments/clas/cop.pdf. Accessed 18 Sep 2008.

Horn, I. S. (2008). Accountable argumentation as a participation structure to support learning through disagreement. In A. J. Schoenfeld & N. Pateman (Eds.), *A study of teaching: Multiple lenses, multiple views: Journal for Research in Mathematics Education, Monograph #14* (pp. 97–126). Reston: NCTM.

Kane, T. J., & Staiger, D. O. (2012). *Gathering feedback for teaching: Combining high-quality observations with student surveys and achievement gains*. Seattle: Bill & Melinda Gates Foundation. http://www.metproject.org/reports.php. Accessed 30 Nov 2012.

Kane, T. J., Taylor, E. S., Tyler, J. H., & Wooten, A. L. (2011). Identifying effective classroom practices using student achievement data. *Journal of Human Resources, 46*(3), 587–613. https://doi.org/10.3386/w15803.

Kilday, C. R., & Kinzie, M. B. (2009). An analysis of instruments that measure the quality of mathematics teaching in early childhood. *Early Childhood Education Journal, 36*(4), 365–372. https://doi.org/10.1007/s10643-008-0286-8.

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

Lampert, M. (2001). *Teaching problems and the problems of teaching*. New Haven and London: Yale University Press.

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

Leinhardt, G. (1993). On teaching. In R. Glaser (Ed.), *Advances in instructional psychology* (pp. 1–54). Hillsdale: Lawrence Erlbaum Associates.

Litke, E. G. (2015). *The state of the gate: A description of instructional practice in algebra in five urban districts.* Unpublished doctoral dissertation, Harvard University, Cambridge, MA.

Lockwood, J. R., Savitsky, T. D., & McCaffrey, D. F. (2015). Inferring constructs of effective teaching from classroom observations: An application of Bayesian exploratory factor analysis without restrictions. *The Annals of Applied Statistics, 9*(3), 1484–1509. https://doi.org/10.1214/15-AOAS833.

Marder, M., & Walkington, C. (2014). Classroom observation and value-added models give complementary information about quality of mathematics teaching. In T. Kane, K. Kerr & R. Pianta (Eds.), *Designing teacher evaluation systems: New guidance from the measuring effective teaching project* (pp. 234–277). New York: Wiley.

Martin-Raugh, M., Tannenbaum, R. J., Tocci, C. M., & Resse, C. (2016). Behaviorally anchored rating scales: an application for evaluating teaching practice. *Teaching and Teacher Education, 59*, 414–419. https://doi.org/10.1016/j.tate.2016.07.026.

Matsumura, L. C., Garnier, H., Slater, S. C., & Boston, M. (2008). Toward measuring instructional interactions 'at-scale.'. *Educational Assessment, 13*(4), 267–300. https://doi.org/10.1080/10627190802602541.

McGuire, P. R., Kinzie, M., Thunder, K., & Berry, R. (2016). Methods of analysis and overall mathematics teaching quality in at-risk prekindergarten classrooms. *Early Education and Development, 27*(1), 89–109. https://doi.org/10.1080/10409289.2015.968241.

Mihaly, K., McCaffrey, D., Staiger, D., & Lockwood, J. R. (2013). A composite estimator of effective teaching (MET Project). The RAND Corporation. http://www.rand.org/pubs/external_publications/EP50155.html. Accessed 21 March 2015.

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

Muijs, D., Kyriakides, L., van der Werf, G., Creemers, B., Timperley, H., & Earl, L. (2014). State of the art–teacher effectiveness and professional learning. *School Effectiveness and School Improvement, 25*(2), 231–256. https://doi.org/10.1080/09243453.2014.885451.

National Council of Teachers of Mathematics [NCTM] (2000). *Principles and standards for school mathematics.* Reston: NCTM.

Nowińska, E., & Praetorius, A. K. (2017). Evaluation of a rating system for the assessment of metacognitive-discursive instructional quality. In *Proceedings of the tenth congress of the European Society for Research in Mathematics Education.* https://keynote.conference-services.net/resources/444/5118/pdf/CERME10_0379.pdf. Accessed 30 Sept 2017.

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

Pinter, H. H. (2013). *Patterns of teacher's instructional moves: What makes mathematical instructional practices unique?* Unpublished Doctoral Dissertation. University of Virginia, Charlottesville, USA.

Posner, T. (2008). Equity in a mathematics classroom: An exploration. In A. J. Schoenfeld & N. Pateman (Eds.), *A study of teaching: Multiple lenses, multiple views: Journal for research in mathematics education, MONOGRAPH #14* (pp. 127–172). Reston: NCTM.

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. https://doi.org/10.1016/j.learninstruc.2013.12.002.

Sammons, P., Lindorff, A. M., Ortega, L., & Kington, A. (2016). Inspiring teaching: Learning from exemplary practitioners. *Journal of Professional Capital and Community, 1*(2), 124–144. https://doi.org/10.1108/JPCC-09-2015-0005.

Sawada, D., Piburn, M. D., Judson, E., Turley, J., Falconer, K., Benford, R., & Bloom, I. (2002). Measuring reform practices in science and mathematics classrooms: The reformed teaching observation protocol. *School Science and Mathematics, 102*(6), 245–253. https://doi.org/10.1111/j.1949-8594.2002.tb17883.

Schaffer, E., Muijs, D., Reynolds, D., & Kitson, K. (1998). *The mathematics enhancement classroom observation recording system, MECORS).* Newcastle: University of Newcastle, School of Education.

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. https://doi.org/10.1007/s11858-016-0765-0.

Schlesinger, L., Jentsch, A., Kaiser, G., König, J., & Blömeke, S. (2018). Subject-specific characteristics of instructional quality in mathematics education. *ZDM Mathematics Education, 50*(3). **(this issue)**.

Schoenfeld, A. H. (2008). On modeling teachers' in-the-moment decision making. In A. J. Schoenfeld & N. Pateman (Eds.), *A study of teaching: Multiple lenses, multiple views: Journal for research in mathematics education, monograph #14* (pp. 45–96). Reston, VA: NCTM.

Schoenfeld, A. H. (2013). Classroom observations in theory and practice. *ZDM—The International Journal on Mathematics Education, 45*(4), 607–621. https://doi.org/10.1007/s11858-012-0483-1.

Schoenfeld, A. J., & Pateman, N. (2008). *(Eds.). A study of teaching: Multiple lenses, multiple views: Journal for research in mathematics education. Monograph #14.* Reston: NCTM.

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. https://doi.org/10.1080/13803610600874067.

Thompson, C. J., & Davis, S. B. (2014). Classroom observation data and instruction in primary mathematics education: Improving design and rigour. *Mathematics Education Research Journal, 26*(2), 301–323. https://doi.org/10.1007/s13394-013-0099-y.

Tsamir, P., Tirosh, D., Tabach, M., & Levenson, E. (2010). Multiple solution methods and multiple outcomes—is it a task for kindergarten children? *Educational Studies in Mathematics, 73*(3), 217–231. https://doi.org/10.1007/s10649-009-9215-z.

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

Walkowiak, T. A., Berry, R. Q., Meyer, J. P., Rimm-Kaufman, S. E., & Ottmar, E. R. (2014). Introducing an observational measure of standards-based mathematics teaching practices: Evidence of validity and score reliability. *Educational Studies in Mathematics, 85*(1), 109–128. https://doi.org/10.1007/s10649-013-9499-x.

Wragg, E. C. (2012). *An introduction to classroom observation.* NY: Routledge.