ZDM – Mathematics Education
https://doi.org/10.1007/s11858-024-01591-x

ORIGINAL PAPER

Integrating subject-generic and subject-specific teaching frameworks: searching for stages of teaching in mathematics

Leonidas Kyriakides[1] · Charalambos Y. Charalambous[1] · Panayiotis Antoniou[1]

Accepted: 11 May 2024

Abstract
Currently there is an attempt to combine subject-generic and subject-specific teaching frameworks to comprehensively capture teaching quality. This study explores the possibility of integrating two widely used and validated frameworks, the subject-generic Dynamic Model of Educational Effectiveness (DMEE) and the subject-specific Mathematical Quality of Instruction (MQI). Toward this end, we drew on data from 38 upper-grade primary school teachers, each observed in six mathematics lessons, which were coded using both frameworks. Data were analyzed using the Extended Logistic model of Rasch to explore whether a common scale of teaching quality with good psychometric properties could be developed. Saltus was then utilized to investigate the possibility of forming levels of effective teaching in mathematics. A common scale encompassing both subject-generic and subject-specific teaching aspects, which had good psychometric properties, was developed. The subject-generic and subject-specific teaching aspects of these frameworks were clustered in five distinct levels. With the exception of the top level that included only subject-generic aspects, all other levels included teaching aspects from both frameworks, thus providing support to the assumption that it is possible to develop levels of effective teaching that combine related subject-generic and subject-specific aspects. In discussing the study findings, we consider their implications for developing an integrated framework of teaching quality and for developing professional development programs that combine subject-generic and subject-specific teaching aspects.

Keywords Subject-generic teaching aspects · Subject-specific teaching aspects · Primary school mathematics · Levels of effective teaching · Teaching quality

1 Introduction

Although for years scholars in the field of teaching quality have been attending to either subject-generic or subject-specific teaching aspects (TAs)—with the former cutting across different subjects and the latter being more germane to teaching particular subjects (cf. Charalambous & Kyriakides, 2017)—the last decade has seen heightened interest in considering both types of TAs (see, for example, such a discussion in the special issue ZDM – Mathematics Education, 50(3). This was fueled by theoretical arguments recognizing the complexity of teaching (Cohen, 2011) and underlining the need to combine theoretical frameworks to better study teaching quality (e.g., Hamre et al., 2013; Charalambous & Praetorius, 2018). Empirical studies reveal that considering both types of TAs can help better capture teacher-student interactions around the content (Praetorius & Charalambous, 2018) and showing either the combination to explain a higher percentage of the unexplained variance in student learning (e.g., Charalambous & Kyriakides, 2017) or working complementarily to predict student learning (e.g., Blazar & Kraft, 2017). Scholars in mathematics education have recently started discussing different ways of combining the two types of TA (Brunner, 2018; Blazar et al., 2017), and developing frameworks attempting to integrate these TAs (e.g., the MAIN-TEACH model, Charalambous & Praetorius, 2020).

✉ Leonidas Kyriakides
kyriakid@ucy.ac.cy

Charalambos Y. Charalambous
cycharal@ucy.ac.cy

Panayiotis Antoniou
antoniou.panayiotis@ucy.ac.cy

[1] Department of Education, University of Cyprus, P. O. Box 20537, Nicosia 1678, Cyprus

Published online: 06 June 2024

Nevertheless, the relationship between subject-generic and subject-specific TAs is still unclear. Can subject-generic and subject-specific TAs be integrated? If so, how? Addressing this question could help catalyze the development of frameworks that bring together subject-generic and subject-specific TAs by not simply juxtaposing them but by exploring their similarities and potential overlaps so that the outcome framework is not simply the sum of its parts. In this paper, we address this question from an empirical perspective, reflecting on whether a common scale encompassing both subject-generic and subject-specific aspects can be developed. Forming such a scale may help examine if and how subject-generic TAs relate to subject-specific TAs, and particularly whether subject-generic and subject-specific TAs belong to a single overarching factor (i.e., teaching quality) and are therefore related to each other. We also examine whether they can be organized into separate distinct groups, or if these groups include both subject-generic and subject-specific TAs. If the first holds, it implies that integrating these two different TAs can largely be equated to simply juxtaposing distinct aspects. If the latter is true, it points to the potential of integrating the TAs in ways that acknowledge the interrelations between these two different types of TAs.

Such empirical work may have theoretical and practical implications. From a theoretical perspective, it can provide insights into developing frameworks that integrate these two types of TAs by considering their similarities and differences. It could also help start reflecting on whether TAs which have for years been considered as "generic" or "specific" fall along a continuum with the boundaries between the two types of practices being more blurred than initially thought. From a practical standpoint, it can inform initial and ongoing professional development programs for pre-service and in-service teachers since it may provide teacher educators with ideas as to how the two types of TAs can be integrated.

In this study we bring together two widely used frameworks, one subject-generic (*the Dynamic Model of Educational Effectiveness* [DMEE], Creemers & Kyriakides, 2008) and one subject-specific (the *Mathematical Quality of Instruction* [MQI], Learning Mathematics for Teaching [LMT] Project, 2011). Before justifying this selection (see Sect. 2), two remarks are in order. First, the exercise undertaken in this study needs to be replicated by drawing on other subject-generic and subject-specific frameworks[1]. Such replication studies may help us understand how frameworks that aim to integrate both types of TAs can be developed. Second, in selecting the two frameworks, we followed the classification of Charalambous and Praetorius (2018), acknowledging that subject-specificity and subject-genericness in frameworks should not be considered as dichotomous but rather as forming a continuum. We therefore based our selection of the two frameworks on their developers' original intentions (i.e., to study teaching quality in a specific subject or across different subjects) and the extent to which the development of these frameworks was informed by subject-specific demands of teaching within a particular discipline.

In pursuing this exercise to work at the intersection of subject-generic and subject-specific frameworks, we recognize that this is not the first endeavor to do so. Several frameworks that combine these two types of TAs—identified as hybrid (see Charalambous & Praetorius, 2018)—have been developed and used over the past decade. These include, among others, TEDS-Instruct (Schlesinger et al., 2018), Teaching for Robust Understanding (Schoenfeld, 2018), and the UTeach Observation Protocol (Walkington & Marder, 2018). Empirical studies utilizing these frameworks suggest that they capture an overarching factor of teaching quality combining both types of TAs (e.g., see Blömeke et al., 2022 for TEDS-Instruct). Recognizing the merit of these frameworks as a promising approach for integration, in this work we follow a rather different approach that brings together two different frameworks and explore the possibility of developing a common scale that encompasses both aspects. This work could also be promising because it allows for collecting more detailed information on teaching quality, given that each framework goes into more depth in capturing either subject-generic or subject-specific aspects.

2 The theoretical frameworks of the study

2.1 The Dynamic Model of Educational Effectiveness (DMEE)

The DMEE (Creemers & Kyriakides, 2008) was developed to establish stronger links between Educational Effectiveness Research (EER) and research on improvement by considering the strengths and limitations of the main integrated models of EER (e.g., Creemers, 1994; Stringfield & Slavin, 1992). The DMEE is multilevel in nature and refers to factors operating at the student, classroom, school, and system levels which are associated with student learning outcomes. Five dimensions are used to measure both quantitative (i.e.,

[1] In this exercise we utilized two widely disseminated frameworks with which we were more familiar. Other subject-generic frameworks (e.g., Three Basic Dimensions [TBD], Praetorius et al., 2018; Classroom Assessment Scoring System [CLASS], Berlin & Cohen, 2018) or subject-specific frameworks (e.g., Instructional Quality Assessment [IQA], Boston & Candela, 2018; Mathematics-Scan [M-Scan], Walkowiak et al., 2018) could also be used toward this end in future replication studies.

frequency) as well as qualitative characteristics of the functioning of each factor (i.e., focus, stage, quality, and differentiation). The dimensions are not only important from a measurement perspective, but also, and even more, from a theoretical point of view. The focus dimension is in line with the synergy theory (Liu & Jiang, 2018) and argues that the specificity and the number of purposes addressed by each task associated with a factor should be examined. Similarly, the stage dimension implies that the factors need to take place over a long period of time to ensure that they have a continuous direct or indirect effect on student learning. Several studies provided empirical support to this argument (e.g., Creemers, 1994; Scheerens, 2013). Using the stage dimension to measure the functioning of a factor can help identify the extent to which there is constancy at each level and flexibility in using each factor. Finally, the differentiation dimension is in line with findings of research into differential effectiveness (Campbell et al., 2003) which reveals that adaptation to the specific needs of each group of students may increase the successful implementation of a factor and ultimately maximize its effect on student learning outcomes (Tomlinson, 2014). Appendix A shows how each dimension is used in measuring the orientation factor (for more information on how each factor is measured see Creemers & Kyriakides, 2008).

At the classroom level, the DMEE takes into account the main findings of teacher effectiveness research and refers to factors concerned with teacher behavior in the classroom found to be associated with student learning outcomes. It also attempts to develop a comprehensive framework of effective teaching by considering different theories of learning and different teaching approaches. Specifically, the model refers to eight factors (hereafter TAs) associated with student learning outcomes in different learning domains (Scheerens, 2013). The main elements of the eight factors are mentioned in Fig. 1 which reveals that the DMEE refers to TAs such as structuring and application (found to be related with student learning outcomes by the early teacher effectiveness studies – see Brophy & Good, 1986) and are associated with the direct and active teaching approach (Joyce et al., 2000) and to modeling and assessment which are in line with constructivism (Schoenfeld, 1998). Moreover, the collaboration technique is considered in defining the elements of the classroom learning environment. Multiple theories of learning are considered in defining the TAs. For example, motivation learning theories and the cognitive load theory are considered in defining orientation and application, correspondingly. Finally, some factors of the DMEE refer to TAs also captured in other subject-generic frameworks. For example, modeling and questioning (i.e., raising process questions) align with cognitive activation included in TBD (Praetorius et al., 2018) and TEDS-Instruct (Schlesinger et al., 2018). Similarly, management of time can be identified in several other frameworks (e.g., CLASS, TBD, TEDS-Instruct). (For a more systematic description of the factors of DMEE and their relationship with other theoretical models, see Kyriakides et al., 2020).

More than 20 large-scale studies and one meta-analysis have been conducted to examine the main assumptions of the DMEE at classroom level (for a review of these studies see Kyriakides et al., 2020). Below, issues of validity, reliability and prediction of student outcomes are briefly discussed. One high-inference and two low-inference observation instruments, as well as a student questionnaire are being used to capture the TAs examined by the DMEE (for more information see Creemers & Kyriakides, 2012; also see Appendix A for a description of the instrument used in the present study). Kyriakides and Creemers (2008) analyzed data emerged from these instruments by using a multi-trait and multi-method model and provided support to the *construct validity* of the instruments. This model was then replicated in more than 20 studies (for a review of these studies see Kyriakides et al., 2020) and confirmatory factor analyses revealed that each TA can be measured in relation to the five dimensions (e.g., Bodroža et al., 2022; Dierendonck, 2023). These studies showed that students were able to provide reliable data on the teaching practices of their teachers. Satisfactory results about the reliability of the observations instruments were also generated (the alpha reliability coefficients for each TAs as captured by the three observation instruments were higher than 0.83, and the reported inter-rater reliability coefficients r^2 were higher than 0.75).

Finally, these studies revealed that TAs are associated with student achievement gains. Cognitive learning outcomes in different subjects (e.g., mathematics, language, science, and religious education) as well as non-cognitive outcomes (e.g., attitudes towards mathematics) were used to measure the impact of the TAs. Thereby some support for the assumption that the TAs are associated with student achievement gains in different learning outcomes has been provided (Chaudhary & Singh, 2022; Polymeropoulou & Lazaridou, 2022). The generic nature of the DMEE is also supported since a synthesis of these studies revealed that the effects of the TAs on different student learning outcomes were similar (i.e., Cohen's *d* values were around 0.20). However, only two studies examined the impact of the teacher factors on non-cognitive outcomes and only one on student metacognitive outcomes.

The DMEE assumes that the eight TA are related to each other. Six studies conducted in different countries (i.e., Canada, Cyprus, Greece, Maldives, and Taiwan) revealed that the TAs can be classified into specific levels of effective teaching. For example, a study focusing on Cypriot primary teachers teaching three different subjects (i.e., Mathematics,

TA	Main elements
1. Orientation	a) Providing the objectives for a specifc task/lesson/series of lessons. b) Challenging students to identify the reason(s) that an activity is taking place in the lesson.
2. Structuring	a) Beginning with an overview and/or review of objectives. b) Outlining the content to be covered. c) Signalling transitions between lesson parts. d) Drawing attention to, and reviewing, main ideas.
3. Questioning	a) Raising different types of question (i.e., process and product) at an appropriate diffculty level. b) Giving students time to respond. c) Dealing with student responses.
4. Teaching modelling	a) Encouraging students to use problem-solving strategies presented by the teacher or other classmates. b) Inviting students to develop their own strategies. c) Promoting the idea of modelling.
5. Application	a) Using seatwork or small-group tasks in order to provide necessary practice and application opportunities. b) Using application tasks as starting points for the next step in teaching and learning.
6. The classroom as a learning environment	a) Establishing on-task behaviour through the interactions that take place (i.e., teacher-student and student-student interactions). b) Dealing with classroom disorder and student competition by establishing rules, persuading students to respect them and implementing the rules.
7. Management of time	a) Organising the classroom environment. b) Maximising engagement rates.
8. Assessment	a) Using appropriate techniques to collect data on students' knowledge and skills. b) Analysing data in order to identify student needs. c) Reporting the assessment results to students and parents. d) Evaluating their own teaching practices.

Fig. 1 Description of the Main Teaching Aspects (TAs) of the Dynamic Model of Educational Effectiveness

Greek language, and Religious Education) revealed five levels of effective teaching (Kyriakides et al., 2009). The first three levels were related to the direct and active teaching approach, by moving from the basic requirements concerning quantitative characteristics of teaching routines to the more advanced requirements concerning the appropriate use of these skills as these are measured by the qualitative characteristics of these TAs. These skills also gradually move from the use of teacher-centered approaches to the active involvement of students. The last two levels were more demanding since teachers are expected to differentiate instruction (Level 4) and demonstrate their ability to use constructivism (Level 5). Teachers situated at higher levels were also found to be more effective in terms of promoting student learning outcomes in each subject. Similar results emerged from studies conducted in other countries. The results of these studies were considered in developing the dynamic approach to teacher improvement (Creemers et al., 2013).

2.2 Mathematical Quality of Instruction (MQI)

Recognizing that the existing classroom observation frameworks did not capture the *mathematical* quality in teaching, the MQI developers sought to develop a framework that would be sensitive to the mathematical nuances in teaching (LMT Project, 2011). Toward this end, they drew on the instructional triangle (cf. Cohen et al., 2003), thinking of instruction as comprised of dynamic interactions among the teacher, the students, and the content, which were situated in educational settings. The framework was developed following both a top-down and a bottom-up approach, through

iterative cycles of examining the literature to identify TAs that are germane to teaching mathematics and a close analysis of video recorded elementary mathematics lessons (see Hill, 2010; LMT Project, 2011 for more information on this process). As such the MQI design fulfils the first perspective of subject-specificity[2] proposed by Mu et al. (2022): that of applicability. The discussion of the derived TAs among different experts in teaching mathematics partly accounted for the second perspective, relevance (see Mu et al., 2022); a discussion with experts in other fields could have offered additional insights about the degree of subject-specificity of the chosen TAs. Since its initial development, the MQI framework has gone through different iterations. In its current form, it includes four dimensions (hereafter, TAs) with twenty items (see Fig. 2). The first two TAs reflect the relationship between the teacher and the content; the third TA focuses on how the teacher facilitates students' interactions with the content, while the fourth captures students' interaction with the content.

MQI has been used in several studies to investigate the relationship between teaching quality and mathematical knowledge for teaching (MKT) and student learning. With respect to the association between MQI and MKT, most studies (Hill et al., 2008; Kelcey et al., 2019; Lee & Santagata, 2020; Santagata & Lee, 2021) have focused on elementary grades. Although employing different teacher populations and dissimilar designs, these studies converge in showing a positive association between MKT and MQI. For example, studies that used small samples of elementary school teachers ($N<10$)—ranging from first-year teachers (Santagata & Lee, 2021) to novice teachers (Lee & Santagata, 2020) to more seasoned teachers (Hill et al., 2008)—showed moderate to strong positive associations between MKT and aspects of MQI (ranging from to r_{rho}=0.65 up to r_{rho}=0.83). Interestingly, Lee and Santagata's (2020) longitudinal study, showed that whereas during the first year of the study these correlations were not significant, they became so in the second year of the study, suggesting that the effects of MKT on teaching quality might not be directly identified during teachers' early career stages. Analogous findings also emerged with larger samples of elementary school teachers (e.g., $N\approx270$). For example, in Kelcey et al. (2019) a one standard deviation difference in teacher knowledge was associated with about a 0.22 standard deviation change in quality in Ambitious teaching (a collective term that encompasses the first, the third, and the fourth MQI TAs, see Fig. 2) and a 0.35 change in *Errors*

TA	Description	Codes
1. Richness of the mathematics	Captures the depth of the mathematics offered to students. Rich mathematics can be established by either focusing on the meaning of facts and procedures or by focusing on key mathematical practices.	a. **Linking and connections**: drawing explicit connections between representations of mathematical concepts b. **Explanations**: teacher/students offer(s) mathematical explanations c. **Mathematical meaning and sense-making**: instruction geared toward supporting students make meaning of mathematical ideas d. **Multiple procedures/solution methods**: teacher/students use(s) and discuss(es) multiple approaches to a problem e. **Patterns and generalizations**: teacher/students identify/ies patterns and use(s) them to develop generalizations f. **Mathematical language**: captures the depth and density of the mathematical language used during instruction g. **Overall code**: a holistic code that captures the overall depth of the mathematics offered to students
2. Errors and imprecision	Explores the degree to which instruction is mathematically incorrect.	a. **Mathematical content errors**: captures instances in which instruction includes major mathematical errors b. **Imprecision in language and notation**: includes instances of teacher's imprecise mathematical language or notation that interfere with the presentation of the content c. **Lack of clarity in presentation of mathematical content**: captures instances when the presentation of mathematical ideas is unclear or confusing, thus obscuring the mathematical content d. **Overall code**: a holistic code that captures the overall presence of teacher errors in doing and talking about mathematics
3. Working with students and mathematics	Refers to teachers' ability to appropriately interpret and respond to students' mathematical ideas and errors, as well as their capability to capitalize on them during instruction.	a. **Remediation of student errors and difficulties**: examines the degree to which the teacher responds to student errors or misunderstandings and offers procedural/ conceptual remediation b. **Teacher uses mathematical contributions**: captures whether and how the teacher responds to and builds on students' mathematical productions to steer the lesson toward a mathematical goal c. **Overall code**: a holistic code that provides an evaluation of the teacher-student interactions around the content
4. Common Core Aligned Student Practices	Focuses on student involvement with the mathematical content through cognitively demanding activities.	a. **Students provide explanations**: assesses students' explanations of ideas, procedures or solutions b. **Student mathematical questioning and reasoning**: explores instances of students engaging in meaning-oriented mathematical practices (e.g., posing mathematically motivating questions, making hypotheses, offering examples) c. **Students communicate about the mathematics of the segment**: captures the extent to which students contribute mathematically to the lesson d. **Task cognitive demand**: examines student engagement in tasks in which they think deeply and reason about mathematics; this code refers to the enactment of the task, regardless of the initial demand of the curriculum/textbook task or how the teacher sets up the task for students. e. **Students work with contextualized problems**: captures the extent to which students work on problems situated in real-world contexts f. **Overall code**: a holistic code capturing evidence of students' involvement in the mathematics of the lesson and the extent to which students participate in and contribute to meaning-making and reasoning

Fig. 2 Description of the Four Teaching Aspects (TAs) of the Mathematical Quality of Instruction

[2] Following Mu et al. (2022), we acknowledge that "subject-specific" is more appropriate for describing MQI rather than the term "content-specific" used in prior publications (e.g., Charalambous & Litke, 2018). Hence, in this paper, we refer to MQI as a subject-specific framework.

and Imprecision. In middle-grades, Hill et al.'s (2012) study with 24 middle-grade teachers showed a moderate correlation between MKT and MQI (r_{rho}=0.58, $p < .01$). Collectively, these results support that MQI satisfies knowledge, the third perspective of subject-specificity (Mu et al., 2022).

Both large scale (Kane & Staiger, 2012) and smaller scale (Blazar & Archer, 2020; Blazar & Kraft, 2017; Blazar et al., 2016; Hill et al., 2011; Kraft & Hill, 2020; Kelcey et al., 2019; Mantzicopoulos et al., 2019) studies have examined the predictive validity of teachers' MQI scores on their student learning. Although the findings of these studies are mixed, they point to a pattern of positive associations between MQI and student learning, thus supporting that MQI *partly* satisfies the last perspective of subject-specificity, predictivity (Mu et al., 2022).

The largest study conducted, the Measures of Effective Teaching [MET] Project, found positive significant, yet low, correlations (from $r = .12$ to $r = .16$) between teachers' MQI scores and students' scores on state tests and a more cognitively demanding project-administered test (Kane & Staiger, 2012). Higher relations were found in smaller-scale studies. Drawing on a sample of 24 middle-school teachers and their 222 students, Hill et al. (2011) found moderate associations between teachers' MQI scores and students' value-added scores ($r_{\text{rho}} = 0.30$ to $r_{\text{rho}} = 0.56$ in the different value-added models employed). Collectively these low to moderate correlations suggest that teaching quality, as measured by MQI is associated with student learning. Other studies that used more advanced analyses than simple correlations provide stronger evidence for this association. In addition, Blazar (2015) showed the overarching TA of *Ambitious Teaching* to positively predict fourth-and fifth-grade students' scores on a low-stakes mathematics test ($\beta = 0.11$, SE = 0.04, $p <$ .05). Focusing on different student learning outcomes and drawing on a sample of 310 fourth- and fifth-grade teachers and their 10,575 students, Blazar and Kraft (2017) showed *Errors and Imprecisions* to be negatively associated with performance on state-tests (β=−0.02, SE = 0.01, $p <$ .10), self-efficacy (β=-0.09, SE = 0.03, $p <$ .01), and happiness (β=-0.18, SE = 0.08, $p <$ .05); however, *Ambitious Teaching* was not significantly related with any of these outcomes. A similar non-significant effect was also found in a randomized field trial utilizing coaching to improve elementary and middle-school teachers' MQI (Kraft & Hill, 2020); although coaching did result in improvements in the mathematical quality in teachers' lessons, this improvement was not reflected on students' learning as captured on formative and summative assessments. In contrast, a study focusing on a younger student population (285 kindergarten students, see Mantzicopoulos et al., 2019) showed scores on *Ambitious Teaching* to predict students' end-of-year progress on kindergarten mathematics standards ($\beta = 1.63$, SE = 0.75, $p <$ .05) but not their mathematical reasoning score ($\beta = 1.78$, SE = 1.81, $p > .05$). The whole lesson MQI scores were also associated with teacher-rated students' interest in mathematics ($\beta = 0.82$, SE = 0.34, $p < .05$).

During the last five years, studies have also focused on examining differential effects of MQI on student learning, exploring for which students and under what conditions higher MQI scores are conducive to student learning. For example, Blazar and Archer (2020) showed *Ambitious Teaching* to be more effective for English language learners ($\beta = 0.07$, SE = 0.03, $p <$ .05) compared to the general student population ($\beta = 0.02$, SE = 0.02, $p >$ .05). Similarly, Kelcey et al. (2019) showed that the significant positive relationship between achievement gains and *Ambitious Teaching* was present only in state districts with coherent instructional guidance and whose state tests were more cognitively challenging ($\beta = 0.11$, SE = 0.04, $p <$ .05) as compared to districts not having these characteristics (β=-0.07, SE = 0.05, $p >$ .05).

2.3 Exploring possibilities for integrating the two frameworks

Although the DMEE and MQI represent two distinct traditions in studying teaching quality in mathematics, with the first focusing on subject-generic TAs (cf. Panayiotou et al., 2021) and the second zooming in on the *mathematical* aspects of teaching (cf. Litke et al., 2021), there are both empirical and theoretical reasons to bring them together and explore possibilities of integration.

Empirically speaking, both frameworks have been validated in the same educational context, which is also the context considered herein. Second, although being a generic framework, DMEE has been used extensively in studying teaching quality in mathematics (cf. Kyriakides et al., 2020). Finally, both frameworks have been used extensively to capture teaching quality in primary grades, which is the focus of this study. Collectively, these three elements suggest that any difficulties that might arise when integrating the two frameworks will not be due to the need of adapting these frameworks to the context of the study.

Theoretically speaking, there exist important similarities and differences between the two frameworks. For example, modeling (DMEE) is intended to provide students with transferable heuristics that can help them move beyond just solving problems in a single lesson. Similarly, the task cognitive demands pertain to structuring a challenging environment for students that can help them develop mathematical reasoning. At the same time, these TAs are distinct in that the first denotes capturing the development of transferable strategies, whereas the second focuses on whether the tasks enacted in the lesson provide students with opportunities to

engage in rich mathematical practices. Consider, also the qualitative characteristics (focus and quality) of questioning and the classroom as a learning environment (DMEE): compared to the more quantitative characteristics (frequency and stage) of these factors, the qualitative characteristics both pertain to providing students with substantive opportunities to interact with the content, through the provision of constructive feedback. The remediation of students' errors in MQI also attends to such opportunities by exploring how the teacher works with students' errors to help them develop mathematical understanding. However, although both frameworks are attending to the feedback provided to students, they consider this aspect from different perspectives: DMEE considers more general characteristics of feedback providing, whereas MQI attends to more mathematical features. Such examples suggest that there are reasons to believe that TAs of DMEE and MQI may co-exist on a single scale. Given that aforementioned DMEE and MQI TAs capture similar manifestations of teacher-student interactions *yet from different perspectives* (as suggested by the two preceding examples), it is likely that these TAs will co-appear in clusters combining generic and specific TAs. To the extent that this holds, it would be informative to examine which TAs from the two frameworks are clustered together and what might be driving their co-existence in the same cluster.

Apart from the similarities identified above, there are important differences between the two frameworks, which, however, suggest that the frameworks may complement each other. For example, whereas MQI does not have any aspect related to orienting students to the importance of what is to be learned, such aspects are covered in DMEE; on the other hand, whereas DMEE does not attend to the mathematical content offered to the students, this is the main focus of the MQI. Similarly, whereas structuring (DMEE) captures connections among the different lesson goals and activities without attending to the mathematical substance of such connections, linking and connections (MQI) pertains to explicitly drawing mathematical connections between different representations and different mathematical ideas. These complementarities also highlight the importance for exploring possibilities of integrating the two frameworks, given that such an integration might show how one framework accounts for the limitations of the other.

In this paper, we explore the possibility of such integration by attempting to develop a common scale that combines subject-specific and subject generic TAs. We use the term "integrate" to denote the combination "of two or more things in order to become more effective" (https://dictionary.cambridge.org/), since our intention is to examine whether the TAs of the two frameworks could be combined into a more functional and comprehensive whole. If such a scale cannot be developed, this would suggest that such an integration is not possible. However, developing such a scale, in and of its own is not sufficient for arguing about such an integration: if the generic TAs are all clustered in certain levels and specific TAs in other distinct levels, the scale would not empirically corroborate the type of integration proposed above, given that the two types of TAs would still be distinct from each other. In this paper we are interested in exploring the possibility of developing a common scale with TAs of both frameworks distributed and mixed all over the continuum.

3 Research questions

This study aimed at addressing the following questions:

1. Can a scale with good psychometric properties that combines the TAs of DMEE and MQI be developed?
2. If such a scale can be developed, can we identify levels of effective teaching that include both subject-generic and subject-specific TAs?

Developing a scale including both types of TAs would provide empirical evidence attesting to the possibility and importance of integrating generic and specific TAs as opposed to simply juxtaposing them. Given the exploratory character of this study, this will also give the opportunity to reflect on what might be driving the co-appearance of TAs from both frameworks in the same cluster, something that could provide important insights about what it means to integrate subject-generic and subject-specific aspects and possible ways of developing frameworks that combine both aspects.

4 Methods

4.1 Participants and setting

Thirty-eight elementary Cypriot school teachers participated voluntarily in the study (see Table 1 for demographic information). Each teacher had six of their mathematics lessons videotaped. Teachers were free to choose which lessons to have videotaped. Teachers' self-selection of recording dates should not be a concern, given prior research suggesting that when teachers were given discretion to choose from among a set of their classroom videos for evaluative purposes, the ranking of teachers in terms of the teaching quality in the chosen videos was similar to the ranking from a random set of videotaped lessons (Ho & Kane, 2013). Each lesson was videotaped by a single camera placed at the back

Table 1 Teacher demographic characteristics

Characteristic	Value
Female	0.63
Master's degree	0.71
Fourth-grade teachers	0.34
Fifth-grade teachers	0.32
Sixth-grade teachers	0.34
Years of experience (mean and SD)	14.89 (5.27)
Years of experience in teaching mathematics (mean and SD)	13.66 (5.86)

Note All values represent proportions unless other units are indicated

of the classroom; students whose parents did not give consent for participation were placed outside of the videotaped cone but participated in the lesson. Ethics permission for the study was obtained from the National Centre of Educational Research (ethics approval number blinded). Lessons averaged about an hour.

4.2 Data coding

Each lesson was coded using both the DMEE and the MQI by raters trained in either of the two frameworks. Training lasted approximately 20 h for each framework and raters were certified only when their ratings were consistent with at least 80% of the ratings of master raters to selected videotaped lesson excerpts. For both DMEE and MQI, each lesson was coded by a pair of raters (N_{DMEE}=3, N_{MQI}=3). The raters (different for DMEE and MQI) first coded the lessons individually and then met to reconcile their ratings. Each rater coded 152 lessons, two from each teacher; all possible pairs of coders were formed and each pair coded two lessons from each teacher. For the purposes of this analysis, we used their reconciled codes that represent the pair's consensus on coding the lesson. Inter-rater reliability *before* reconciliation were higher than 0.70 for both DMEE and MQI.

Raters were asked to complete both a low-inference and a high-inference instrument. Because of differences in the coding procedures utilized for the low-inference DMEE and MQI instruments, we utilized the two high-inference instruments. The DMEE high-inference instrument measures all eight TAs but assessment. Observers were expected to complete a Likert scale comprising 34 items at the end of each lesson to indicate how often each teacher behavior was observed (for examples, see Appendix A1). The MQI high-inference instrument was developed taking into consideration the segment-level MQI codes which were transferred at the lesson level: at the end of each lesson, the raters were asked to evaluate the mathematical quality of teaching for the MQI codes using a scale from 0 (not at all) to 4 (to a great extent) (see Appendix A2 for an example of this scale). At the time the study was conducted, the existing MQI version included six codes for *Richness* (see Fig. 2, except for c), four codes of *Errors and Imprecision*, three codes of *Working with students and mathematics*, and four codes for *Common Core-Aligned Student Practices* (except for c and e).

4.3 Data analysis

Descriptive analysis was conducted to identify TAs suffering from ceiling and/or floor effects. Two items on the differentiation dimension of two DMEE TAs (i.e., structuring and dealing with misbehavior) had to be excluded since they were infrequently observed (1.31%). Similarly, all the four *Errors and Imprecision* codes of the MQI were dropped because they appeared very infrequently (e.g., 1.7% for mathematical errors).

The Extended Logistic model of Rasch (Andrich, 1988) was first used separately for each framework, to examine whether its corresponding data could form a scale measuring its respective TAs. We treated each lesson separately meaning that 228 person estimates (i.e., 38 teachers X 6 lessons) were generated by using Quest (Adams & Khoo, 1996). The Rasch model is appropriate for the specification of such scales because it enables testing whether the data meets the requirement that both teachers' lesson performance on the framework items and the difficulties of the items form a stable sequence (within probabilistic constraints) along a single continuum (Bond & Fox, 2001).

Once developing two separate scales, the Rasch model was then utilized to find out whether a common scale can be established (we also checked whether the data had better fit to more complex models, such as a multidimensional IRT scale; see more on these analyses in Appendix B1). Having established the reliability of the common scale, we then employed the procedure for detecting pattern clustering in measurement designs developed by Marcoulides and Drezner (1999) to examine if the various TAs are systematically grouped by difficulty level (see more in Appendix B2). This procedure enables segmenting the observed measurements into constituent groups (or clusters) so that the members of any group are similar to each other, according to the selected criterion (i.e., the difficulty level of each item).

The Rasch model and the clustering method cannot provide answers on how deep the divide is separating the levels of the cluster analysis. Wilson (1989) developed a variant of the Rasch model, the so-called Saltus model, as a method that can differentiate between different levels[3]. Specifically,

[3] According to the Oxford English Dictionary, by level we imply "a position on an imaginary scale in respect to a given amount" (in this case the difficulties of the TAs). In this study, the levels are empirically formed by considering the difficulty level of the TAs considered. The levels are formed by grouping together those TAs with similar

Table 2 The psychometric properties of the three scales developed for the DMEE and MQI separately and their combination

Statistic	DMEE (L=32)	MQI (L=10)	DMEE & MQI (L=40)
Mean			
Item	0.00	0.00	0.00
Cases	0.14	0.09	0.13
Standard deviation			
Item	1.10	0.71	0.98
Cases	0.69	1.40	0.78
Separability			
Item	0.87	0.79	0.87
Cases	0.82	0.83	0.86
Mean Infit Mean Square			
Item	1.00	1.00	0.99
Cases	1.02	1.01	1.01
Mean Outfit Mean Square			
Item	1.02	1.03	0.99
Cases	1.01	1.03	1.00
Infit t			
Item	−0.03	−0.08	−0.02
Cases	0.04	0.01	0.02
Outfit t			
Item	0.05	−0.03	−0.01
Cases	−0.01	0.08	−0.01

the Saltus model allows to differentiate between major and less pervasive changes in moving from one level to the other without sacrificing the idea of one common underlying continuum. Readers interested in the technical details of this model are referred to Appendix B3. Thus, the Saltus model was used to differentiate between major and less pervasive changes in moving from one level to the other without sacrificing the idea of one common underlying continuum.

5 Results

5.1 Developing a common scale of teaching quality

The Rasch model was used to analyze teachers' performance on the 32 DMEE items *and* the 10 MQI items. After dropping two DMEE items on the focus dimension of orientation and the stage dimension of dealing with misbehavior which did not fit the model, the remaining 40 items of DMEE and MQI fit the model well (see Table 2, Column 4). Specifically, all TAs had item infit within the range 0.81 up to 1.22, and item outfit within the range of 0.77 up to 1.19 (see Appendix C). Moreover, the results of this analysis revealed that these TAs were well targeted against the teachers' lesson performance since their scores ranged from −1.76 to 2.04 logits and the difficulties of the 40 items (i.e., TAs) ranged from −1.88 to 2.36 logits. Furthermore, the indices of the separation of persons and TAs were higher than 0.85, suggesting satisfactory scale reliability (Bond & Fox, 2001). Finally, Yen's (1993) procedure (see Appendix B1) was used to test for Local independence, a central assumption of Item Response Theory models; violations of this assumption are usually tested using test statistics based on item pairs (Debelak & Koller, 2020). Local independence was not violated. Collectively, these findings suggested that subject-generic and subject-specific TAs could form a single scale with good psychometric properties (for how the fitting of the Rasch model compared to alternative more complex IRT models, see Appendix B1).

5.2 Developing levels of teaching quality

The application of Marcoulides and Drezner's (1999) analysis to cluster the 40 TAs based on their Rasch item difficulties showed that they could be optimally organized into five groups (i.e., levels of teaching, see Table 3). The cumulative D for the five-cluster solution was 41% whereas the sixth gap added only 3.9% and the seventh gap added even less (3.2%). Given that explaining at least 40% of the observed variance is considered satisfactory in cluster analysis (Romesburg, 1984), we further examined this solution using the Saltus model.

To apply the Saltus model, we assumed that the 40 TAs are structured into the five groups of the cluster analysis. The Saltus solution had better fit to the actual data than the Rasch model and offered a statistically significant improvement over the Rasch model which was equal to 1087 chi-square units at the cost of 30 additional parameters; this solution was also found to fit better to the data compared to fitting saltus with a set of alternative solutions with fewer or more clusters (see Appendix B2). Table 3 presents the Rasch difficulty parameters of the 40 TAs along with the Saltus difficulty parameters, starting from the easiest level (i.e., Level 1 shown in Column 3) and moving up to the most difficult level (i.e., Level 5, Column 7). A comparison of the Rasch and Saltus parameters and a justification of the choice of the latter over the former are presented in Appendix B2. Based on this comparison it can be claimed that the spectrum of TAs measured through the DMEE and MQI is discontinuous rather than continuous. A description of the different levels is given below.

A first observation on how the DMEE and MQI TAs have been clustered into levels is that, except for the fifth level (i.e., the most demanding one), all other levels include both subject-generic and subject-specific TAs. A second observation is that there is not only conceptual homogeneity within

difficulty. The benefit of forming such levels is that it allows clustering TAs into groups. To the extent that the clustered TAs within a group represent homogeneous TAs, these groups can point to certain teacher professional development needs.

Table 3 Rasch and saltus parameter estimates for factor scores measuring the teaching aspects of the DMEE and the MQI

Teaching aspects	Rasch	Implied within-level difficulty (Saltus)				
	All	Level 1	Level 2	Level 3	Level 4	Level 5
Level 1: Structuring a basic learning environment						
Dealing with misbehavior -frequency	-1.88	-2.92	-2.92	-2.92	-2.92	-2.92
Questioning frequency	-1.63	-2.86	-2.86	-2.86	-2.86	-2.86
Application frequency	-1.54	-2.77	-2.77	-2.77	-2.77	-2.77
Management of time frequency	-1.53	-2.68	-2.68	-2.68	-2.68	-2.68
Structuring frequency	-1.36	-2.59	-2.59	-2.59	-2.59	-2.59
Application stage	-1.35	-2.51	-2.51	-2.51	-2.51	-2.51
Structuring stage	-1.10	-2.48	-2.48	-2.48	-2.48	-2.48
WSM: using students' contributions	-1.09	**-2.38**	*-2.38*	-2.38	-2.38	-2.38
Level 2: Providing opportunities for active student engagement						
Questioning focus	-0.88	**-1.52**	*-2.29*	-2.31	-2.21	-2.36
Questioning quality	-0.75	-1.41	-2.21	-2.28	-2.30	-2.32
CLE – Interactions frequency	-0.66	-1.34	-2.11	-2.17	-2.19	-2.28
Questioning stage	-0.29	-1.29	-2.18	-2.24	-2.17	-2.21
Dealing with misbehavior quality	-0.29	-1.21	-2.05	-2.12	-2.11	-2.18
CCASP: student explanations	-0.27	-1.18	-2.01	-2.07	-2.03	-2.05
CCASP: questioning/reasoning	-0.21	-1.12	-1.91	-2.01	-1.93	-2.12
Application quality	-0.18	-1.06	-1.94	-1.91	-1.88	-1.91
RM: mathematical language	-0.16	-0.99	-1.85	-1.82	-1.81	-1.85
CLE – Interactions stage	-0.13	-0.88	-1.81	-1.77	-1.72	-1.77
CLE – Interactions focus	0.01	-0.82	**-1.76**	*-1.74*	-1.75	-1.72
Level 3: Building a (mathematically) rich learning environment						
RM: Linking and connections	0.09	0.22	**-0.98**	*-1.71*	-1.72	-1.69
Structuring Focus	0.10	0.27	-0.77	-1.66	-1.67	-1.66
CLE – Interactions quality	0.12	0.38	-0.73	-1.63	-1.61	-1.61
CLE – Interactions differentiation	0.21	0.44	-0.68	-1.55	-1.54	-1.55
WSM: remediation	0.23	0.48	-0.59	-1.51	-1.49	-1.49
RM: Mathematical explanations	0.27	0.54	-0.51	-1.46	-1.44	-1.41
Structuring Quality	0.29	0.59	-0.45	-1.42	-1.40	-1.37
Orientation frequency	0.31	0.66	-0.41	-1.38	-1.37	-1.34
RM: overall	0.38	0.73	-0.33	-1.31	-1.32	-1.31
Modeling frequency	0.53	0.79	-0.36	**-1.28**	*-1.25*	-1.28
Level 4: Engaging students in demanding tasks						
Orientation Stage	0.66	1.92	0.52	**-0.47**	*-1.23*	-1.26
CCASP: cognitive demands	0.68	1.93	0.61	-0.42	-1.17	-1.22
CCASP: overall	0.74	1.96	0.58	-0.40	-1.11	-1.20
Modeling focus	0.74	2.06	0.66	-0.32	-1.06	-1.18
Modeling stage	0.84	2.14	0.69	-0.27	-1.01	-1.14
Questioning quality	0.96	2.21	0.72	-0.19	**-0.92**	*-1.05*
Level 5: Differentiating teaching						
Application differentiation	1.21	3.12	1.75	0.81	**0.18**	*-1.03*
Orientation differentiation	1.23	3.23	1.79	0.88	0.21	-0.95
Questioning differentiation	1.46	3.31	1.84	0.92	0.30	-0.91
Modeling quality	1.88	3.39	1.90	0.99	0.38	-0.88
Modeling differentiation	2.36	3.44	1.95	1.03	0.47	-0.84

Notes. For DMEE: CLE: Classroom as a Learning Environment; MQI: RM: Richness of Mathematics; WSM: Working with students and mathematics; CCASP: Common Core Aligned Student Practices

the TAs of each level, but there also seems to be a progression in how the TAs are organized into levels, starting from TAs that impose fewer demands on teachers and moving up to TAs that impose increasingly more demands. We unpack this argument while describing each level.

A common threat linking the subject-generic and the subject-specific TAs of the first level is that both pertain to structuring *a basic learning environment* rather than providing more quality opportunities for student learning. This level includes subject-generic TAs related to the quantitative

characteristics (i.e., frequency) of factors associated with the direct teaching approach such as structuring and application. The subject-specific TA of this level also pertains to providing a basic learning environment, since it expects the teacher largely to acknowledge and respond to students' contributions.

The TAs of the second level relate to providing students with opportunities to *actively engage in learning*, interacting with their classmates and the content. In terms of the subject-generic TAs, the second level concerns qualitative characteristics of the two aspects of the classroom learning environment factor (i.e., encouraging interactions and dealing with misbehavior) and the appropriate use of questioning, including the provision of constructive feedback. The subject-specific TAs of this level largely concern the opportunities that the teacher provides to students to offer explanations, raise questions, offer examples, and engage in reasoning.

The generic and specific TAs of the third level place even more demands on teachers to build an environment that does not only support active student participation, but it is (mathematically) rich. This level includes subject-generic TAs that pertain only to the quantitative characteristics of factors associated with constructivist approaches (i.e., orientation and modeling). Unlike the specific TAs of the previous level that largely relate to providing students with opportunities for engagement with the content, the specific TAs of this level put more demands on teachers, since they expect them to structure mathematically rich environments through the provision of explanations, the linking of representations, and the identification and remediation of students' errors.

Further increasing the challenge, the types of TAs included in the fourth level expect teachers to not only offer a rich environment but afford students challenging learning opportunities that can have an impact on both cognitive and meta-cognitive learning outcomes. This level includes generic TAs that relate to the qualitative characteristics of factors associated with constructive teaching. Compared to the generic TAs included in the previous level, those of this level do not capture only the provision of related opportunities but require that these opportunities are suitable for students and promote learning. The specific TAs stipulate that the teacher selects and enacts challenging tasks, thus providing students not only with rich mathematical experiences, but also experiences with demanding content.

The fifth level places the most demands on teachers by expecting them to differentiate their teaching to meet different student needs. This level includes only generic TAs related to the differentiation dimension of the DMEE and reveals that differentiation of teaching is very challenging for primary mathematics teachers.

6 Discussion

Drawing on two widely used and validated frameworks, the subject-generic DMEE and the subject-specific MQI, this study showed that it is possible to develop a common scale with good psychometric properties which encompasses both generic and specific TAs. This finding, which is in line with a basic assumption of the DMEE supporting that TAs are interrelated (Kyriakides et al., 2020), implies that both subject-generic and subject-specific aspects can be thought to form an overarching construct, namely teaching quality. The study also provides further support to those indicating the limitations of using exclusively either generic TAs (e.g., Panayiotou et al., 2021) or specific TAs (e.g., Litke et al., 2021) to describe teaching quality.

In arguing about the importance of forming a common scale of TAs, we acknowledge certain limitations. First, our sample consisted only of primary teachers. Although teaching mathematics in primary grades requires strong mathematical knowledge for teaching (cf. Ball et al., 2008), replication studies are needed with secondary school teachers. Second, during the analysis 11 items were dropped. Dropping nine of these items (e.g., misbehavior from DMEE and *Errors and Imprecision* from MQI) was unsurprising, given that they were infrequently observed (apparently due to the self-selected sample of the study and that the lessons were videotaped); had these items been observed more frequently they could have been included in the scale formed—an open issue to address in future research. Two additional items not necessarily related to the teacher sample and mode of lesson observation were also dropped (see Sect. 5.1). We argue, however, that doing so might be unavoidable when trying to develop a common scale measuring both types of TAs, as long as the main constructs of each framework are well represented in the common scale; interestingly, leaving these two items in and running more complex IRT models (see Appendix B1) resulted in much worse fit indices. Despite having to drop these 11 items, it is important to note that all DMEE TAs (factors and dimensions) and the three main MQI dimensions (which form the overarching dimension of *Ambitious Teaching*) were represented in the scale formed.

Another important study finding was that the subject-generic and subject-specific TAs of the two frameworks could be optimally clustered in five distinct levels. Although the grouping of generic TAs into levels has already been demonstrated (Kyriakides et al., 2020), this was the first time to search for grouping both generic and specific TAs in levels. Equally important, except for the last level, all preceding levels included both subject-generic and subject-specific TAs. This implies that the quality of the lessons observed appears to be a function of both generic and specific TAs;

hence, it seems unlikely to have lessons that excel in one type of TAs and are particularly poor in the other type.

The development of the common scale and the establishment of levels combining subject-generic and subject-specific TAs have important theoretical implications. Although one approach for integrating both types of TAs would be to start from developing a theoretical model that integrates both (cf. Charalambous & Praetorius, 2020), this study proposes and examines another approach: starting from existing frameworks that have been considered to fall into two different clusters (generic vs. specific) and exploring the potential of integrating them not by juxtaposing them, but by investigating which aspects of them can co-exist in a single level and whether this co-existence provides a meaningful description of teaching quality. To further explore this line of integration advanced herein more work is needed, including utilizing other subject-generic and subject-specific frameworks, other grade levels (e.g., lower elementary, secondary), and different educational contexts. Finding levels of teaching quality through a cross-sectional study has certain limitations (Kyriakides et al., 2009). Therefore, the levels formed cannot be considered developmental. To develop such levels, longitudinal studies are needed to figure out whether some teachers move from one level to the other in a stepwise manner and/or other teachers remain at the same level. Such studies could provide empirical evidence on the extent to which these levels can define certain professional needs for particular groups of teachers, which, in turn, can be used for designing professional development programs. Future studies could also employ more interpretive work (e.g., Bookmark method), to examine the interpretive validity of the levels formed and how practitioners themselves understand the similarities and differences of the items clustered within each level.

Despite these limitations, at least two initial insights could be gleaned. First, the clustering of the 40 TAs into distinct levels calls for more critical examination of what unites aspects that are called generic or specific. One possibility could be that this co-existence of subject-generic and subject-specific aspects in the same levels challenges the very notion of distinguishing between these two types of TAs and calls for considering them as a unified whole. Another could be that these TAs are indeed distinct given that different disciplines impose different demands on teachers, thus rendering certain TAs more specific than others. More thoroughly examining these possibilities both theoretically and empirically represents an important open issue. At the same time, another question is equally important: If these TAs are indeed distinct, what can explain their co-appearance in the same levels? Is it just a statistical artifact or are there more fundamental similarities among them? We argue that unpacking *the demands (in terms of knowledge and practice) that different TAs* (identified as generic or specific) *impose on teachers* represents a productive path for addressing this question. Such an analysis may have implications for teacher initial and ongoing professional development. Second, to the extent that such levels are replicated in future studies, we could identify if the teachers' lessons are consistently clustered into levels and the extent to which they relate to student learning gains. This could indicate the importance of treating such levels as a heuristic for identifying teachers' needs to better support their students' learning.

The identification of the five levels in this study is in line with the use of stage models in teacher professional development (e.g., Berliner, 1994; Sternberg et al., 2000). Specifically, the five levels illustrate not only the complex nature of teaching quality but could help develop specific teacher education courses, considering the needs of each teacher group situated at different levels. Given that subject-generic and subject-specific TAs were integrated in most of the levels emerging, the study findings provide teacher educators with ideas as to how the two types of TAs can be integrated in initial and ongoing teacher education.

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/s11858-024-01591-x.

Funding Open access funding provided by the Cyprus Libraries Consortium (CLC).

References

Adams, R. J., & Khoo, S. (1996). *Quest: The interactive test analysis system, Version 2.1*. ACER.

Andrich, D. (1988). A general form of Rasch's extended logistic model for partial credit scoring. *Applied Measurement in Education*, *1*(4), 363–378. https://doi.org/10.1207/s15324818ame0104_7.

Ball, D. L., Thames, M. H., & Phelps, G. (2008). Content knowledge for teaching: What makes it special? *Journal of Teacher Education*, *59*(5), 389–407. https://doi.org/10.1177/0022487108324554.

Berlin, R., & Cohen, J. (2018). Understanding instructional quality through a relational lens. *Zdm*, *50*(3), 367–379. https://doi.org/10.1007/s11858-018-0940-6.

Berliner, D. (1994). Expertise: The wonder of exemplary performances. In J. Mangieri, & C. Block (Eds.), *Creating powerful*

thinking in teachers and students: Diverse perspectives (pp. 161–186). Harcourt Brace College.

Blazar, D. (2015). Effective teaching in elementary mathematics: Identifying classroom practices that support student achievement. *Economics of Education Review, 48,* 16–29. https://doi.org/10.1016/j.econedurev.2015.05.005

Blazar, D., & Archer, C. (2020). Teaching to support students with diverse academic needs. *Educational Researcher*, *49*(5), 297–311. https://doi.org/10.3102/0013189X20931226.

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.

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.

Blömeke, S., Jentsch, A., Ross, N., Kaiser, G., & König, J. (2022). Opening up the black box: Teacher competence, instructional quality, and students' learning progress. *Learning and Instruction*, *79*, 101600. https://doi.org/10.1016/j.learninstruc.2022.101600.

Bodroža, B., Teodorović, J., & Jošić, S. (2022). Validation of scales for measuring factors of teaching quality from the dynamic model of Educational Effectiveness. *Psihologija*, *55*(2), 169–190. https://doi.org/10.2298/PSI200915010B.

Bond, T. G., & Fox, C. M. (2001). *Applying the Rasch model: Fundamental measurement in the human sciences*. Lawrence Erlbaum Associates.

Boston, M. D., & Candela, A. G. (218). The Instructional Quality Assessment as a tool for reflecting on instructional practice. *ZDM Mathematics Education, 50*(3), 427–444. https://doi.org/10.1007/s11858-018-0916-6.

Brophy, J., & Good, T. (1986). Teacher behavior and student achievement. In M. C. Wittrock (Ed.), *Handbook of research on teaching* (3rd ed.). McMillan.

Brunner, E. (2018). Qualität Von mathematikunterricht: Eine Frage Der Perspektive. *Journal för Mathematik-Didaktik*, *39*(2), 257–284. https://doi.org/10.1007/s13138-017-0122-z.

Campbell, R. J., Kyriakides, L., Muijs, R. D., & Robinson, W. (2003). Differential teacher effectiveness: Towards a model for research and teacher appraisal. *Oxford Review of Education*, *29*(3), 347–362. https://www.jstor.org/stable/3595446.

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. https://www.journals.uchicago.edu/doi/.

Charalambous, C. Y., & Litke, E. (2018). Studying instructional quality by using a content-specific lens: The case of the Mathematical Quality of instruction framework. *Zdm*, *50*(3), 445–460. https://doi.org/10.1007/s11858-018-0913-9.

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

Charalambous, C. Y., & Praetorius, A. K. (2020). Creating a forum for researching teaching and its quality more synergistically. *Studies in Educational Evaluation*, *67*, 100894. https://doi.org/10.1016/j.stueduc.2020.100894.

Chaudhary, P., & Singh, R. K. (2022). A meta analysis of factors affecting teaching and student learning in higher education. *Frontiers in Education*. https://doi.org/10.3389/feduc.2021.824504. *6*.

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

Cohen, D., Raudenbush, S., & Ball, D. (2003). Resources, instruction, and research. *Educational Evaluation and Policy Analysis*, *25*(2), 1–24. https://doi.org/10.3102/01623737025002119.

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

Creemers, B. P. M., & Kyriakides, L. (2008). *The dynamics of educational effectiveness: A contribution to policy, practice and theory in contemporary schools*. Routledge. https://doi.org/10.4324/9780203939185.

Creemers, B. P. M., & Kyriakides, L. (2012). *Improving quality in education: Dynamic approaches to school improvement*. Routledge. https://doi.org/10.4324/9780203817537.

Creemers, B. P. M., Kyriakides, L., & Antoniou, P. (2013). *Teacher professional development for improving quality in teaching*. Springer. https://link.springer.com/book/10.1007/978-94-007-5207-8.

Debelak, R., & Koller, I. (2020). Testing the local Independence Assumption of the Rasch Model with Q3-Based nonparametric model tests. *Applied Psychological Measurement*, *44*(2), 103–117. https://doi.org/10.1177/0146621619835501.

Dierendonck, C. (2023). Measuring the classroom level of the dynamic model of Educational Effectiveness through teacher self-report: Development and validation of a new instrument. *Frontiers in Education*, 8. https://doi.org/10.3389/feduc.2023.1281431.

Hamre, B. K., Pianta, R. C., Downer, J. T., DeCoster, J., Mashburn, A. J., Jones, S. M., Brown, J., Cappella, E., Atkins, M., Rivers, S. E., Brackett, M. A., & Hamagam, A. (2013). Teaching through interactions: Testing a developmental framework of teacher effectiveness in over 4,000 classrooms. *The Elementary School Journal*, *113*(4), 461–487. https://doi.org/10.1086/669616.

Hill, H. C. (2010). *The Mathematical Quality of Instruction: Learning Mathematics for Teaching* Paper presented at the 2010 annual meeting of the American Educational Research Association, Denver, CO.

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*, 430–511. https://doi.org/10.1080/07370000802177235.

Hill, H. C., Kapitula, L., & Umland, K. (2011). A validity argument approach to evaluating teacher value-added scores. *American Educational Research Journal*, *48*(3), 794–831. https://doi.org/10.3102/0002831210387916.

Hill, H. C., Umland, K., Litke, E., & Kapitula, L. R. (2012). Teacher quality and quality teaching: Examining the relationship of a teacher assessment to practice. *American Journal of Education*, *118*(4), 489–519. https://doi.org/10.1086/666380.

Ho, A. D., & Kane, T. J. (2013). The reliability of classroom observations by school personnel. Bill & Melinda Gates Foundation. Retrieved from http://k12education.gatesfoundation.org/download/?Num=2520&filename=MET_Reliability-of-Classroom-Observations_Research-Paper.pdf

Joyce, B., Weil, M., & Calhoun, E. (2000). *Models of teaching*. Allyn & Bacon.

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 May 2013.

Kelcey, B., Hill, H. C., & Chin, M. J. (2019). Teacher mathematical knowledge, instructional quality, and student outcomes: A multilevel quantile mediation analysis. *School Effectiveness and School Improvement*, *30*(4), 398–431. https://doi.org/10.1080/09243453.2019.1570944.

Kraft, M. A., & Hill, H. C. (2020). Developing ambitious mathematics instruction through web-based coaching: A randomized field trial. *American Educational Research Journal*, *57*(6), 2378–2414. https://doi.org/10.3102/0002831220916840.

Kyriakides, L., & Creemers, B. P. M. (2008). Using a multidimensional approach to measure the impact of classroom level factors upon student achievement: A study testing the validity of the dynamic model. *School Effectiveness and School Improvement*, *19*(2), 183–205. https://doi.org/10.1080/09243450802047873.

Kyriakides, L., Creemers, B. P. M., & Antoniou, P. (2009). Teacher behaviour and student outcomes: Suggestions for research on teacher training and professional development. *Teaching and Teacher Education*, *25*(1), 12–23. https://doi.org/10.1016/j.tate.2008.06.001.

Kyriakides, L., Creemers, B. P. M., Panayiotou, A., & Charalambous, E. (2020). *Quality and equity in education: Revisiting theory and research on educational effectiveness and improvement*. Routledge. https://doi.org/10.4324/9780203732250.

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

Lee, J., & Santagata, R. (2020). A longitudinal study of novice primary school teachers' knowledge and quality of mathematics instruction. *ZDM-Mathematics Education*, *52*, 295–309. https://doi.org/10.1007/s11858-019-01123-y.

Litke, E., Boston, M., & Walkowiak, T. A. (2021). Affordances and constraints of mathematics-specific observation frameworks and general elements of teaching quality. *Studies in Educational Evaluation*, *68*, 100956. https://doi.org/10.1016/j.stueduc.2020.100956.

Liu, J., & Jiang, Z. (2018). The synergy theory of economic growth. In J. Liu & Z. Jiang (Eds.), *The synergy theory on economic growth: Comparative study between China and developed countries* (pp. 57–90). Springer. https://link.springer.com/book/https://doi.org/10.1007/978-981-13-1885-6.

Mantzicopoulos, P., French, B. F., & Patrick, H. (2019). The quality of mathematics instruction in kindergarten: Associations with students' achievement and motivation. *Elementary School Journal*, *119*(4), 651–676. https://doi.org/10.1086/703176.

Marcoulides, G. A., & Drezner, Z. (1999). A procedure for detecting pattern clustering in measurement designs. In M. Wilson, & G. Engelhard, Jr. (Eds.), *Objective measurement: Theory into practice (Vol. 5)* Ablex Publishing Corporation.

Mu, J., Bayrak, A., & Ufer, S. (2022). Conceptualizing and measuring instructional quality in mathematics education: A systematic literature review. *Frontiers in Education*, *7*. https://doi.org/10.3389/feduc.2022.994739.

Panayiotou, A., Herbert, B., Sammons, P., & Kyriakides, L. (2021). Conceptualizing and exploring the quality of teaching using generic frameworks: A way forward. *Studies in Educational Evaluation*, *70*, 101028. https://doi.org/10.1016/j.stueduc.2021.101028.

Polymeropoulou, V., & Lazaridou, A. (2022). Quality teaching: Finding the factors that foster student performance in junior high school classrooms. *Education Sciences*, *12*(5), 1–20. https://doi.org/10.3390/educsci12050327.

Praetorius, A. K., & Charalambous, C. Y. (2018). Classroom observation frameworks for studying instructional quality: Looking back and looking forward. *Zdm*, *50*(3), 535–553. https://doi.org/10.1007/s11858-018-0946-0.

Praetorius, A. K., Klieme, E., Herbert, B., & Pinger, P. (2018). Generic dimensions of teaching quality: The German framework of three Basic dimensions. *Zdm*, *50*(3), 407–426. https://doi.org/10.1007/s11858-018-0918-4.

Romesburg, H. C. (1984). *Cluster analysis for researchers*. Lifetime Learning Publication and Wadsworth Inc.

Santagata, R., & Lee, J. (2021). Mathematical knowledge for teaching and the mathematical quality of instruction: A study of novice elementary school teachers. *Journal of Mathematics Teacher Education*, *24*(1), 33–60. https://doi.org/10.1007/s10857-019-09447-y.

Scheerens, J. (2013). The use of theory in school effectiveness research revisited. *School Effectiveness and School Improvement*, *24*(1), 1–38. https://doi.org/10.1080/09243453.2012.691100.

Schlesinger, L., Jentsch, A., Kaiser, G., König, J., & Blömeke, S. (2018). Subject-specific characteristics of instructional quality in mathematics education. *Zdm*, *50*(3), 475–490. https://doi.org/10.1007/s11858-018-0917-5.

Schoenfeld, A. H. (1998). Toward a theory of teaching in context. *Issues in Education*, *4*(1), 1–94. https://doi.org/10.1016/S1080-9724(99)80076-7.

Schoenfeld, A. H. (2018). Video analyses for research and professional development: The teaching for robust understanding (TRU) framework. *Zdm*, *50*(3), 491–506. https://doi.org/10.1007/s11858-017-0908-y.

Sternberg, R. J., Forsythe, G. B., Hedlund, J., Horvath, J. A., Wagner, R. K., Williams, W. M., Snook, S. A., & Grigorenko, E. L. (2000). *Practical intelligence in everyday life*. Cambridge University Press.

Stringfield, S. C., & Slavin, R. E. (1992). A hierarchical longitudinal model for elementary school effects. In B. P. M. Creemers, & G. J. Reezigt (Eds.), *Evaluation of educational effectiveness* (pp. 35–69). ICO.

Tomlinson, C. A. (2014). *The Differentiated Classroom: Responding to the needs of all Learners* (2nd Ed.).). ASCD.

Walkington, C., & Marder, M. (2018). Using the UTeach Observation Protocol (UTOP) to understand the quality of mathematics instruction. *Zdm*, *50*(3), 507–519. https://doi.org/10.1007/s11858-018-0923-7.

Walkowiak, T. A., Berry, R. Q., Pinter, H. H., & Jacobson, E. D. (2018). Utilizing the M-Scan to measure standards-based mathematics teaching practices: Affordances and limitations. *Zdm*, *50*(3), 461–474. https://doi.org/10.1007/s11858-018-0931-7.

Wilson, M. (1989). Saltus: A psychometric model of discontinuity in cognitive development. *Psychological Bulletin*, *105*(2), 276–289. https://doi.org/10.1037/0033-2909.105.2.276.

Yen, W. (1993). Scaling and performance assessments: Strategies for managing local item dependence. *Journal of Educational Measurement*, *30*, 187–213. https://doi.org/10.1111/j.1745-3984.1993.tb00423.x.