Blikstad-Balas, M., Klette, K. & Tengberg, M. (Eds.) (2021).
Ways of Analyzing Teaching Quality. Potentials and Pitfalls.
Scandinavian University Press.
DOI: https://doi.org/10.18261/9788215045054-2021-05

5. Instructional Quality: A Review of Conceptualizations, Measurement Approaches, and Research Findings

Bas Senden, Trude Nilsen and Sigrid Blömeke

Abstract This chapter reviews and discusses conceptualizations, measurement approaches, and research findings of instructional quality. Although agreement on how to conceptualize and measure instructional quality is rare, some common ground can be found. In addition, research findings indicate that the role of instructional quality for student learning might vary across contexts, hinting towards the importance of differential effectiveness for instructional quality.

Keywords Instructional Quality | Framework synthesis | Operationalization | Student Outcomes | Differential effectiveness

INTRODUCTION

For the last decades, research findings have consistently supported the importance of teachers and their instruction for student learning (e.g., Burroughs et al., 2019; Hattie, 2009; Konstantopoulos & Chung, 2011; Muijs et al., 2014). However, the data also revealed that teachers differ in their influence on student learning, and it is the highly effective teachers that make the difference (Hattie, 2009; Konstantopoulos & Chung, 2011). However, "what characterizes highly effective teachers?" and "how do these teachers create environments through which student learning is enhanced?" are questions that are only partly answered.

In line with these two questions, indicators for teacher quality can be divided into what teachers *bring* to the classroom and what teachers *do* in the classroom. When teachers enter the classroom, they *bring* their competence with them, consisting of teacher characteristics and teacher qualifications (Blömeke et al., 2016;

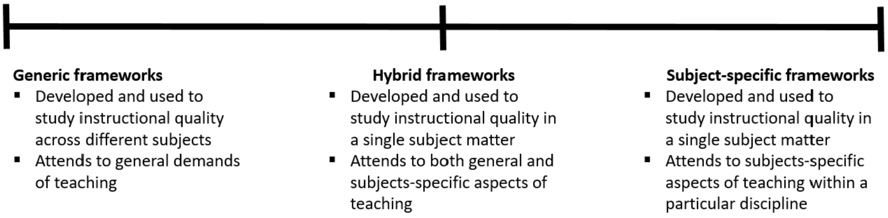
Goe & Stickler, 2008). What teacher *do* in the classroom refers to the classroom practices teachers employ to accomplish specific teaching tasks (Goe & Stickler, 2008). Whereas research findings on the impact of teacher competence on student learning revealed inconsistent results, what teachers *do* in the classroom—their actual teaching—has been more consistently shown to be an essential indicator for students' learning outcomes (Blömeke & Olsen, 2019; Burroughs et al., 2019; Hattie, 2009; Muijs et al., 2014). However, also in this respect, inconsistencies are found regarding which classroom practices are important, for which outcomes, and in which context. Reviewing these inconsistencies is one of the primary purposes of this chapter.

The core components of teaching that are considered indicative of student learning are reflected in the construct of instructional quality (Kunter & Voss, 2013; Nilsen. Gustafsson & Blömeke, 2016). Various frameworks and instruments were developed and used to conceptualize instructional quality and consequently study its influence on student outcomes. However, these frameworks differ considerably in the teaching aspects covered, conceptualizations, operationalizations, and measurement. In addition, findings relating instructional quality to student outcomes vary across countries, student age, subject domains, and type of outcome (Blömeke & Olsen, 2019). These differences have proven problematic for finding consensus on which aspects of instruction support student learning.

Therefore, this study seeks to contribute to disentangling these challenges by, first, discussing the differences between generic and subject-specific aspects of instructional quality and how they contribute to measuring instructional quality, second, evaluating the different ways in which frameworks of instructional quality are conceptualized, third, discussing how instructional quality is operationalized and measured and finally discussing how relations between instructional quality and student outcomes may differ across countries, cohorts, subject domains, and type of outcomes by showcasing findings from studies of some of the authors.

INSTRUCTIONAL QUALITY: FROM GENERIC TO SUBJECT-SPECIFIC

Frameworks developed and used to study instructional quality intend to assess either generic or subject-specific (also named domain or content-specific) aspects of instructional quality or a combination of both (Blömeke et al., 2020). This has led to three general categories of frameworks: (1) generic frameworks, (2) subject-specific frameworks and (3) hybrid frameworks. Charalambous and Praetorius (2018) provided a continuum from generic to subject-specific in which they visualized different frameworks. For the present chapter, we adapted and adjusted the continuum to include three categories (Figure 5.1). The following section will discuss each category placed on the continuum and provide several examples of existing frameworks within that category.



Generic frameworks
- Developed and used to study instructional quality across different subjects
- Attends to general demands of teaching

Hybrid frameworks
- Developed and used to study instructional quality in a single subject matter
- Attends to both general and subjects-specific aspects of teaching

Subject-specific frameworks
- Developed and used to study instructional quality in a single subject matter
- Attends to subjects-specific aspects of teaching within a particular discipline

Figure 5.1. A continuum from generic to subject-specific for the classification of instructional-quality frameworks (adapted from Charalambous & Praetorius, 2018, p. 356).

Generic Frameworks

On the complete left of this continuum, we find the generic frameworks. Generic frameworks include aspects of instructional quality without specifying them specifically for the subject under investigation. Thus, these frameworks address aspects of instructional quality on a more general level as, in essence, a generic framework is developed and used to investigate instructional quality and its effect on learning across different subjects (Charalambous & Praetorius, 2018; E. Kyriakides et al., 2018).

Upon reviewing the hitherto existing generic frameworks, we find that multiple frameworks have been developed in the last two decades. Examples of popular frameworks are the Classroom Assessment Scoring System (CLASS; Pianta et al., 2008), the framework of the Three Basic Dimensions (TBD; Klieme et al., 2001), the Framework for Teaching (FfT; Danielson, 2007), the Tripod 7cs framework (7Cs; Ferguson, 2012) and the Dynamic Model of Educational Effectiveness (DMEE; Creemers & Kyriakides, 2008). Because they intend to generalize aspects of instruction across subjects, these frameworks are widely used for teacher evaluation, research purposes, or international large-scale assessments such as the Program for International Student Assessment (PISA).

However, the extent to which generic frameworks can be used across different subjects is controversial. Researchers have argued that some of the generic frameworks include aspects that might be considered subject-dependent. For example, the dimension of cognitive activation in the framework of the three basic dimensions, and the instructional support dimension in CLASS, are often discussed as being more closely related to subject-specific aspects of instructional quality (Charalambous & Praetorius, 2018; Schlesinger & Jentsch, 2016).

Another critique is that generic frameworks might be generalizable across particular, more related, subjects, but not across others (e.g., Charalambous et al., 2019; Cohen et al., 2018; Praetorius et al., 2016). For example, generic frameworks might cut across German and English (Praetorius et al., 2016), but less so across mathematics and physical education (Charalambous et al., 2019) or mathematics and language arts (Cohen et al., 2018).

Subject-Specific Frameworks

Some researchers have stressed the importance of considering that each subject is unique and requires different teaching qualities (e.g., Charalambous & Kyriakides, 2017; Cohen et al., 2018; Schlesinger & Jentsch, 2016). This has led to the development and use of frameworks that are considered more subject-specific. In contrast to the generic frameworks, these frameworks are developed and used to analyze instructional quality within one specific subject and are informed by the subject-specific demands of teaching within a specific discipline. Therefore, these frameworks contain aspects of instructional quality unique to teaching in a specific subject (Charalambous & Kyriakides, 2017; E. Kyriakides et al., 2018).

Because these frameworks are restricted to one subject, many different frameworks exist within and across specific disciplines. However, most subject-specific frameworks were developed within mathematics education and include aspects such as richness of mathematics or mathematical errors and imprecisions (for a complete overview, see Schlesinger & Jentsch, 2016). Frameworks used in mathematics include, among others: the Mathematical Quality of Instruction framework (MIQ; Learning Mathematics for Teaching Project, 2011), the Instructional Quality Assessment (IQA; Junker et al., 2006) and the Mathematics Scan measure (M-Scan; Berry et al., 2012). Other frameworks are found within science (Carlson et al. 2019) or physical education (E. Kyriakides et al., 2018).

Using Both Generic and Subject-Specific Frameworks

The most straightforward way to include generic and subject-specific aspects of instructional quality is to combine generic and subject-specific frameworks. For example, in physical education (PE), E. Kyriakides et al. (2018) proposed generic and subject-specific aspects of instructional quality that may contribute to student psychomotor learning by combining generic aspects from the Dynamic Model of

Educational Effectiveness (DMEE; Creemers & Kyriakides, 2008) framework with subject-specific teaching practices from a modified version of the Task Structure Systems (TSS; Siedentop et al., 1994). Similarly, the Measures of Effective Teaching (MET) project combined two generic frameworks and two subject-specific frameworks in their investigations of instructional quality (Kane & Staiger, 2012).

In addition, some studies have investigated the commonalities between frameworks and propose to find more common ground across frameworks to cover a broader range of instructional aspects that attend to both general and subject-specific demands (e.g., Blazar et al., 2017; Charalambous & Kyriakides, 2017; Praetorius & Charalambous, 2018). In this context, Praetorius and Charalambous (2018) analyzed and reviewed twelve different frameworks: four generic, three mathematics specific and four hybrid frameworks and proposed a common structure of instructional quality including both generic and subject-specific aspects.

Finally, some frameworks combine both generic and subject-specific frameworks into a single, hybrid framework. For example, the TEDS-Instruct framework from Schlesinger et al. (2018) combines generic aspects of instruction with subject-specific aspects by using the three dimensions of the TBD framework and extending it with subject-specific aspects of instruction in mathematics education found by reviewing several mathematic specific frameworks. Other frameworks, such as the Teaching for Robust Understanding (TRU; Schoenfeld, 2013) framework, were developed as generic frameworks but later included subject-specific dimensions, whereas the UTeach Observation Protocol (UTOP; Walkington & Marder, 2015) mainly assesses subject-specific practices but adds additional generic elements (Charalambous & Praetorius, 2018). Frameworks that combine generic and subject-specific practices would be placed in the middle of the continuum and are referred to as hybrid frameworks (Charalambous & Praetorius, 2018; Praetorius & Charalambous, 2018).

CONCEPTUALIZATIONS OF INSTRUCTIONAL QUALITY

Frameworks for instructional quality do not only vary regarding the type of instructional practices they cover; they are also conceptualized differently. During the development process, top-down or bottom-up approaches, or a combination of both, are used (Praetorius & Charalambous, 2018). A framework developed through a top-down approach builds on existing literature and expert judgment. This approach can also include building on existing frameworks (e.g., the previously mentioned TEDS-Instruct framework). A bottom-up approach, on the other hand, includes watching and analyzing video lessons, undertaking interviews with students/teachers, and conducting exploratory analyses such as exploratory factor analysis. Many existing frameworks use both approaches (Praetorius & Charalambous, 2018). To understand the variety of frameworks that differ; (1) in their dimensionality and; (2) in the depth and breadth to which they cover their dimensions, we elaborate on the process of identifying and conceptualizing key aspects of high-quality instruction.

To this end, this section will highlight several existing frameworks and discuss their background and conceptualizations. Table 5.1 depicts a selection of frameworks and their dimensions and sub-dimensions and will be used as a reference point for the discussion. Included are several frameworks mentioned earlier, namely: three generic frameworks (CLASS, TBD, and the 7C's), one hybrid framework (TEDS-Instruct) and one subject-specific framework (MQI).

Background of the Frameworks

The first framework, CLASS, was introduced at the beginning of the 21st century in the US by Hamre and Pianta (2007) and is defined as: "a standardized observation measure of global classroom equality" (Pianta & Hamre, 2009, p. 111). Pianta and Hamre (2007) proposed a latent structure, assessing three major dimensions, with each dimension housing several more specific sub-dimensions assumed to support student learning outcomes (see Table 5.1 for elaboration on the dimensions and sub-dimensions). Within these sub-dimensions are posited several behavioral indicators reflective of that sub-dimensions (Pianta & Hamre, 2009). The conceptualization of the framework can be traced back to previous literature on classroom teaching and educational effectiveness.

CLASS was initially developed for pre-kindergarten classrooms but has been expanded for classrooms ranging from infants to secondary education (Pianta & Hamre, 2009). The dimensions mentioned in Table 5.1 are from the original Pre-K framework, but slight differences exist between frameworks for different target groups (e.g., CLASS covers fewer dimensions in frameworks assessing younger groups). Several studies, with data from pre-kindergarten to 12th grade, have tested and validated the three-dimensional factor structure (Bihler et al., 2018; Pakarinen et al., 2010; Pianta & Hamre, 2009; Virtanen et al., 2017).

Around the same time, Klieme et al. (2001) introduced the TBD framework in Germany. For the development of the TBD framework, the German national centre for the Trends In Mathematics and Science Study (TIMSS) in 1995 enhanced the study's research design by adapting 21 questionnaire scales developed by Gruehn (2000) to assess students' perception of teaching. The first TIMSS video study complemented the questionnaire scales, and both were followed up one year later, resulting in a full-size longitudinal study in Germany (Praetorius et al., 2020). Clausen (2002) developed a high-inference observation protocol. Exploratory factor analysis was conducted on these high-inferences rating by Klieme et al. (2001), resulting in a clear three-dimensional factor structure. The three basic dimensions (TBD) of instructional quality (see Table 5.1) found in this study were largely similar to those from the CLASS framework. The framework originated within mathematics instructions, but the three dimensions were proposed as generic dimensions of instructional quality. In addition, the measurement of the three basic dimensions was not connected to a single instrument, but different studies developed their own operationalization using different perspective (e.g., student, teacher, or observer) (Praetorius et al., 2018). The operationalization of the TBD in research studies, therefore, varies widely. The extent to which it varies became clearer in a recent study done by Praetorius et al. (2018). The authors provided a comprehensive overview of the framework and identified sub-dimensions based on operationalizations used in previous studies. To this end, they identified four sub-dimensions for classroom management, ten for student support, and seven for cognitive activation (see Table 5.1 for a complete overview). Thus, while the TBD framework allows researchers a great deal of autonomy and flexibility, exactly these features make it increasingly challenging to compare results across studies.

CLASS and TBD are parsimonious frameworks for instructional quality, including three similar aspects intended to represent high-quality instruction. In contrast, the 7Cs framework consists of seven dimensions (see Table 5.1). The 7C framework originated from a workshop Ronald Ferguson designed with educators from Shakers Height, Ohio. The workshop eventually developed the Tripod Engagement Framework and the Tripod Survey to measure student perceptions. The Tripod referred to content, pedagogy, and relationships (Ferguson & Danielson, 2015; Ferguson, 2012). In the next few years, the Tripod Survey continued to develop further by including the interests educators expressed, the research literature on student engagement, and teaching practices (Ferguson & Danielson, 2015; Rowley et al., 2019). Eventually, they developed a survey for teachers and three separate versions of the tripod student survey. One version focused on early elementary students' (kindergarten to 2nd grade), one on upper elementary (third to fifth grades), and one on secondary education (6th to 12th grade). In 2009, the Measures for Effective Teaching (MET) project selected the Tripod surveys as their student perception measure. The student survey was tested and validated extensively within the project and has since gained increasing popularity, mainly in the United States (Ferguson & Danielson, 2015; Wallace et al., 2016).

The seven dimensions of the framework have been grouped into either two or three categories (see Table 5.1). The two categories employed by Ferguson and Danielson (2015) include measures of Press (P; consisting of Challenge and Classroom management) and measures of Support (S; consisting of Care, Confer, Captivate, Clarify, and Consolidate). Additionally, the Guide to the Tripod 7Cs framework proposed three categories that include measures of Personal Support (PS; Care and Confer), Curricular Support (CS; Captivate, Clarify, and Consolidate), and Academic Press (AP; Challenge and Classroom management) (Tripod Education Partners, 2016).

Table 5.1. Showcased frameworks and respective dimensions and sub-dimensions

Frameworks	Dimensions	*Sub*-dimensions
CLASS[a]	Classroom Organization	Behavior management, productivity, instructional learning formats
	Emotional Support	Positive climate, negative climate, sensitivity, regard for student perspective
	Instructional Support	Concept development, quality of feedback, Language modeling
TBD[b]	Classroom Management	(Lack of) disruptions and discipline problems, (Effective) time use/time on task, Monitoring/whithiness, Clear rules and routines
	Student Support	Differentiation and adaptive support, Pace of instruction, Constructive approach to errors, Factual constructive feedback/ appreciation, Interestingness and relevance, Performance pressure and competition (negative indicator), Individual choice options, Teacher → student interactions, Student → teacher interactions and Student → student interactions
	Cognitive Activation	Support of social relatedness experience, Challenging tasks and questions, Exploring and activating prior knowledge, Exploration of the students' ways of thinking/elicit student thinking, Receptive/transmissive understanding of learning of the teacher (negative indicator), Discursive and co-constructive learning, Genetic-Socratic teaching and Supporting metacognition
TEDS-Instruct[c]	Mathematics educational structuring	Dealing with mathematical errors of students, Teachers' mathematical correctness, Teachers' mathematical explanations, mathematical depth of the lesson, support of mathematical competencies, Using multiple representations, deliberate practice, appropriate mathematical examples, Relevance of mathematics for students

Frameworks	Dimensions	*Sub*-dimensions
Tripod 7Cs[d]	Care (S/PS)	Build relationships, address learning needs
	Confer (S/PS)	Respect perspective, promoting discussion and inviting input
	Captivate (S/CS)	Designing simulating lessons and facilitating active participation
	Clarify (S/CS)	Explain Cleary, checking for understanding and providing constructive feedback
	Consolidate (S/CS)	Review and summarize and connect ideas
	Challenge (P/AP)	Press for rigorous thinking, press for quality work and press for persistence
	Classroom Management (P/AP)	Manage activities and manage behavior
MQI[e]	Common core-aligned student practices	Students ask mathematical questions and reasons, give multiple explanations and get cognitively challenging tasks
	Working with students and mathematics	The teacher accurately interprets and responds to students' mathematical ideas and remediates students' errors thoroughly. The teacher includes multiple solution methods, uses fluent and precise mathematical language and develops mathematical generalizations from specific examples
	Richness of mathematics	The teacher explains mathematical ideas and draws connections among different mathematical ideas. The teacher includes multiple solution methods
	Errors and imprecision	The teacher (does not) make content errors, shows imprecision in language and notation, or teaches with a lack of clarity
	Classroom work is connected to mathematics	The classroom work has a mathematical point and instructional time is not spend on activities that do not develop mathematical ideas

[a] the Classroom Assessment Scoring System as indicated by Pianta and Hamre (2009)
[b] the Three Basic Dimensions as indicated by Praetorius et al. (2018)
[c] the TEDS-Instruct as indicated by Jentsch et al. (2020)
[d] the Tripod 7Cs as indicated by the Guide to Tripod's 7Cs framework (2016)
[e] the Mathematical Quality of Instruction framework as indicated by the Center for Education Policy Research at Harvard University (2020)

The hybrid TEDS-Instruct arose from questioning whether generic aspects of the TBD framework were sufficiently reflecting instruction in mathematics teaching (Jentsch et al., 2020). To this background, the generic aspects of instruction from the TBD framework were extended with mathematics-specific aspects of instruction. These aspects were obtained from a previous systematic literature review of existing frameworks in mathematics (Schlesinger & Jentsch, 2016), which led to the sub-dimensions of subject-related mathematics education quality and teaching-related quality. Accordingly, Schlesinger et al. (2018) developed an observational instrument to assess the quality of mathematics instructions in lower-secondary education. Exploratory factor analysis was used on the high-inference ratings obtained by the observational instrument and revealed a four-dimensional factor structure: the three dimensions of the TBD and one subject-specific dimension (see Table 5.1). Thus, the two subject-specific sub-dimensions found earlier were collapsed into the subject-specific dimension of mathematics educational structuring (Jentsch et al., 2020).

Last, The Mathematical Quality of Instruction (MQI) framework was developed by Heather Hill and colleagues to measure the mathematical quality of mathematics instruction. As such, the framework intends to measure aspects of mathematical content available for students during instruction (Learning for Mathematics Project, 2011). In the development process of the MQI, researchers drew on their own experiences, the analysis of video recordings, and existing literature on effective instruction in mathematics. The developers of the MQI followed an iterative process, using the observational data and literature to revise the instrument repeatedly. The framework that followed consisted of several major constructs and corresponding scales. A complete overview of the development process is found in the Learning for Mathematics Project (2011) paper, whereas Charalambous and Litke (2018) provide an overview of the empirical support for the framework. Since the initial development of the framework, there have been several iterations. These have led to five dimensions that intend to "capture the nature of the mathematical content available to students during instruction, as expressed in teacher-student, teacher-content, and student-content interactions" (Center for Education Policy Research from Harvard University, 2020) (see Table 5.1 for an overview of the five dimensions and its sub-dimensions).

Conceptualization: Similarities and Differences

The frameworks presented in Table 5.1 vary regarding their dimensionality. The variety in dimensionality is closely related to how frameworks classify their categories (e.g., dimensions, sub-dimensions, and indicators). With classification, we refer to the systematic arrangement of the different categories or aspects of instructional quality. Many instructional quality frameworks use three levels, specifying their dimensions into more specific sub-dimensions and eventually indicators. However, frameworks vary on several points concerning the classification used, making it increasingly difficult to compare them.

First, frameworks differ in the wording they use for their categories. On the first level, dimensions may also be referred to as domains, elements, or components. On the second level, sub-dimensions are also referred to as subscales, dimensions, or indicators. Second, the different categories are used interchangeably and do not necessarily indicate the degree of specificity.

Table 5.2. An example of the categorization process of the showcased framework

Frameworks	Example Dimensi-ons	Example Sub-dimensions	Example Indicator
CLASS	Emotional support	Positive climate	Relationships
TBD	Student Support	Differentiation and adaptive support	The teacher provides exercises with different difficulty levels
7Cs	Care (Personal Support)	Build relationships	My teacher in this class really tries to understand how students feel about things
	Confer (Personal Support)	Respect perspective	My teacher welcomes my ideas and suggestions

This becomes clearer by observing Table 5.2, which provides an example of the classification of three showcased frameworks. Here, the dimensions *care* and *confer* (grouped into the conceptual category *personal support*) in the 7Cs framework are similar to the dimension of *student support* in TBD or *emotional support* in CLASS. Another example is the sub-dimension "*build relationships*" of the 7Cs framework, which is similar to the indicator "*relationships*" of CLASS. Examples of differences in the classification of categories can be found across all frameworks.

In order to compare several frameworks, it can therefore be necessary to compare across different levels of categories. When we compare the frameworks in Table 5.1 when they are grouped in the most parsimonious way, we can distinguish three core aspects of high-quality instruction that are apparent in all these frameworks, namely: (1) how a classroom is managed, (2) the socio-emotional support of students in classrooms, and (3) teaching that is clear and cognitively challenging. These three overarching constructs are also found when the dimensions of other frameworks of instructional quality are compared to each other (see Wisniewski et al., 2020 for a comparison of several frameworks).

All three key aspects are, to some extent, apparent in all generic frameworks. In contrast, the subject-specific frameworks generally contain instructional practices that focus on the core aspect of teaching that is clear and cognitively challenging and less on how a classroom is managed or the socio-emotional support provided (as seen in the MQI framework and the subject-specific dimension of the TEDS-

Instruct in Table 5.1). This indicates that clear and cognitively challenging instructions may be regarded as more dependent on the subject. In contrast, managing a classroom and providing socio-emotional support are considered aspects of instruction less dependent on the subject under investigation. This is in line with previous studies considering cognitive activation to be a more subject-specific dimension (Dorfner et al., 2018; Praetorius et al., 2016; Schlesinger & Jentsch, 2016).

Thus, to a large extent, scholars seem to agree on several core aspects of instruction reflected in frameworks measuring instructional quality. However, when comparing the content covered by the different frameworks showcased in Table 5.1, it becomes clear that dimensions that cover similar core aspects across frameworks can differ regarding the depth and breadth to which they cover them. An indicator of this difference is the sheer number of sub-dimensions included in similar dimensions. For example, TBD covers ten sub-dimensions for *student support*, whereas the dimension *emotional support* in CLASS or *care* and *confer* in the 7Cs each cover five. To some extent, this is due to the classification and the depth and breadth of these sub-dimensions themselves. However, the choices made in the development process and the purpose of the framework play an essential role. Some frameworks include a large variety of instructional aspects depending on these choices, whereas others choose a more specific focus. Thus, generic frameworks of instructional quality indeed cover, to a large extent, three similar core aspects of instruction. However, a certain degree of caution is required when comparing different frameworks because the extent of instructional practices they cover across seemingly similar dimensions can vary.

OPERATIONALIZATION AND MEASUREMENT

In addition to conceptualizing instructional quality, there is a need to describe precisely how to measure it, which is also referred to as the process of operationalization (DeCarlo, 2018). Several frameworks are already connected to an instrument, thus providing guidelines for measuring the proposed aspects of instruction. In such cases, the operationalization process can be dependent on the instrument. However, researchers could decide to use their own operationalizations. Doing so would change how they measure the aspects of instruction proposed by the framework but would not change the underlying purpose, the theoretical foundation, or conceptualization of the framework in question (Praetorius & Charalambous, 2018).

Instruments and Indicators

Measuring instructional quality is generally done through one or several of three different perspectives: observers, students, and teachers. Some frameworks are connected to instruments that use a specific perspective, whereas others are not connected to any instrument or perspective in particular. For example, the 7C framework is connected to the Tripod Student Survey, and CLASS and MQI are essentially developed as observational instruments. In contrast, the TBD is not connected to a single instrument but is associated with measurements collecting data from all three perspectives (Praetorius et al., 2018). By using a framework and its associated instrument, the choice on which indicators and perspectives to include becomes largely predetermined. However, frameworks such as CLASS and the 7Cs provide different instruments depending on the age group, whereas for TBD, the indicators can differ depending on the study. Both ways offer advantages and disadvantages: using an instrument with predefined indicators creates better opportunities for comparability but is limited by a specific perspective. On the other hand, using multiple perspectives has been debated as well, as previous research has shown that different perspectives (e.g., observers, students, and teachers) do not always agree on how they perceive instructional quality (see, e.g., De Jong & Westerhof, 2001; Kunter & Baumert, 2006; Maulana & Helms-Lorenz, 2016; van Der Scheer et al., 2019; Wagner et al., 2016). Therefore, it has been argued that different perspectives could complement each other by providing several viewpoints (Kunter & Voss, 2013).

In the following sections, we will discuss the different perspectives regarding their strength and weaknesses and their implementation to measure instructional quality in classrooms.

Observations

Classroom observations are often seen as an objective method and, therefore, commonly used (Clare et al., 2001). Observations can be undertaken in a context where the observer is present in the classroom (internal) or where observers rate videotaped lessons (external) (Schlesinger & Jentsch, 2016). In both cases, several trained raters use rating scales and observation sheets to assess classroom behavior.

Before conducting observations, several choices with regards to the study design must be made. Such decisions include the amount and quality of observers that conduct the ratings, the number of lessons to observe, the number of ratings per lesson, the length of a rating period, the degree of inference, and the way of scoring. In addition, segments of observations are nested within lessons and lessons are nested within classes, giving observations a hierarchical structure that needs to be considered in the analysis. Given that the data is hierarchical in nature, a multilevel approach is warranted, such as three-level structural equation modeling with segments at the first level, lessons at the second, and classes at the third level (Lei, 2018). These decisions influence the reliability of the observations and are partly dependent on the framework used. For example, how many lessons should be observed to ensure reliable data depends on the construct under investigation. This is seen in a study done by Praetorius et al. (2014), who found that within the TBD framework, classroom management and a supportive climate were stable across lessons and, therefore, needed only one lesson per teacher. Cognitive activation, on the other hand, showed high variability and needed nine lessons.

Regarding the strength and weaknesses of observations, one of the most common counterarguments is that they are expensive and timely. Observers are also limited to a number of lessons over a short period and thus might not know whether the classroom behavior represents the whole period. In addition, the behavior of teachers might be influenced when they know that they are being observed in the classroom. Last, when the observations are internal, there is a possibility that the observers have to participate actively in the classroom and may therefore concentrate less on observing classroom behavior (De Jong & Westerhof, 2001; Pianta & Hamre, 2009). On the other hand, one of the main strengths is the possibility to estimate inter-rater reliability. This is possible as different observers rate one teacher, creating a source of variance between observers (De Jong & Westerhof, 2001; Lüdtke et al., 2009).

Student Ratings

In contrast to observations, student ratings of teachers obtained through student questionnaires are cost-effective and less time-consuming. They are commonly used in educational research, but whether they can effectively assess teacher behavior has long been debated. Critics have mentioned, among other things, that students lack the competence or stability to rate teacher behavior, and they could be influenced by teacher popularity, the attractiveness of the subject, and interests or grades (Aleamoni, 1999). Moreover, students' age has been mentioned as a factor that might influence ratings as there may be limits to what extent young students can distinguish between details of instructional quality. However, recent research has refuted most critique by providing evidence that student ratings from as young as 3rd grade are a reliable and valid data source to measure instructional quality (De Jong & Westerhof, 2001; Fauth et al., 2014; Kunter & Baumert, 2006; L. Kyriakides et al., 2014; Rowley et al., 2019; van Der Scheer et al., 2019; Wisniewski et al., 2020). Moreover, a strong argument for student ratings is that students' perceptions are based on their day to day experience with different teachers, teaching styles, across subjects, and for an extended period, making them "experts" in the field (De Jong & Westerhof, 2001; Maulana & Helms-Lorenz, 2016).

Student ratings allow for measurement on multiple levels: first at the student level, reflecting the individual students' perception, then at the classroom level, representing students' shared perception in the classroom, or even at the school level, representing the shared perceptions of students within a school (Lüdtke et al., 2006). It is vital to consider the hierarchical or nested structure where students are nested within classrooms and classroom are nested within schools when using student ratings. Because instructional quality is a construct that reflects teacher behavior, the primary unit of analysis for investigating instructional quality should be at the classroom level (Lüdtke et al., 2009). However, using student data at the classroom level also requires several considerations. These include whether the number of students per class is sufficient, whether aggregated student ratings can be reliably evaluated, and whether students within a class agree with the quality of instruction provided by their teacher (Lüdtke et al., 2006). In addition, the design of the study also influences the conclusions that can be derived from the data. For example, if the classroom level is not represented in the data (e.g., if students are randomly sampled within schools like in PISA, rather than sampling whole classes), it is impossible to explain the differences in performance between classes caused by teachers' instructional quality.

TEACHER SELF-REPORTS

Last, teacher self-reports are based on teachers' describing their own instructional quality. Therefore, it has been criticized for being biased by self-serving strategies or teaching ideals (Kunter & Baumert, 2006; Nilsen, Gustafsson & Blömeke, 2016). If teachers answer what they think they are expected to answer, this is often referred to as social desirability (van de Mortel, 2008). Teacher surveys also have a nested structure, with teachers nested in schools. However, in contrast to student questionnaires, teacher surveys provide data directly on the classroom level. Last, both student ratings as teacher self-reports are regularly employed in international large-scale assessment studies such as the Trend in Mathematics and Science Study (TIMSS) and the Progress in International Reading Literacy Study (PIRLS).

Multiple Perspectives

Studies investigating instructional quality collect data using either one specific perspective or a combination of multiple perspectives. When a combination of instruments is used, the results are analyzed independently and compared to each other. To reliably compare results, the different perspectives ideally measure the same construct. In other words, there is a degree of agreement on how observers, students, or teachers perceive teaching behavior associated with instructional quality. The agreement between different types of data sources is debated, but most studies have indicated that the different perspectives do not agree very well (see, e.g., De Jong & Westerhof, 2001; Kunter & Baumert, 2006; Maulana & Helms-Lorenz, 2016; van Der Scheer et al., 2019; Wagner et al., 2016). One reason could be that indicators of high-quality instruction are experienced differently by teachers, students, and observers (Kunter & Voss, 2013). Thus, the different instruments tap on instructional quality from different perspectives. From this point of view, there is no single optimal approach, but different approaches can provide additional perspectives.

INSTRUCTIONAL QUALITY: FINDINGS

In the first part of the present chapter, we showed how conceptualizations, operationalizations, and measurement of instructional quality might vary substantially from one study to the next. In this second part of the chapter, we will show that the same goes for findings; the effect of instructional quality on student outcome varies across countries, students' age, and subject domains (Blömeke & Olsen, 2019; Nilsen, Gustafsson & Blömeke, 2016). The effect further depends on the type of outcome—whether the outcome is cognitive or affective (Blömeke & Olsen, 2019; Nilsen et al., 2018). In addition, the effect on student outcomes also depends on the type of data used, especially whether the data is cross-sectional or longitudinal (Nilsen, Gustafsson & Blömeke, 2016).

Whether a predictor has a different effect depending on the group (e.g., country or age) belongs to the area of educational effectiveness, and more specifically, differential effectiveness (Hall & Lindorff, 2020; Scherer and Nilsen, 2019). There has been a consistent call for research within this area, as educational research increasingly has become aware that predictors, such as instructional quality, may not have the same effect on all groups. It could, for instance, be that some aspects of instructional quality are more important to certain groups of students depending on factors such as country, age or even socioeconomic status (Bergem, Nilsen & Scherer, 2016; Blömeke and Olsen, 2019).

Through four showcases consisting of studies undertaken by the authors, this second part of the present study seeks to illustrate how the relationship between instructional quality and student outcomes varies according to the factors mentioned (countries, type of outcome, grade-level or students' age, subject domain, and type of data), even when the same conceptual framework is used. The findings presented from these four showcases (Blömeke et al., in press; Blömeke & Olsen, 2019; Nilsen et al., 2016; Nilsen et al., 2018) will be discussed in light of research within this field.

Findings Across Countries and Type of Outcomes

Blömeke et al. (2016) examined the relationship between teacher competence, instructional quality, and fourth-graders' achievement in mathematics across the countries that participated in the international large-scale assessment TIMSS 2011. Teacher quality was represented by three dimensions: teacher education background, participation in professional development activities, and teachers' sense of preparedness. Teacher education background was described by teachers' years of experience and initial formal education, including the highest level of formal education and their specialization in mathematics (major or main area(s) of study). Professional development was measured by a set of questions including broad professional development activities (e.g., in mathematics content) and professional development activities preparing for specific challenges (e.g., "integrating information technology into mathematics"). Another set of questions covered collaborative activities representing continuous school-based professional development (e.g., "Visit another classroom to learn more about teaching"). The third teacher quality dimension was teachers' self-efficacy measured as their self-reported sense of preparedness to teach specific topics in mathematics within the three content domains of number, geometric shapes and measures, and data display (e.g., "Adding and subtracting with decimals"). The measure of instructional quality was based on the teacher questionnaire. The teachers were asked six questions on how often they perform various activities in this class. These items tapped into three dimensions of instructional quality: (1) clear instruction (e.g., "Use questioning to elicit reasons and explanations"); (2) cognitive activation (e.g., "Relate the lesson to students' daily lives"); and (3) supportive climate (e.g., "Praise students for good effort").

One of the most interesting results was the considerable variation of relations between instructional quality and achievement across countries. For some countries, the relations were insignificant, and for the rest, the strength and sign of the regression coefficient varied substantially. Such variations of findings are in line with other recent studies (Bellens et al., 2019; Blömeke & Olsen, 2019; Nilsen et al., 2018). Often, the explanation given revolves around cultural differences, especially in international large-scale assessment, where a large number of heterogeneous countries participate from all over the world. Teachers' practices are likely different in, for instance, Western countries and Asian countries and students may respond differently to these practices in Confucian and non-Confucian countries (Blömeke & Olsen, 2019; Van de Vijver & Tanzer, 2004). Following this line of logic, including samples from a homogenous set of countries such as the Nordic countries may produce less variation in results, especially if the same conceptual framework, the same scales, the same data, and the same type of outcomes are used. However, variations are expected if the type of outcomes differs; according to previous research, relations between instructional quality and cognitive outcomes may differ from relations between instructional quality and affective outcomes (Fauth et al., 2014; Yi & Lee, 2017).

Nilsen et al. (2018) investigated the relations between instructional quality and student achievement and motivation for the Nordic countries. Less variation was hypothesized between countries, while variations across outcomes (cognitive and affective) were expected. Teacher quality was measured through qualification and competence. Qualifications included teachers' highest level of formal education, their specialization in science or science education (major or main areas studied), professional development activities in seven science content areas (e.g., "Science pedagogy/instruction"), and the number of hours of professional development activities. Teacher competence was measured by: (1) seven items measuring the extent to which teachers collaborate with other science teachers (e.g., "Discuss how to teach a particular topic"); (2) seven statements measuring how motivated teachers are for their work (e.g., "I am proud of the work I do"); (3) ten statements measuring teachers' self-efficacy in pedagogical content knowledge (e.g., "Making science relevant for students"); and (4) self-efficacy in content knowledge within 22 topics covering the range of all science topics in the TIMSS framework. Instructional quality was measured by teachers' self-reports of seven practices that pertained to cognitive activation (e.g., "Ask students to complete challenging exercises that require them to go beyond the instruction") and teacher support (e.g., "Encourage students to express their ideas in class").

All Nordic countries (25 916 students and 2093 classes/teachers) who participated in TIMSS 2015 were included to investigate the relations between teacher quality, instructional quality and student achievement and motivation in science in grade 4 and 8. Two-level (students and classes) structural equation modeling was used to investigate whether instructional quality mediated the relation between teacher quality and student outcomes. Because the aim was to identify the degree to which teacher quality and instructional quality could explain differences between classes, the focus was on the class level. At the same time, students' individual differences in their ratings of instructional quality were controlled for at the student level. Comparability of the constructs was assured, and a multi-group approach was implemented.

Table 5.3. Standardized regression coefficient for the relation between instructional quality and achievement in science in grade 4

	Denmark	Finland	Sweden	Norway
Achievement	Not significant	0.18*	0.16*	0.33*
Intrinsic motivation	0.25*	Not significant	0.31*	0.33*

Source: Nilsen, Scherer & Blömeke (2018), p. 72 & 90–91.

The results unexpectedly varied between the Nordic countries for the relations between instructional quality and achievement at the class level (see Table 5.3). For Denmark, the relation was not significant; for Finland and Sweden, the relations were weak but significant, while for Norway, there was a moderate relation. For the relations between instructional quality and intrinsic motivation, the results also varied between the Nordic countries. For Finland, the relation was not significant, while the relations were moderately strong for the rest of the countries, albeit a bit weaker for Denmark.

While relations were not expected to vary much between countries, the relations were expected to vary substantially using different outcome (Fauth et al., 2014). Upon comparing the regression coefficients *within* countries for achievement and motivation, the relations were stronger for affective than cognitive outcome (except for Finland).

Findings Across Age

As previously discussed, some studies used different instruments for different age groups of students. There has, however, been some evidence on the applicability of, for instance, the TBD scales developed by Klieme and colleges (2009) across different grades. These studies have validated the scales in grades three and nine in Germany (Fauth et al., 2014; Wagner et al., 2016). However, there is still a clear need for validation studies that provide evidence of measurement invariance across grades. There are several potential causes of measurement non-invariance that could make instructional quality incomparable across grades. These include, for instance, differences in the implementation of instructional quality (e.g., certain types of classroom management could be different so that items do not function similarly). Moreover, perceptions of environments may change over time with students getting older and thus more able to differentiate between aspects of instructional quality (e.g., Wagner et al., 2016; Wittmann & Lehnhoff, 2005). For instance, one age group uses the extreme ends of the scales (e.g., Likert scales) more than the other one by strongly agreeing or disagreeing with statements. It could also be that one group of students understands the questions differently than the other.

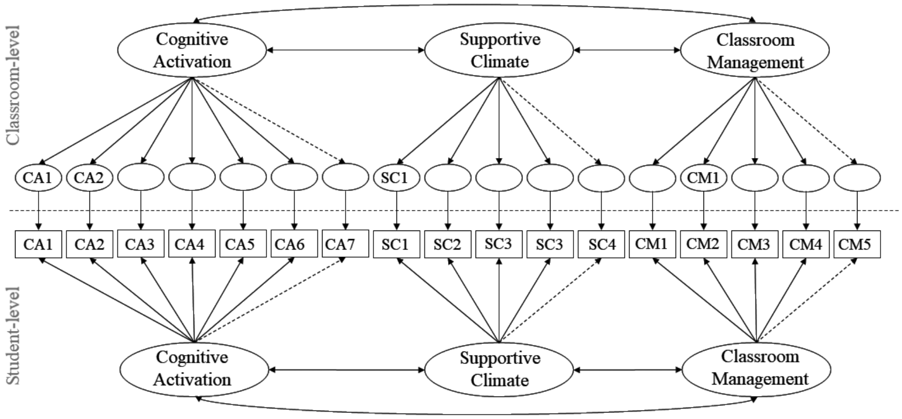


Figure 5.2. The proposed factor structure of instructional quality.

The following showcase (Nilsen et al., 2016) investigated the comparability across grades 4, 5, 8, 9, and 13 in Norwegian schools (20, 353 students) using the TBD framework. In Norway, grade 4 and 5 students attend primary school, grade 8 and 9 students attend lower-secondary school, while grade 13 students attend the last year of upper-secondary school. Data from TIMSS 2015 and TIMSS Advanced 2015 were used. The student questionnaire included questions pertaining to three dimensions, as illustrated in Figure 5.2, of instructional quality in mathematics lessons: Cognitive Activation (7 items, Fauth et al., 2014), Supportive Climate (4 items, Fauth et al., 2014), and Classroom Management (5 items, Baumert et al., 2010).

The results of the analyses revealed that it was possible to replicate the intended factor structure across grades. Furthermore, the data could support metric invariance at the student level and scalar invariance at the class level. This means that, at the student level, the means of the constructs were not comparable across grades while the relations to other variables (e.g., student achievement) were. However, the means of the constructs were also comparable at the class level (as well as relations to other variables).

Figure 5.3 shows the means of the three dimensions or factors at the class level. The factor means of the three dimensions in grade 4 are set to zero, and the means of the other grades are compared to grade 4. Any differences between grade 4 and any of the other grades are significant. Interestingly, the pattern shows that the perception of the quality of instruction seems to decrease with increasing grades, although it seems to improve again in upper-secondary school (grade 13). The only dimension that deviates from the pattern is classroom management in grade 13, which is higher compared to grade 4. However, when interpreting these results, one needs to consider that the target population in grade 13 is very different from grades 4 through 9, so that one cannot compare the results directly.

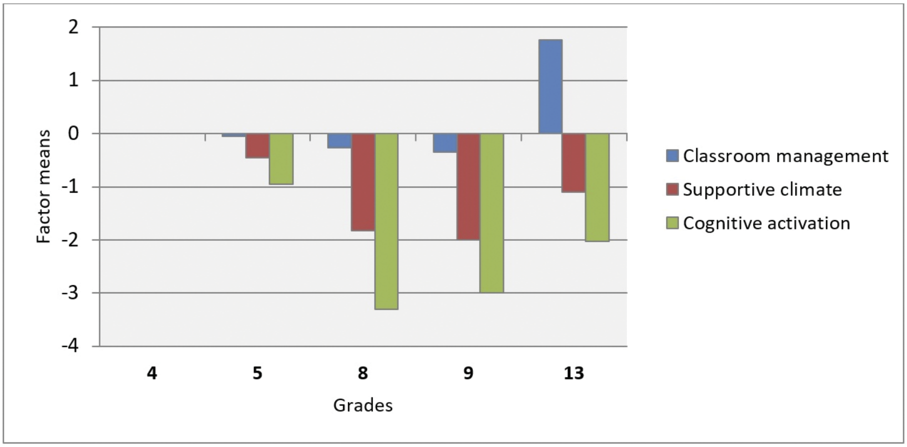


Figure 5.3. Factor mean comparisons at the class level.

The important conclusion of this study is that these scales (Klieme et al., 2009) may be used across students' age in Norway as long as one is careful with respect to mean comparisons. The evidence is the following: 1) the assumption on the applicability of the three-dimensional model was confirmed, 2) the data supported metric invariance both on the individual and the class level, and 3) the invariance at the class level indicated that mean comparisons across grades may be possible.

Using the same data and the same methodology, Bergem et al. (2016) found that the strength of the relationships between all aspects of instructional quality and student achievement decreased from primary to lower-secondary school. This is in line with other studies (e.g., Nilsen et al., 2018). However, whether or not instructional quality is more critical to student outcome in primary than lower-secondary school still requires further research. It would, for example, be important to check whether the difference in strength of relation is a function of differential variance. It would also be important to include other predictors and control variables as previous research has shown that it is essential to include the whole picture: for example, where instructional quality mediates the relation between teacher competence and student outcome (Baumert et al., 2010; Blömeke et al., 2013).

Across Subject Domains

Many of the best-known instructional quality frameworks and observation protocols were implemented in a specific subject domain, such as the TBD in mathematics (Klieme et al., 2009) and the PLATO manual based on the (Grossman et al., 2013; Klette & Blikstad-Balas, 2018) but are developed with the intention to assess generic aspects of instruction that apply to all subjects (Klette & Blikstad-Balas, 2018; Praetorius et al., 2018).

Some studies have investigated to what extent the instruments used are comparable ("measurement invariant") across subjects. This is a crucial measurement question and concerns the structure of an instrument and how the items of its different scales function. For instance, Rieser et al. (2016) examined the comparability of the TBD measured by student ratings in Germany through national extensions of TIMSS 2015. They found that all three dimensions were measurement invariant across science and mathematics.

Having confirmed that the TBD are comparable across science and math, we present a third showcase where the relations between instructional quality and student outcomes in mathematics and science were estimated (Blömeke & Olsen, 2019). Using data from TIMSS 2011, multilevel structural equation models were applied to the datasets of five countries in multiple-group analyses. The five countries were selected to test consistency across heterogeneous contexts. The countries represent different cultures that could potentially affect relationships between instructional quality and student outcomes. South Korea and Thailand were chosen to represent Asian countries with both Confucian and non-Confucian heritage. In addition, England and Norway were chosen to represent English-speaking and non-English-speaking western countries, and Tunisia was chosen to represent an Arabic speaking country.

Table 5.4. The relationships between instructional quality and mathematics and science achievement in grade 4 and 8

	G4 Math	G4 Science	G8 Math	G8 Science
England	−0.16	−0.15	0.32***	0.19**
Norway	0.14	0.37**	0.32**	0.22
South Korea	0.27*	0.23*	0.52***	0.28**
Thailand	0.14	0.24**	−0.17	−0.23**
Tunisia	0.38***	0.29**	−0.12	0.04

Source: Blömeke & Olsen (2019), p. 178.

The study included both grade 4 and grade 8 students and cognitive and affective outcome to test consistency of relations across different age groups and types of outcomes. However, in the present showcase, we focus only on cognitive outcomes to facilitate comparisons between mathematics and science (see Table 5.4). The findings show that the relationships between instructional quality and achievement varied across mathematics and science, grade 4 and 8 students, and countries. The authors concluded that it does not seem to be justified to generalize results across these groups.

Findings Across Time

The studies presented above are cross-sectional studies where the data was collected at one time point only. There are several threats to causal inferences drawn from such data (Gustafsson, 2013; Gustafsson & Nilsen, in press). One of these threats is reversed causality, an issue that occurs when there is no way of telling whether the predictor affects the outcome or the other way around. Another threat is omitted variables where a variable not part of the model is the actual cause behind an effect. There are several ways to address omitted variables, reversed causality and other issues related to causal inference drawn based on cross-sectional studies if, for example, data from several rounds are available. Using causal methodology such as examining how differences in one variable are related to differences in another one over such rounds or taking advantage of the longitudinal design at the country level are approaches that will reduce or even remove the threats to causal inferences.

Longitudinal studies on the individual level would also be able to circumvent several threats to causality. However, there are few longitudinal studies because they require a large amount of resources—especially if with large or representative samples that enable the generalization of/across populations. Within the field of teacher instruction, the COACTIV project was a 1-year longitudinal extension of PISA in Germany designed to study the importance of Mathematical Pedagogical Content Knowledge and Mathematical Content Knowledge for instructional quality and student achievement (Baumert et al., 2010). A sample of Grade 10 Mathematics students and their teachers were included in the study, and instructional quality was measured through three basic dimensions. The findings showed that both cognitive activation and classroom management had moderate effects on achievement (0.32 and 0.30, respectively), while there was no significant effect of student support. These results align with the hypothesis that cognitive activation and classroom management are mainly related to cognitive outcomes, whereas student support is considered more important to affective outcomes (Klieme et al., 2009; Praetorius et al., 2018).

A similar idea was behind the Teachers' Effects on Student Outcome (TESO) project: a longitudinal extension of TIMSS 2019 funded by the Research Council of Norway. TESO has representative samples of students at the Norwegian national level. The aim is to investigate the effect of teacher quality and instructional quality on the development of student achievement and motivation in mathematics and science. Instructional quality is conceptualized in line with the TBD framework of Klieme and colleagues (2009) and measured by three types of data: video observations, student, and teacher questionnaires.

Linking Instruction and Student Achievement (LISA) is another longitudinal project whose aim is to identify effects of teachers' instructional quality on student achievement in mathematics and reading. Student achievement is measured through national tests. This project was the first classroom study in Norway to include large samples (Klette et al., 2017). No effects of instructional quality on student outcome have yet been identified.

The fourth and final study being showcased in the present chapter is a recent longitudinal study by Blömeke et al. (in press). This study collected data at two time points, with 1.5 to 2 years in between. The aim was to investigate the effect of teachers' mathematical content, and pedagogical content, knowledge and teachers' skills (perception, interpretation, and decision-making skills) on student learning progress and whether these were mediated by instructional quality. The sample included 3,496 eighth grade students from 154 classrooms and their teachers in Germany. Two-level mediation models with students on the first and classes on the second level were employed. Using the TEDS-Instruct instrument (Blömeke et al., 2020), teachers' mathematical content and pedagogical content knowledge were measured by standardized knowledge tests. Teachers' cognitive skills were measured through teachers' reactions to typical classroom situations presented in three video clips. Classroom observations were used to measure instructional quality, and student achievement in mathematics was measured by tests based on national standards. To reduce the complexity of the model, which included a large chain of constructs, instructional quality was measured by one latent construct consisting of all three basic dimensions and one subject-specific dimension.

The results showed that teachers' mathematical content knowledge predicted their mathematical pedagogical content knowledge, which again affected teachers' skills. Neither mathematical content knowledge nor mathematical pedagogical content knowledge had a direct relation to students' learning progress. However, both teachers' skills and instructional quality significantly affected student learning progress, where the effect size of instructional quality in the full model was 0.18. Compared to Baumert et al. (2010), the effect size in the study by Blömeke et al. (in press) was smaller. This could be related to the high correlation between teachers' skills and instructional quality, meaning that the two constructs measured much of the same. This is an interesting finding in itself because it points to a potential starting point for educational treatments intended to improve instructional quality. Moreover, Blömeke et al. (in press) used a different operationalization with one latent construct to measure instructional quality, which could explain the difference in results compared to Baumert et al. (2010). Nonetheless, both studies show that, even when previous achievement is controlled for, instructional quality positively affects student achievement in mathematics.

SUMMARY AND CONCLUSION

Despite all the differences found across the field of instructional quality, we would like to start with some common ground and recent advances. First, instructional quality is most frequently conceptualized in a generic way, independent of the specific content taught. Such a conceptualization often includes three core aspects: (1) aspects related to the management of the classroom; (2) aspects of socio-emotional support in classrooms; and (3) aspects of teaching that is clear and cognitively challenging.

Recently, a repeated call for bringing together generic and subject-specific conceptualizations of instructional quality has been made (Blazar et al., 2017; Charalambous & Praetorius, 2018; Praetorius & Charalambous, 2018). The general idea behind those efforts is that a complex multidimensional construct such as instructional quality needs to consider both general and subject-specific aspects of instruction to form a complete picture of what happens in the classroom (Blazar et al., 2017; Charalambous & Kyriakides, 2017).

Ideally, researchers would agree on one framework measuring generic aspects of instructional quality and one specific framework for each subject that neither has the issue of overrepresentation nor underrepresentation. However, such consensus seems far away, and subject-specific frameworks are primarily available for particular subjects such as mathematics and science and less for subjects such as arts, languages, or physical education. For now, efforts in bringing together both perspectives are primarily being made by building on already existing frameworks, which has the benefit of tapping into decades of scholarly work already done in the field. While ultimately, the choice for a specific framework(s) might depend on the research question at hand (e.g., whether one wants to examine subject-specific or generic aspects of instruction, or a combination of both), we expect the current efforts in bringing both types of conceptualizations to continue both in theory and practice.

After deciding what framework(s) to use, there is a range of choices concerning the operationalizations. These choices partly depend on the framework at hand and the instrument(s) they are connected to. Existing instruments with predetermined items and scales lead to more easily comparable studies. However, these instruments often offer a single perspective (e.g., observations, students' or teachers' report) and, to our knowledge, not a single framework offers validated and widely used instruments covering all three perspectives. The use of a single perspective to measure instructional quality might be one of its pitfalls as it is, to date, unclear how to deal with the disagreement between different perspectives. As for now, the consensus seems to be that all perspective offer specific advantages (Kunter & Voss, 2013). Including experiences from multiple perspectives could therefore lead to a more complete picture of instructional quality in the classroom. Yet, they should have the same theoretical underlying framework.

In addition, measuring instructional quality through multiple perspectives over an extended period would enhance the robustness of inferences and causal claims. Thus, triangulation of multiple measures of instructional quality (e.g., observations and questionnaires) and longitudinal designs would: (1) facilitate comparisons across studies; (2) advance the knowledge of the field; (3) enable more reliable and robust inferences; and (4) provide better advice to educational policy and practice. Unfortunately, such research designs are expensive and time-consuming and, therefore, not often found. However, the inclusion of measures of instructional quality in international large-scale assessments, such as TIMSS and PISA, provides a possibility for such a design when followed up with a longitudinal extension and complimented with classroom observations.

In the second part of the review, we showcased the findings of several studies, indicating that the effects of instructional quality on student outcomes vary across countries, students' age, and subject domains. Despite differences in how the three dimensions of instructional quality were operationalized, the four showcases indicate that the effect of instructional quality on student outcomes may depend on the cultural context, students' age, and subject domain. As such, this chapter empirically validates issues of differential effectiveness, which have attracted increased interest in the field (see, e.g., Hall & Lindorff, 2020). However, more research is needed, especially with instruments that are largely similar so that differential effects of instructional quality can reliably be identified.

Concerning the cultural context, it could be that certain aspects of instructional quality have different effects in different countries, for instance, in Confucian versus non-Confucian countries. Alternatively, certain aspects of instruction might be deemed important in one part of the world but not in the other (Berliner, 2005). Most frameworks of instructional quality originated in the context of the global north and are conceptualized according to those cultural standards. Moreover, they are scarcely tested and used outside of this specific context. More research is needed where conceptual frameworks are created and tested in collaboration across cultural contexts to learn from each other. For this to happen, researchers need to collaborate across nations and fields, share and disseminate openly, and strive to bring the field forward together.

The differential effect may also apply to students' age and subject domain. Several studies have even shown that instructional quality may have different effects on different subgroups of students, for instance, depending on their gender, ethnicity, and socioeconomic status (Baumert et al., 2010; Gustafsson et al., 2018; Nilsen & Bergem, 2020; Rjosk et al., 2014). If so, the question arises whether each country should have its own scale or items, and in addition, different scales or items for different ages of students, for different subject domains, and even for different groups of students. Still, it would be beneficial to research, especially meta-analyses, and practice if the core of all scales of instructional quality could be retained and that minor adjustment could be made according to needs and context. This is a complex measurement challenge that needs to be examined further.

Thus, even though instructional quality has gained increased attention and the number of studies addressing this topic is becoming substantial, further research is still needed to address the above issues. The field needs a shared understanding of what constitutes good teaching, for whom, and how to measure it.

REFERENCES

Aleamoni, L. (1999). Student rating myths versus research facts. *Journal of Personnel Evaluation in Education, 13*(2), 153–166. https://doi.org/10.1023/A:1008168421283

Baumert, J., Kunter, M., Blum, W., Brunner, M., Voss, T., Jordan, A., Klusmann, U., Kraus, S., Neubrand, M., & Tsai, Y.-M. (2010). Teachers' mathematical knowledge, cognitive activation in the classroom, and student progress. *American Educational Research Journal, 47*(1), 133–180. https://doi.org/10.3102/0002831209345157

Bellens, K., Van Damme, J., Van Den Noortgate, W., Wendt, H., & Nilsen, T. (2019). Instructional quality: Catalyst or pitfall in educational systems' aim for high achievement and equity? An answer based on multilevel SEM analyses of TIMSS 2015 data in Flanders (Belgium), Germany, and Norway. *Large-scale Assessments in Education, 7*(1), 1–27. https://doi.org/10.1186/s40536-019-0069-2

Bergem, O. K., Nilsen, T., & Scherer, R. (2016). Undervisningskvalitet i Matematikk [Instructional quality in Mathematics]. In O. K. Bergem, H. Kaarstein, & T. Nilsen (Eds.), *Vi kan lykkes i realfag [We can succeed in mathematics and science]* (pp. 120–136). Universitetsforlaget. https://doi.org/10.18261/97882150279999-2016-08

Berliner, D. (2005). The near impossibility of testing for teacher quality. *Journal of Teacher Education, 56*(3*)*, 205–213. https://doi.org/10.1177/0022487105275904

Berry, R. Q., Rimm-Kaufman, S. E., Ottmar, E. M., Walkowiak, T. A., Merritt, E., & Pinter, H. H. (2012). *The Mathematics Scan (M-Scan): A measure of standards-based mathematics teaching practices (Unpublished measure)*. University of Virginia.

Bihler, L.-M., Agache, A., Kohl, K., Willard, J. A., & Leyendecker, B. (2018). Factor analysis of the classroom assessment scoring system replicates the Three Domain Structure and reveals no support for the Bifactor Model in German preschools. *Frontiers in Psychology, 9*, 1–13. https:/ /doi.org/10.3389/fpsyg.2018.01232

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

Blömeke, S., Gustafsson, J. E., & Shavelson, R. (2013). Assessment of competencies in higher education. *Zeitschrift für Psychologie, 221*(3), 202–202. https://doi.org/10.1027/2151-2604/a000148

Blömeke, S., Jentsch, A., Ross, N., Kaiser, G., & Koenig, J. (in press). Opening up the black box: Teacher competence, instructional quality and students' learning progression. *Learning and Instruction*.

Blömeke, S., Kaiser, G., König, J., & Jentsch, A. (2020). Profiles of mathematics teachers' competence and their relation to instructional quality. *ZDM, 52*(2), 329–342. https://doi.org/10.1007/978-3-319-41252-8_2

Blömeke, S., & Olsen, R. V. (2019). Consistency of results regarding teacher effects across subjects, school levels, outcomes and countries. *Teaching and teacher education, 77*, 170–182. https://doi.org/10.1016/j.tate.2018.09.018

Blömeke, S., Olsen, R. V., & Suhl, U. (2016). Relation of student achievement to the quality of their teachers and instructional quality. In T. Nilsen & J.E. Gustafsson (Eds.), *Teacher quality, instructional quality and student outcomes: Relationships across countries, cohorts and time* (pp. 21–50). Springer. https://doi.org/10.1007/978-3-319-41252-8_2

Burroughs, N., Gardner, J., Lee, Y., Guo, S., Touitou, I., Jansen, K., & Schmidt, W. (2019). A review of the literature on teacher effectiveness and student outcomes. In N. Burroughs, J. Gardner, Y. Lee, S. Guo, I. Touitou, K. Jansen, & W. Schmidt (Eds.), *Teaching for excellence and* *equity: Analyzing teacher characteristics, behaviors and student outcomes with TIMSS* (pp. 7–17). Springer. https://doi.org/10.1007/978-3-030-16151-4_2

Carlson J., Daehler, K.R., Alonzo, A.C., Barendsen, E., Berry, A., Borowski, A., Carpendal, J., Kam Ho Chan, K., Cooper, R., Friedrichsen, P., Gess-Newsome, J., Henze-Rietveld, I., Hume, A., Kirschner, S., Liepertz, S., Loughran, J., Mavhunga, E., Neumann, K., Nilsson, P., Park, S., Rollnick, M., Sickel, A., Schneider, R.M., Kjung Suh, J., van Driel, J., Wilson, C.D. (2019) The refined consensus model of pedagogical content knowledge in science education. In A. Hume, R. Cooper, A. Borowski (Eds.), *Repositioning pedagogical content knowledge in teachers' knowledge for teaching science* (pp. 77–94). Springer. https://doi.org/10.1007/978-981-13-5898-2_2

Center for Education Policy Research from Harvard University. (2020). *Mathematical quality of instruction (MQI): MQI domains.* https://cepr.harvard.edu/mqi-domains

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

Charalambous, C. Y., Kyriakides, E., Kyriakides, L., & Tsangaridou, N. (2019). Are teachers consistently effective across subject matters? Revisiting the issue of differential teacher effectiveness. *School Effectiveness and School Improvement, 30*(4), 353–379. https://doi.org/10.1080/09243453.2019.1618877

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

Clare, L., Valdes, R., Pascal, J., & Steinberg, J. R. (2001). Teachers' assignments as indicators of instructional quality in elementary schools (CSE technical report 545). Center for the study of evaluation. National center for research on evaluation, standards, and student testing. http://d-scholarship.pitt.edu/26215/1/TR545.pdf

Clausen, M. (2002). *Unterrichtsqualität: Eine Frage der Perspektive? Empirische Analysen zur Übereinstimmung, Konstrukt- und Kriteriumsvalidität.* Waxmann.

Cohen, J., Ruzek, E., & Sandilos, L. (2018). Does teaching quality cross subjects? Exploring consistency in elementary teacher practice across subjects. *AERA Open, 4*(3), 1–16. https://doi.org/10.1177/2332858418794492

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

Danielson, C. (2007). *Enhancing professional practice: A framework for teaching (2nd ed.).* Association for supervision and curriculum development.

De Jong, R., & Westerhof, K. (2001). The quality of student ratings of teacher behaviour. *Learning Environments Research, 4*(1), 51–85. https://doi.org/10.1023/A:1011402608575

DeCarlo, M. (2018). *Scientific Inquiry in Social Work.* Open Social Work Education.

Dorfner, T., Förtsch, C., & Neuhaus, B. J. (2018). Effects of three basic dimensions of instructional quality on students' situational interest in sixth-grade biology instruction. *Learning and Instruction, 56*, 42–53. https://doi.org/10.1016/j.learninstruc.2018.03.001

Fauth, B., Decristan, J., Rieser, S., Klieme, E., & Büttner, G. (2014). Student ratings of teaching quality in primary school: Dimensions and prediction of student outcomes. *Learning and Instruction, 29*, 1–9. https://doi.org/10.1016/j.learninstruc.2013.07.001

Ferguson, R. (2012). Can student surveys measure teaching quality? *Phi Delta Kappan Magazine, 94*(3), 24–28. https://doi.org/10.1177/003172171209400306

Ferguson, R., & Danielson, C. (2015). How framework for teaching and tripod 7Cs evidence distinguish key components of effective teaching. In T. Kane, K. Kerr, & R. Pianta (Eds.), *Designing teacher evaluation systems* (pp. 98–143). Jossey-Bass. https://doi.org/10.1002/9781119210856.ch4

Goe, L., & Stickler, L. M. (2008). Teacher quality and student achievement: Making the most of recent research. TQ research & policy brief. National comprehensive center for teacher quality.

Grossman, P., Loeb, S., Cohen, J., & Wyckoff, J. (2013). Measure for measure: The relationship between measures of instructional practice in middle school English language arts and teachers' value-added scores. *American Journal of Education, 119*(3), 445–470. https://doi.org/10.1086/669901

Gruehn, S. (2000). *Unterricht und schulisches Lernen: Schüler als Quellen der Unterrichtsbeschreibung*. Waxmann.

Gustafsson, J.E., Nilsen, T., & Hansen, K. Y. (2018). School characteristics moderating the relation between student socioeconomic status and mathematics achievement in grade 8. Evidence from 50 countries in TIMSS 2011. *Studies in Educational Evaluation, 57*, 16–30.

Gustafsson, J.E. (2013). Causal inference in educational effectiveness research: A comparison of three methods to investigate effects of homework on student achievement. *School Effectiveness and School Improvement, 24*(3), 275–295. https://doi.org/10.1016/j.stueduc.2016.09.004

Gustafsson, J. E., & Nilsen, T. (In press). Methods of causal analysis with ILSA data. In T. Nilsen, A. Stancel-Piątak, & J. E. Gustafsson (Eds.), *International handbook of comparative large-scale studies in education.* Springer International Handbooks of Education.

Hall, J., & Lindorff, A. (2020). *International perspectives in educational effectiveness research.* Springer. https://doi.org/10.1007/978-3-030-44810-3

Hamre, B. K., & Pianta, R. C. (2007). Learning opportunities in preschool and early elementary classrooms. In R. C. Pianta, M. J. Cox, & K. L. Snow (Eds.), *School readiness and the transition to kindergarten in the era of accountability* (pp. 49–83). Paul H Brookes Publishing.

Hattie, J. (2009). *Visible learning: A synthesis of over 800 meta-analyses relating to achievement.* Routledge. https://doi.org/10.4324/9780203887332

Jentsch, A., Schlesinger, L., Heinrichs, H., Kaiser, G., König, J., & Blömeke, S. (2020). Erfassung der fachspezifischen Qualität von Mathematikunterricht: Faktorenstruktur und Zusammenhänge zur professionellen Kompetenz von Mathematiklehrpersonen [Measuring the subject-specific quality in mathematics instruction: factor structure and relations to mathematics teachers' professional competence]. *Journal für Mathematik-Didaktik, 42*, 97–121. https://doi.org/10.1007/s13138-020-00168-x

Junker, B., Weisberg, Y., Matsumura, L.C., Crosson, A., Wolf, M.K., Levison, A., & Resnick, L. (2006). Overview of the instructional quality assessment (CSE technical report 671). Center for the study of evaluation. National center for research on evaluation, standards, and student testing. https://cresst.org/wp-content/uploads/R671.pdf

Kane, T. J., & Staiger, D. O. (2012). *Gathering feedback for teaching: Combining high quality observations with student surveys and achievement gains*. Bill & Melinda Gates Foundation.

Klette, K., & Blikstad-Balas, M. (2018). Observation manuals as lenses to classroom teaching: Pitfalls and possibilities. *European Educational Research Journal, 17*(1), 129–146. https://doi.org/10.1177/1474904117703228

Klette, K., Blikstad-Balas, M., & Roe, A. (2017). Linking instruction and student achievement: Research design for a new generation of classroom studies. *Acta Didactica Norge – tidsskrift for fagdidaktisk forsknings- og utviklingsarbeid i Norge, 11*(3), 19. https://doi.org/10.5617/adno.4729

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

Klieme, E., Schümer, G., & Knoll, S. (2001). Mathematikunterricht in der Sekundarstufe I: "Aufgabenkultur" und Unterrichtsgestaltung. In E. Klieme, & J. Baumert (Eds.), *TIMSS – Impulse für Schule und Unterricht* (pp. 43–57). Bundesministerium für Bildung und Forschung.

Konstantopoulos, S., & Chung, V. (2011). The persistence of teacher effects in elementary grades. *American Educational Research Journal, 48*(2), 361–386. https://doi.org/10.3102/0002831210382888

Kunter, M., & Baumert, J. (2006). Who is the expert? Construct and criteria validity of student and teacher ratings of instruction. *Learning Environments Research, 9(3)*, 231–251. https://doi.org/10.1007/s10984-006-9015-7

Kunter, M., & Voss, T. (2013). The model of instructional quality in COACTIV: A multicriteria analysis. In M. Kunter, J. Baumert, W. Blum, U. Klusmann, S. Krauss, & M. Neubrand (Eds.), *Cognitive activation in the mathematics classroom and professional competence of teachers* (pp. 97–124). Springer. https://doi.org/10.1007/978-1-4614-5149-5_6

Kyriakides, E., Tsangaridou, N., Charalambous, C., & Kyriakides, L. (2018). Integrating generic and content-specific teaching practices in exploring teaching quality in primary physical education. *European Physical Education Review, 24*(4), 418–448. https://doi.org/10.1177/1356336x16685009

Kyriakides, L., Creemers, B., Panayiotou, A., Vanlaar, G., Pfeifer, M., Cankar, G., & McMahon, L. (2014). Using student ratings to measure quality of teaching in six European countries. *European Journal of Teacher Education, 37*(2). https://doi.org/10.1080/02619768.2014.882311

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.

Lei, Xiaoxuan, Li, Hongli & Leroux, Audrey. (2018). Does a teacher's classroom observation rating vary across multiple classrooms? *Educational Assessment Evaluation and Accountability, 30*, 27–46. https://doi.org/10.1007/s11092-017-9269-x

Lüdtke, O., Robitzsch, A., Trautwein, U., & Kunter, M. (2009). Assessing the impact of learning environments: How to use student ratings of classroom or school characteristics in multilevel modeling. *Contemporary Educational Psychology, 34*(2), 120–131. https://doi.org/10.1016/j.cedpsych.2008.12.001

Lüdtke, O., Trautwein, U., Kunter, M., & Baumert, J. (2006). Reliability and agreement of student ratings of the classroom environment: A reanalysis of TIMSS data. *An International Journal, 9*(3), 215–230. https://doi.org/10.1007/s10984-006-9014-8

Maulana, R., & Helms-Lorenz, M. (2016). Observations and student perceptions of the quality of preservice teachers' teaching behaviour: construct representation and predictive quality. *An International Journal, 19*(3), 335–357. https://doi.org/10.1007/s10984-016-9215-8

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

Nilsen, T., Gustafsson, J.E., & Blömeke, S. (2016). Conceptual framework and methodology of this report. In T. Nilsen & J. E. Gustafsson (Eds.), *Teacher quality, instructional quality, and student outcomes: Relationships across countries, cohorts and time* (pp. 1–19). Springer. https://doi.org/10.1007/978-3-319-41252-8

Nilsen, T., & Bergem, O. K. (2020). Teacher competence and equity in the nordic countries: Mediation and moderation of the relation between SES and achievement. *Acta Didactica Norden, 14*(1), 1–26. https://doi.org/10.5617/adno.7946

Nilsen, T., Scherer, R., Bergem, O. K., & Kaarstein, H. (2016). *Student ratings of instructional quality: How valid are they across grades? [Conference presentation]* ECER 2016, Dublin.

Nilsen, T., Scherer, R., & Blömeke, S. (2018). The relation of science teachers' quality and instruction to student motivation and achievement in the 4th and 8th grade: A nordic perspective. In The nordic council of ministers (Ed.), Northern lights on TIMSS ansd PISA 2018 (pp. 61–94). Nordic Council of Ministers. https://doi.org/10.6027/TN2018-524

Pakarinen, E., Lerkkanen, M.-K., Poikkeus, A.-M., Kiuru, N., Siekkinen, M., Rasku-Puttonen, H., & Nurmi, J.-E. (2010). A validation of the classroom assessment scoring system in finnish kindergartens. *Early Education & Development, 21*(1), 95–124. https://doi.org/10.1080/10409280902858764

Pianta, R. C., & 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

Pianta, R. C., La Paro, K. M., & Hamre, B. K. (2008). *Classroom assessment scoring system™: Manual K-3.* Paul H Brookes Publishing.

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., Grünkorn, J., & Klieme, E. (2020). Towards developing a theory of generic teaching quality: Origin, current status, and necessary next steps regarding the three basic dimensions model. *Zeitschrift für Pädagogik*, *66*(1), 15–36. https://doi.org/10.3262/ZPB2001015

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

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

Praetorius, A.-K., Vieluf, S., Saß, S., Bernholt, A., & Klieme, E. (2016). The same in German as in English? Investigating the subject-specificity of teaching quality. *Zeitschrift für Erziehungswissenschaft, 19*(1), 191–209. https://doi.org/10.1007/s11618-015-0660-4

Rieser, S., Wendt, H., Hole, A., & Grønmo, L.S. (2016, August 24). *Student ratings of instructional quality: How valid are they across subjects?* [Conference symposium paper]. ECER 2016, Dublin, Ireland. https://eera-ecer.de/ecer-programmes/conference/21/contribution/39198/

Rjosk, C., Richter, D., Hochweber, J., Lüdtke, O., Klieme, E., & Stanat, P. (2014). Socioeconomic and language minority classroom composition and individual reading achievement: The mediating role of instructional quality. *Learning and Instruction, 32*, 63–72. https://doi.org/10.1016/j.learninstruc.2014.01.007

Rowley, J. F. S., Phillips, S. F., & Ferguson, R. F. (2019). The stability of student ratings of teacher instructional practice: Examining the one-year stability of the 7Cs composite. *School Effectiveness and School Improvement, 30*(4), 549–562. https://doi.org/10.1080/09243453.2019.1620293

Scheerens, J., Luyten, J. W., Steen, R., & de Thouars, Y. C. H. (2007). *Review and meta-analyses of school and teaching effectiveness.* Universiteit Twente, Afdeling Onderwijsorganisatie en management.

Scherer, R., & Nilsen, T. (2019). Closing the gaps? Differential effectiveness and accountability as a road to school improvement. *School Effectiveness and School Improvement, 30*(3), 255–260. https://doi.org/10.1080/09243453.2019.1623450

Schlesinger, L., & Jentsch, A. (2016). Theoretical and methodological challenges in measuring instructional quality in mathematics education using classroom observations. *ZDM, 48*, 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, 50*(3), 475–490. https://doi.org/10.1007/s11858-018-0917-5

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

Seidel, T., & Shavelson, R. J. (2007). Teaching effectiveness research in the past decade: The role of theory and research design in disentangling meta-analysis results. *Review of Educational Research, 77*(4), 454–499. https://doi.org/10.3102/0034654307310317

Siedentop, D., Doutis, P., Tsangaridou, N., Ward, P., & Rauschenbach, J. (1994). Don't sweat gym! An analysis of curriculum and instruction. *Journal of Teaching in Physical Education, 13*, 375–394 https://doi.org/10.1123/jtpe.13.4.375

Tripod Education Partners (2016). *Guide to Tripod's 7Cs^{TM} Framework of effective teaching.* Cambridge Innovation Center. https://tripoded.com/teacher-toolkit/

van de Mortel, T. F. (2008). Faking it: Social desirability response bias in self-report research. *Australian journal of advanced nursing, 25*(4), 40–48. https://www.ajan.com.au/archive/Vol25/Vol_25-4_vandeMortel.pdf

van de Vijver, F., & Tanzer, N. (2004). Bias and equivalence in cross-cultural assessment. *European review of applied psychology, 54*(2), 119–135. https://doi.org/10.1016/j.erap.2003.12.004

van Der Scheer, E. A., Bijlsma, H. J. E., & Glas, C. A. W. (2019). Validity and reliability of student perceptions of teaching quality in primary education. *School Effectiveness and School Improvement, 30*(1), 30–50. https://doi.org/10.1080/09243453.2018.1539015

Virtanen, T., Pakarinen, E., Lerkkanen, M.-K., Poikkeus, Siekkinen, M., & Nurmi, J.-E. (2017). A validation study of classroom assessment scoring system secondary in the finnish school context. *The Journal of Early Adolescence, 38*(6), 849–880. https://doi.org/10.1177/0272431617699944

Wagner, W., Göllner, R., Werth, S., Voss, T., Schmitz, B., & Trautwein, U. (2016). Student and teacher ratings of instructional quality: Consistency of ratings over time, agreement, and predictive power. *Journal of Educational Psychology, 108*(5), 705–721. https://doi.org/10.1037/edu0000075

Walkington, C., & Marder, M. (2015). 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* (pp. 234–277). Jossey-Bass.

Wallace, T. L., Kelcey, B., & Ruzek, E. (2016). What can student perception surveys tell us about teaching? Empirically testing the underlying structure of the tripod student perception survey. *American Educational Research Journal, 53*(6), 1834–1868. https://doi.org/10.3102/0002831216671864

Wisniewski, B., Zierer, K., Dresel, M., & Daumiller, M. (2020). Obtaining secondary students' perceptions of instructional quality: Two-level structure and measurement invariance. *Learning and Instruction, 66.* https://doi.org/10.1016/j.learninstruc.2020.101303

Wittmann, M., & Lehnhoff, S. (2005). Age effects in perception of time. *Psychological Reports, 97*(3), 921–935. https://doi.org/10.2466/pr0.97.3.921-935

Yi, H., & Lee, Y. (2017). A latent profile analysis and structural equation modeling of the instructional quality of mathematics classrooms based on the PISA 2012 results of Korea and Singapore. *Asia Pacific Education Review, 18*(1), 23–39. https://doi.org/10.1007/s12564-016-9455-4