Obtaining secondary students' perceptions of instructional quality: Two-level structure and measurement invariance

Benedikt Wisniewski[a,*], Klaus Zierer[a], Markus Dresel[b], Martin Daumiller[b]

[a] Department of School Pedagogy, University of Augsburg, Germany
[b] Department of Psychology, University of Augsburg, Germany

1. Introduction

Instructional quality is a strong predictor of students' learning outcomes (Helmke, 2012; Seidel & Shavelson, 2007). To assess aspects of instructional quality in classrooms, various sources of information can be used. While classroom observations by colleagues or trained observers are time consuming and restricted to narrow time frames (Van der Scheer et al., 2019), the assessment of instructional quality by students has the advantages of providing aggregated data that can cover extended periods of time and does not require additional resources. When acquiring information in the form of standardized assessments of students' perceptions of instructional quality (SPIQ), it is necessary to discuss whether these can be used as a reliable and valid information source. SPIQ have frequently been examined in the context of higher education (college or university students' evaluations of teaching, SET, Clayson, 2008; Cohen, 1980; Marsh, 1980), but, due to differences with respect to learners' as well as teaching characteristics, this research is not readily transferable to a secondary education context, which requires specific investigation.

Marsh, Dicke, and Pfeiffer (2019) note that questionnaires used in schools to obtain standardized perceptions have received little attention in scientific research while rarely being used for teaching evaluation in OECD countries. In order to obtain SPIQ in an objective, reliable, and valid way and based on sound theory, there is a growing body of research on construct and factorial evidence (Fauth et al., 2014, Ferguson, 2012, Göllner et al., 2016, Kunter & Baumert, 2007, Kyriakides et al., 2014, Lüdtke et al., 2006, Lüdtke et al., 2007, Marsh et al., 2019, Praetorius et al., 2018, Scherer et al., 2016, Wagner et al., 2013, Wagner et al., 2016, Wallace et al., 2016, Fauth et al., 2014; Ferguson, 2012; Ferguson & Danielson, 2014; Gaertner & Brunner, 2018; Göllner et al., 2016; Kunter & Baumert, 2007; Kyriakides et al., 2014; Lüdtke et al., 2006; Lüdtke et al., 2007; Marsh et al., 2019; Praetorius et al., 2018; Scherer et al., 2016; Wagner et al., 2013; Wagner et al., 2016; Wallace et al., 2016). To conceptualize perceptions of instructional quality, previous research on SPIQ has used different frameworks, the most prominent being the model of three basic dimensions, namely classroom management, student support, and cognitive activation (Klieme et al., 2006; Praetorius et al., 2018).

To obtain SPIQ, school teachers partly use ad-hoc instruments that are mainly based on everyday assumptions and not on sound theory (Ory & Ryan, 2003), but even for instruments that are based on sound theory and for which validity checks have been conducted, in many cases, studies do not take the nested structure of student data into account (Lüdtke et al., 2007). The majority of previous studies examining the factorial structure of questionnaires for SPIQ in primary and secondary education have only considered individual student responses as the unit of analysis (e.g., Coffey & Gibbs, 2001; Jackson et al., 1999; Lenske, 2016; Shevlin et al., 2000; Watkins, Marsh & Young, 1987).

[*] Corresponding author. Lehrstuhl für Schulpädagogik, Universität Augsburg, Universitätsstr. 10, 86159, Augsburg, Germany.
E-mail address: benedikt.wisniewski@phil.uni-augsburg.de (B. Wisniewski).

Therefore, Marsh et al. (2019) describe the use of doubly latent multilevel models for the examination of SPIQ as a major research desideratum. Beyond the necessity of this approach to adequately model student responses, we consider it an important research goal to show that doubly latent multi-level models can also be applied to contexts such as school subjects, school types, or grade levels when examining instructional quality.

Considering the growing body of research on instructional quality based on the student perspective, there is a need for knowledge regarding the reliability and validity of SPIQ and about the properties of the used data. In this article—after a brief outline of the state of research on students' ability to provide meaningful assessments of teaching characteristics and the dimensionality of instructional quality—we investigate the factorial structure and measurement invariance of SPIQ in secondary education across different subject groups, school types, and grade levels, considering a two-level structure of the data both for confirmatory factor analysis and measurement invariance testing, and relate them to teachers' self-assessments of their instructional quality.

1.1. Can students adequately assess instructional quality?

When students are asked about characteristics of instructional quality, it is essential to ensure that their answers to measures of SPIQ are reliable, valid, and fair. In the following, we sum up previous findings from secondary education and supplement them with research from higher education. In doing so, we also consider differences between both contexts—such as in teaching and learning methods, the learning environment, and the personal responsibility of learners for their achievement (Hassel & Ridout, 2018; Helmke et al., 2008).

1.1.1. Reliability

In the K-12 context, correlations of SPIQ between several student responses regarding the same teacher or course are large ($r = .70$ to $r = .87$, Kyriakides et al., 2014). There are strong similarities between responses to different courses offered by the same teacher, but little to no similarities between responses to the same courses offered by different teachers, suggesting that SPIQ depend more on a teacher's behavior in a given course than on particular subjects and their content (Richardson, 2005). While specific aspects, such as the content of the items, depend on the questionnaire at hand, these findings strongly support the notion that secondary students' perceptions of instructional quality can be reliably assessed.

1.1.2. Validity

Evidence for the validity of the interpretability of student perceptions as a measure of instructional quality can be based, among others, on item content, relations to other variables, and the internal structure of students' answers to the measure (AERA, APA, & NCME, 2005).

An important aspect of validity based on item content is unfairness as a source of construct-irrelevant variance, favoring subgroups or individuals. K-12 SPIQ are often criticized regarding fairness (Goe et al., 2008), assuming that they are influenced by grades and interests, teacher popularity, or the attractiveness of the subject rather than observable teacher behavior. Such fairness concerns have been widely refuted for college and university students (Aleamoni, 1987; Feldman, 2007; Marsh, 2007; Ory & Ryan, 2001; Richardson, 2005) and also for K-12 students, who are able to discriminate between effective and ineffective teaching and are no more prone to grading leniency effects or other distortions and bias than university students (Follman, 1992, 1995).

Some studies have provided evidence for relations to student achievement that speak for the predictive validity of K-12 SPIQ (Ferguson, 2012; Kane et al., 2014; Kuhfeld, 2016), with the most predictive aspects of student perceptions being related to a teacher's ability to control a classroom and to challenge students with rigorous work. Moreover, students' motivations are influenced by perceived teacher support and cognitive activation (Dietrich et al., 2015; Fauth et al., 2014; Klieme, Pauli, & Reusser, 2009; Scherer et al., 2016). Nonetheless, the existing findings are not conclusive: Praetorius, Klieme, Herbert, and Pinger (2018) reviewed results across different studies on predictive evidence for SPIQ regarding student achievement and student motivation and found inconsistent results. They name multiple reasons for the inconsistency of findings and emphasize construct validity as a requirement for predictive validity.

Apart from predictive evidence, comparisons of SPIQ with others' perspectives constitute another aspect required to validly interpret SPIQ. Agreement between teachers, students, and external observers when assessing instructional quality are low to moderate, ranging from $r = -.28$ to .50 (Clausen, 2002, Van der Scheer et al., 2019, Wagner et al., 2013). Agreement between teachers' and students' assessments is higher for dimensions that are easy to observe such as classroom management ($r = .64$), and lower for less observable dimensions such as cognitive activation ($r = .09$) or interaction pace ($r = .10$; Kunter & Baumert, 2007). Again, as predictive validity, this aspect of criteria validity also depends on the construct underlying the measurement of SPIQ.

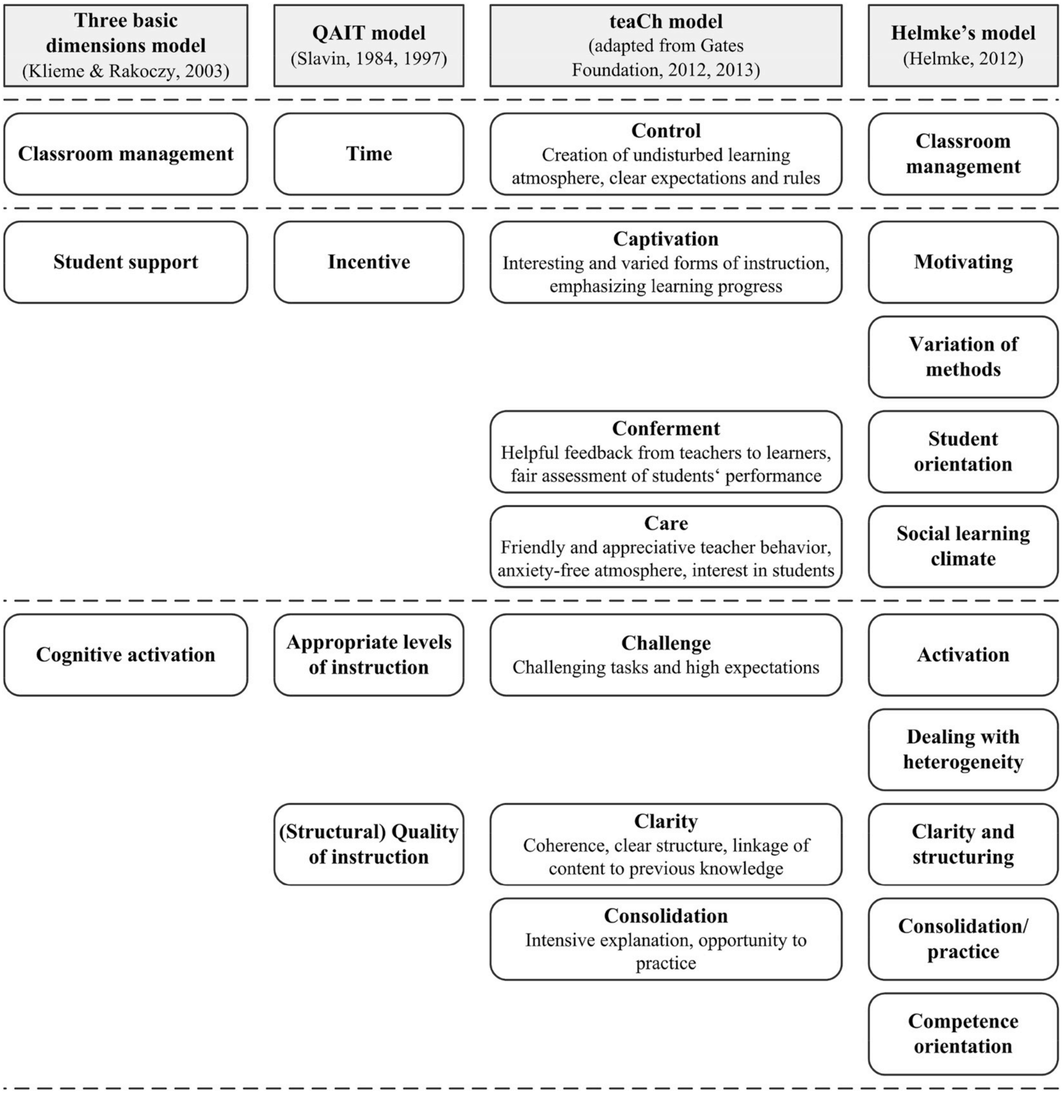
Existing research has already identified different models that describe the internal structure of SPIQ in a given measure and presented evidence for construct validity (see section 1.2 for further details on these models). However, in some cases, theoretically assumed models could not be statistically confirmed and in other cases, models were confirmed, but without considering the nested data structure. Taken together, existing research suggests that K-12 SPIQ assessments sufficiently fulfill a series of basic validity aspects. However, more research is needed on the internal structure, particularly based on the premise that the nested data has to be adequately modelled, which has often been ignored in previous research.

1.1.3. Generalizability

Generic instructional quality is understood as a concept that is supposed to be relatively stable (Wagner et al., 2016) and that can be assessed in any lesson, no matter the subject, irrespective of school types or grade levels.

Correlations of SPIQ between one school year and the next are large ($r > .80$, Kyriakides et al., 2014), which is consistent with findings from the higher education context, where students are able to provide assessments with high test-retest-reliability even for extended periods between test times (Carle, 2009; Marsh, 2007). Considering the differential stability (or generalizability across time) of SPIQ, Kunter and Voss (2013) distinguish surface structures (characteristics that are directly observable, e.g., social forms, forms of teaching, methods, media use) from deep structures of teaching (characteristics that become visible through the interpretation of the teaching process, teaching-learning processes, and interaction). While deep structures are assumed to be more stable over time, surface structures can vary heavily from lesson to lesson (Praetorius et al., 2014). Teaching characteristics that are defined by students' preconditions, teacher-student interactions (i.e., motivation, orientation towards student interests, comprehensibility), or situational factors (Gaertner & Brunner, 2018; Kane, 2013) are less suitable for comparisons across different contexts (Göllner et al., 2016; Wagner et al., 2013). Given this background, configural and metric invariance of SPIQ have been attested for certain instructional characteristics such as lesson structuring and classroom management across different subjects and measurement time (Gaertner & Brunner, 2018; Göllner et al., 2016). However, indications for impaired generalizability have also been reported, with humanities and languages tending to be assessed more favorably than natural and social sciences (Feldman, 2007) and stability being moderated by grade level (Gaertner & Brunner, 2018).

Taken together, these findings show the limitations of SPIQ comparisons across time and different contexts. However, generalizability is



Three basic dimensions model (Klieme & Rakoczy, 2003)	**QAIT model** (Slavin, 1984, 1997)	**teaCh model** (adapted from Gates Foundation, 2012, 2013)	**Helmke's model** (Helmke, 2012)
Classroom management	**Time**	**Control** Creation of undisturbed learning atmosphere, clear expectations and rules	**Classroom management**
Student support	**Incentive**	**Captivation** Interesting and varied forms of instruction, emphasizing learning progress	**Motivating**
			Variation of methods
		Conferment Helpful feedback from teachers to learners, fair assessment of students‘ performance	**Student orientation**
		Care Friendly and appreciative teacher behavior, anxiety-free atmosphere, interest in students	**Social learning climate**
Cognitive activation	**Appropriate levels of instruction**	**Challenge** Challenging tasks and high expectations	**Activation**
			Dealing with heterogeneity
	(Structural) Quality of instruction	**Clarity** Coherence, clear structure, linkage of content to previous knowledge	**Clarity and structuring**
		Consolidation Intensive explanation, opportunity to practice	**Consolidation/ practice**
			Competence orientation

Fig. 1. Comparisons of models of instructional quality.

a crucial prerequisite for using SPIQ to compare instructional quality characteristics obtained in different contexts. While it stands to reason that low-inference dimensions of instructional quality are more suitable for comparisons, more research is required, especially as evidence on generalizability can be influenced by the clustered structure of the data. The, to the best of our knowledge, hitherto underinvestigated combination of multi-level analysis and measurement invariance testing is essential in order to meet this requirement.

1.2. Models of instructional quality in secondary education

Seeking high construct validity, assessing instructional quality requires models that adequately map the very construct. SPIQ can be used to assess the deep structure of teaching, that is, how interesting, difficult, or understandable instruction generally is (Gaertner & Brunner, 2018). Most pertinent models for the secondary education context consist of a common core of instructional characteristics, including classroom management, structuring, motivation, and support, but these are grouped differently and vary in their grade of differentiation and dimensionality, with each comprising specific instructional aspects (see Fig. 1 for an overview, and Praetorius & Charalambous, 2018, or Seidel & Shavelson, 2007, for more detailed comparisons). Particularly prominent is the concept of three basic dimensions of instructional quality (Klieme et al., 2006, Praetorius et al., 2018, Klieme et al., 2006; Praetorius et al., 2018). Additionally, a 4-factor-model (Slavin, 1994; 1997), a 10-factor-model (Helmke, 2012), and a 7-factor-model (Ferguson & Danielson, 2014; Gates Foundation, 2012, 2013) were proposed.

The model of three basic dimensions (Klieme et al., 2006; Praetorius et al., 2018) has a parsimonious structure and a very strong theoretical foundation based on general theories of instruction and psychological theories of student cognition and motivation alike. The three dimensions of classroom management, student support, and cognitive activation have been verified by confirmatory factor analyses (Fauth et al., 2014, Künsting et al., 2016, Kunter & Voss, 2013, Fauth et al., 2014; Künsting et al., 2016; Kunter & Voss, 2013) and Praetorius et al. (2018) presented evidence that dimensions found in studies that identify more than the three factors are often close or identical to these. Moreover, the measurement of the three basic dimensions has been found to be invariant across different countries (Scherer et al., 2016).

Two conceptually similar 7-factor models (Ferguson & Danielson, 2014; Gates Foundation, 2012, 2013; Wisniewski & Zierer, 2020) reflect these three basic dimensions and further distinguish three sub-dimensions each for cognitive activation and student support. While having a weaker theoretical and empirical foundation than the model of three basic dimensions, they offer the advantage of providing practitioners with an easy to interpret dimensionality while standardized instruments exist to gather feedback from students: the Tripod questionnaire (Gates Foundation, 2012, 2013) and the teaCh questionnaire (Wisniewski & Zierer, 2020). The latter was developed for German-speaking students and previous results from conventional exploratory and confirmatory factor analyses suggest that it measures SPIQ on the basis of a measurement model that is very similar to the 7 Cs from the Measures of Effective Teaching (MET) project, sharing the seven category titles, but not being completely identical to it (see Fig. S1 in the supplement).

The differentiation of seven factors as sub-dimensions of the three basic dimensions from a theoretical perspective can be justified based on different existing frameworks and conceptualizations. The included seven dimensions are rooted in extensive literature on the relevance of time spent on task as a central prerequisite of effective learning processes and the essential necessity to create a learning atmosphere enabling this (dimension “control” as related to the basic dimension “classroom management”; e.g., Brophy, 2000; Kuger, 2016). Moreover, they refer to comprehensive research documenting the fundamental role that motivational and social dimensions of instruction as well as feedback play in self-regulated and co-constructive learning processes (dimensions “captivation”, “conferment”, and “care” as related to the basic dimension “student support”; e.g., Benning, Praetorius, Janke, Dickhäuser, & Dresel, 2019; Hattie & Timperley, 2016; Patrick, Kaplan & Ryan, 2011; Rakoczy, 2008). According to Taut and Rakoczy (2016), student support can be divided into organizational choices (provision of choice, individualization) and supportive social aspects (teacher-student relationship), a separation of aspects closely reflected in the dimensions of captivation and care. Further, the dimension of conferment as an additional aspect of student support is characterized by instructional behavior that enables teachers and students to judge progress toward learning goals (Seidel & Shavelson, 2007; Taut & Rakoczy, 2016) and the teacher being sensitive to individual needs (Praetorius et al., 2018). Finally, the dimensions included in this model build on work demonstrating the importance of challenging and stimulating problems for enabling active cognitive processing, on the one hand, and clarity, structuring, and practicing, on the other hand, for developing understanding, knowledge, and their long-term availability (dimensions “challenge”, “clarity”, and “consolidation” as related to the basic dimension “cognitive activation”; e.g., Brophy, 2000; Klieme, Pauli, & Reusser, 2009; Lipowsky et al., 2009). The theoretical understanding of learning that underlies the model refers to learning as an active, constructive, self-regulated, and social process that depends on cognitive, motivational, and emotional characteristics of the learner, the described instructional characteristics of the learning environment (as a learning opportunity that is used more or less), and their interaction (e.g., Helmke, 2012). Praetorius et al. (2018) define cognitive activation by teachers exploring and building on students’ prior knowledge and ways of thinking and by providing students with challenging problems and questions in order to engage them in higher-level thinking processes. These two aspects are reflected in the dimensions of clarity and challenge. Clarity is apprehended as a distinct factor by several researchers (Clausen, Reusser, & Klieme, 2003; Schlesinger & Jentsch, 2016; Seidel & Shavelson, 2007) who define it as a dimension focusing on structural transparency and goal orientation. Furthermore, the aspect of consolidating knowledge can be identified as a separate factor (Schlesinger & Jentsch, 2016; Taut & Rakoczy, 2016), characterizing how teachers help students organize material by practicing and connecting ideas, leading to better retention, multiple pathways for knowledge retrieval, and more effective reasoning (Ferguson, 2012).

At first glance, the separability of some of the aforementioned dimensions may not be obvious. In particular, the dimensions of “captivation” and “challenge” both refer to motivational prerequisites of teaching. However, “captivation” includes specific characteristics of the instruction (interesting, varied, relevant, demanding, useful) that enable students to experience self-efficacy (Ryan & Deci, 2000) and expand their competence (Benning et al., 2019), while “challenge” incorporates what teachers demand from their students, reflecting their beliefs about students’ academic capabilities (Rubie-Davies, 2010).

1.3. Two-level modeling of SPIQ

SPIQ are, by definition, focused on the behaviors of individual teachers. In order to map these perceptions in a methodologically adequate way, it is necessary to consider the statistical peculiarities of this form of data. The assumption that students within and between classes are independent of each other and share no common perceptions is violated for nested data—i.e., when students are taught within groups (here: classrooms)—because substantial similarities within groups (here: a common teacher) lead to inaccurate estimates of model parameters, standard errors, and fit indices, as well as an increased risk of type I errors (D'Haenens, Van Damme & Onghena, 2010; Dyer, Hanges & Hall, 2005; Schweig, 2014). Lüdtke et al. (2011) showed that nested structures are particularly salient when students are asked to assess aspects of the school context. Using aggregated student data at the classroom level is crucial when investigating characteristics of instructional quality because students' individual perceptions of this classroom-level construct are exposed to multiple sources of interference at the student level (Marsh et al., 2012) and the shared perception of students, adjusted for these interferences, is of interest. As such, it is necessary to confirm that all findings based upon individual students as the unit of analysis can also be demonstrated at the classroom level (Marsh, 1983), requiring the observation of student perceptions simultaneously on an individual student level and a clustered classroom level, which can be done using multilevel confirmatory factor analysis.

Even when considering different levels of analysis, educational researchers run the risk of reporting distorted results when applying models to different contexts (Schwab & Helm, 2015). A model that works for one specific school type or one specific school subject may not be suitable for a different context (e.g., that can exhibit different instructional characteristics) or different groups of people (e.g., who might understand items differently). Therefore, the analysis of instructional quality not only requires the consideration of different levels, but also the consideration of different contexts and groups of respondents. To make sure that a model can be applied to them, measurement invariance has to be confirmed.

1.4. Research questions and theory-based expectations

The purpose of the present study was to illuminate central aspects of reliability and validity of SPIQ in secondary education. We wanted to test whether reliable assessments of instructional quality are possible with classroom-sized samples. Based on previous research, the assumptions of a multi-dimensional construct and students' general ability to provide meaningful assessments thereof are plausible. However, research requires an empirical demonstration of a factorial structure that fits theoretical considerations on a student and classroom level and that can be applied to multiple contexts (different school types, school subjects, or grade levels) on both levels. Finally, as a criterion of validity from relations to other variables, we wanted to demonstrate that SPIQ sufficiently reflect teachers’ self-perceptions.

To this end, we put forward five specific hypotheses:

Hypothesis 1. Average-sized classes (25 students) are large enough to obtain reliable assessments of instructional quality based on the underlying 7 factors as facets of three basic dimensions.

Hypothesis 2. SPIQ can adequately be described with seven factors and three superordinate factors on the student level and the corresponding factors on the classroom level. Specifically, we expected both the 7-factor model and the 7-factor model with 3 superordinate factors to describe the data well and better than alternative models (that subsume or disentangle different facets of instructional quality).

Hypothesis 3. A large part of the variation of SPIQ lies on the between-classroom level (given the fact that student perceptions are substantially related to instructional characteristics of different teachers and to student-teacher relationships).

Hypothesis 4. The measurement model is invariant across school types, subject types, and grade levels.

Hypothesis 5. SPIQ are positively correlated with teachers' self-perceptions of instructional quality.

2. Method

2.1. Procedure and sample

We surveyed a sample of 15,005 student perceptions from 690 classes in grades five to twelve in eight German schools from three different school types ($n = 6{,}005$ from three university preparatory high schools/Gymnasium, $n = 6212$ from three intermediate secondary schools/Realschule, $n = 1{,}192$ from two vocational schools/Berufsschule, and $n = 504$ with no specification) during the period of September 2017 to October 2018 via an online feedback portal. Teachers were able to conduct surveys on their students' perceptions of instructional quality and compare these to their self-perceptions. The data originated from an everyday school context, rather than being obtained for research purposes.

With the approval of their school administrations, the participating schools provided consent to use their data in an anonymized form and acquired the written consent of teachers and parents. Due to the technical nature of the online portal, no personalized student data were obtained. Any personalized teacher data was anonymized before it was transferred to us for analysis.

Students participated regarding the subjects of humanities and languages (4,209 students), math, sciences, and IT (4,671 students), and social studies (2,559 students). For 1,171 students, the subject was not specified. Data from 892 students had to be excluded from the analyses as they answered less than one third (9 of the 29) of the items (in which case, serious participation was unlikely). Two hundred and seventy-one teachers took part in the survey. A total of 172 teachers participated with more than one class and 421 classes rated more than one teacher. The average cluster size of the classes was 22.6 students. For 392 classes, teachers also made self-assessments. The data include identification codes for teachers and information on school type, subject, and grade level.

The assessed school types provide a heterogenous database reflecting the variety of different school types in the German school system—however, they do not constitute a representative sample on the school level in a narrow sense. Because teachers decided to use the online feedback portal voluntarily, the sample is restricted to such teachers who were willing to reflect their teaching based on student feedback. This might have led to a positive selection of teachers. All participating schools are from the same German federal state (Bavaria), which furthermore restricts the representativity due to Germany's federal structure.

2.2. Measures

To measure SPIQ, we used the German questionnaire for student perceptions in secondary education (teaCh). This questionnaire comprises 29 items that refer to seven categories reflecting the three basic dimensions of instructional quality (see Fig. 1), with three items for “care”, six items for “control”, four items for “conferment”, four items for “clarity”, two items for “challenge”, four items for “consolidation”, and six items for “captivation”. All 29 items and their reliability indices are listed in the appendix. They were presented with four-point Likert-type scales, ranging from 1 (*I don't agree) to* 4 (*I agree*). The items for teachers were identical to the student version but formulated from the teachers' perspective.

2.3. Statistical analyses

For data analysis we used R-Studio (version 1.0.136 for Mac; R Core Team, 2014) and the multilevel package for R (version 2.6; Bliese, 2016) for the calculation of intraclass correlations and interrater agreement, and Mplus (version 8.1; Muthén & Muthén, 2018) for multilevel CFA and measurement invariance testing.

First, we conducted a conventional one-level confirmatory factor analysis on the student level in which we tested the seven-factor model by using the total sample covariance matrix and correcting the standard errors with Taylor-series linearization (Muthén, 1994). Then, we examined the extent of variation within (student level) and between classes (classroom level). For this purpose, the intraclass correlations (ICC1 and ICC2) were computed as a measure for the degree of dependence or clustering of the data within classrooms of different teachers with ICC1 representing the proportion of variance of individual-level outcomes explained by group membership and ICC2 representing the observed total variance in classroom average scores occurring at the classroom level (and therefore the inter-individual reliability of aggregated mean scores; Shieh, 2016). ICC1 values exceeding .05 are indicative of substantial correlations among variables between and within, implying that a multilevel analysis is necessary to describe the data well (Dyer et al., 2005) and ICC2 values above .60 allow a meaningful aggregation of individual-level data on a group level (Bliese, 2000; Chen, Mathieu & Bliese, 2004). Because for some teachers, more than one course was assessed (courses nested within teachers) and because teachers were nested in schools, level-2-ICCs and level-3-ICCs (Hedges et al., 2012) were calculated for these additional levels.

The aggregation of data requires confirmation that the underlying construct has an identical meaning on the level of the student and on the level of the classroom (Lüdtke et al., 2006). To investigate the extent of absolute agreement between individual student ratings within a classroom, we used two indicators of interrater agreement, $r_{WG(j)}$ (James et al., 1984) and $AD_{W(j)}$ (Burke and Dunlap, 2002). For four-level Likert-type scales as used in the present research, $r_{WG(j)} > .70$ and $AD_{W(j)} \leq .67$ indicate satisfactory interrater agreement (Lüdtke et al., 2006).

Regarding the required number of students for reliable assessments, we calculated the theoretical sample sizes that are required to get reliable information from the used items. Because ICC1 and ICC2 are related to each other as a function of group size per item (Bliese, 2000), they can be used to calculate the theoretically assumed sample sizes that provide different levels of reliability for each item, using Cicchetti's (1994) suggested guidelines for interpretation (fair reliability from .40, good reliability from .60, excellent reliability from .75).

To examine the factorial structure of our data, while controlling for the nested data structure, we subsequently conducted a two-level confirmatory factor analysis (MCFA), using separate within- and between-group covariance matrices. For the assessment of the model fit, conservative fit statistics were used ($\chi^2/df \leq 3$, CFI ≥ .95, TLI ≥ .95, RMSEA ≤ .08, SRMR ≤ .08, Hu & Bentler, 1999). The hypothesized models (seven factors and seven factors with three superordinate factors) were compared to each other and against different alternative models (that were based on theoretical as well as statistical considerations). For model comparisons, we investigated Satorra-Bentler-scaled χ^2-difference tests, and the common cut-off values of ΔCFI = .01 and ΔRMSEA = .015 (Chen, 2007). As indicated in Fig. 1, loadings from the first-order factors captivation, conferment, and care on the second-order factor student support, loadings from the first-order factors challenge, clarity, and consolidation on the second-order factor cognitive activations and a loading from the first-order factor control on the second-order factor classroom management were modelled. To adequately model the second-order factor “classroom management” (that consists of only one first-order factor), we set the loading of the first-order factor on the second-order factor and the variance of the second-order factor to one.

In a next step, we tested measurement invariance of both assumed models to see if the supposed factorial structure can be applied independently of respective school types, subjects, and grade levels. We report a detailed description of our approach in the electronic supplement. In contrast to previous research on SPIQ measurement invariance (Scherer, Nilsen, & Jansen, 2016), we used a two-level model to test for measurement invariance of the hypothesized models. Accordingly, we met the demand to examine whether a common factorial structure is tenable not only on different levels, but also across different groups.

In order to add an external criterion of validity to our considerations, we investigated the convergence between students' and teachers' perceptions (by estimating correlations between students' perceptions and teachers’ self-perceptions of the aspects of instructional quality on the between level, as well as conducting multi-group analyses to investigate whether the covariances between these aspects differed between students and teachers).

3. Results

3.1. Descriptive results

All item means were slightly above the theoretical mean (see appendix) and the data contained a significant portion of inter-individual variance (Table 1). Responses were approximately normally distributed with skewness ranging from −0.97 to −0.57 and kurtosis values ranging from −0.49 to 0.13.

The ICCs for all of the items were above the conventionally used cut-off values (Chen, Mathieu, & Bliese, 2004) with a median of .11 for ICC1 (range: .08–.15) and a median of .75 for ICC2 (range: .66–.81, see appendix). These values indicate a between-level variability that requires multilevel analysis. Considering a third and fourth level of analysis (courses within teachers and teachers within schools) indicated that the observed variance of the average intra-teacher-scores and intra-school-scores did not require further levels of analysis (ICC1s < .001).

Interrater agreement was satisfactory for all scales except for “captivation”, which had a $r_{WG(j)}$ value slightly below the critical cut-off value of .70. Furthermore, our results indicated that at least 9 students are needed for fair reliability, 19 students for good reliability and 30 students for excellent reliability of all 29 items. For example, the ICC1 for the item “care 1” is .11. Using the Spearman-Brown-formula (which describes ICC2 as a function of ICC1 and sample size), the sample size needs to be at least 30 to reach the cut-off of ICC2 = .75 indicating excellent reliability.

Table 1
Descriptive statistics for factors on the student level.

Factor	Items	*M*	*SD*	ω	ICC1	ICC2	$r_{WG(J)}$	$ADw_{(J)}$
Care	3	3.12	0.90	.82	.08	.87	.83	.63
Control	6	3.22	0.86	.84	.09	.88	.92	.58
Conferment	4	3.10	0.87	.85	.07	.85	.87	.63
Clarity	4	3.10	0.86	.74	.07	.84	.88	.55
Challenge	2	3.18	0.87	.77	.08	.86	.80	.60
Consolidation	4	3.02	0.90	.83	.08	.86	.85	.67
Captivation	6	3.05	0.86	.84	.07	.84	.68	.58

Notes. ω reflects McDonald's Omega on the subscale level. ICC1 and ICC2 represent intra-class correlations on the subscale level, $r_{WG(J)}$ and $ADw_{(J)}$ represent absolute interrater agreement. Descriptive statistics for single items can be found in the appendix.

Results of a conventional one-level CFA, in which only the individual student level was considered, indicated a good fit for the seven-factor model on the student level (χ^2/df = 3.40, $p < .01$; CFI = .984; TLI = .982; RMSEA = .013; SRMR = .017). As this type of analysis does not reflect the nested data structure adequately, we subsequently estimated two-level CFAs.

3.2. Two-level confirmatory factor analyses

Results of the two-level CFAs showed an excellent fit to the data of the 7-factor model (Table 2). Adding the three superordinate factors yielded a slight decrease in model fit (which was not substantial regarding the CFI and RMSEA values, however the χ^2-difference tests were statistically significant). All factor loadings were significantly different from zero ($p < .01$). The standardized loadings for all items ranged from .43 to .60 on the student level and from .60 to .97 on the classroom level, demonstrating that the seven factors explain large proportions of the variation in item responses on the aggregated level. Intercorrelations between factors were large and positive (within level: .56–.89, between level: .40–.93). All factor loadings and intercorrelations between factors for the 7-factor-model are shown in Fig. 2. The intercorrelations of the superordinate factors were .25 for classroom management with student support, .24 for classroom management with and cognitive activation, and .95 for student support with cognitive activation.

Following theoretical consideration, we compared four additional models (see Table 2). Regrouping our items according to Helmke's (2012) 10-factor model, Slavin's 4-factor model (1984, 1987), the three basic dimensions of instructional quality (Klieme et al., 2006), and an undifferentiated single factor model described the data significantly worse. Furthermore, based on statistical considerations (large factor correlations), we estimated another set of alternative models that also described the data inferiorly (see Table S2 in the supplement).

3.3. Measurement invariance for subject type, school type, and grade levels

Next, we investigated the measurement invariance of the hypothesized models. We summarize the results for the 7-factor model in Table 3 (and the results of the 7-factor model with the three superordinate factors in Table S1 in the supplement). The analyses for subject groups, school types, and grade levels showed that the more parsimonious models did not fit the data worse than the models allowing for between-group variations (as indicated by the statistically non-significant χ^2-difference tests and the CFI and RMSEA differences). As the more restrictive models did not describe the data statistically significantly worse than the less restrictive models, strict measurement invariance can be assumed across the investigated subject groups (humanities/languages, math/sciences, and social sciences), school types (university preparatory high school, secondary intermediate school, vocational school), and grade levels (5–7, 8–10, 11–12).

3.4. Convergence between students' perceptions and teachers’ self-perceptions

Latent correlations between students' perceptions and teachers’ self-perceptions of instructional quality were substantial for all seven factors on the classroom level of analysis, with a median of .59 (range: .49–.78, see Table 4). Students and teachers shared between 24.0% and

Table 2
Fit statistics and comparisons between expected and alternative models of SPIQ.

	χ^2	*df*	χ^2/df	CFI	TLI	RMSEA	SRMR within	SRMR between	ΔCFI	ΔRMSEA	TRd	Δ*df*	*p*
Hypothesized models													
7-factor model	1491	712	2.90	.984	.982	.009	.017	.074					
7-factor model with 3 s order factors	2052	736	2.76	.972	.969	.011	.021	.071	.006	.002	10,692	22	<.001
Alternative models													
10-factor model	3004	664	4.52	.953	.943	.015	.026	.091	.031	.006	1015	48	<.001
4-factor model	4409	742	5.94	.927	.920	.018	.032	.101	.057	.009	13,083	30	<.001
3-factor model	3487	748	4.66	.945	.940	.015	.029	.098	.039	.006	11,264	36	<.001
1-factor model	4560	745	6.12	.924	.918	.032	.018	.100	.060	.023	17,371	33	<.001

Notes. All models were compared with the 7-factor-model; TRd, Satorra-Bentler-scaled χ^2-difference test. See Table S2 in the electronic supplement for further alternative models.

60.8% of common variance, with particularly high correlations for the factors “clarity” and “control”.

Finally, as a second aspect of convergence between students' and teachers’ perceptions, we tested if the distinguished aspects of instructional quality were similarly related to each other for students and teachers. Multi-group modeling indicated no statistically significant differences in between-factor correlations between students and teachers (see electronic supplement for a detailed description of these findings).

4. Discussion

The aim of the present study was to investigate the characteristics of secondary students' perceptions of instructional quality, specifically their factorial structure and generalizability. Multilevel methods were applied to explore the factorial structure of SPIQ at an individual and a classroom level and across different contexts (school type and subject). Strengths include the consistent consideration of both levels of analysis, the focus on secondary education students as a rather underinvestigated

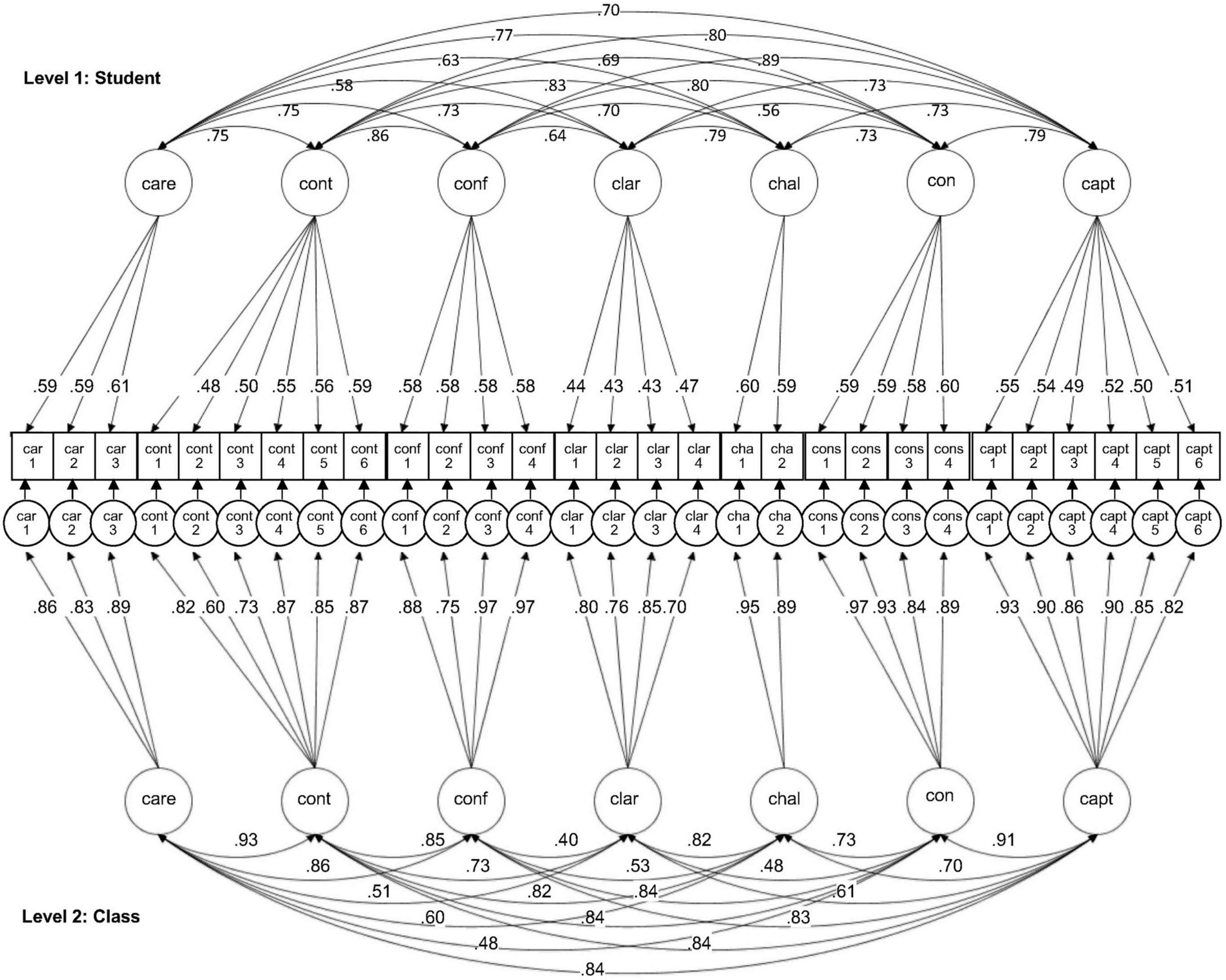


Fig. 2. Measurement model of the seven dimensions of SPIQ (care, care; cont, control; conf, conferment; clar, clarity; chal, challenge; con, consolidate; capt, captivation). Residual variances are not reported.

Table 3
Measurement invariance of the 7-factor model for subject types, school types, and grade levels.

	χ^2	*df*	χ^2/df	CFI	TLI	RMSEA	SRMR within	SRMR between	ΔCFI	ΔRMSEA	TRd	Δ*df*	*p*
Subject type													
Configural Invariance	3296	2136	1.54	.968	.968	.019	.019	.070					
Metric invariance	3356	2224	1.51	.970	.970	.020	.020	.077	.002	.001	80.25	88	.71
Scalar invariance	3421	2282	1.50	.970	.970	.020	.020	.078	.000	.000	65.56	58	.23
Strict invariance	3533	2396	1.48	.972	.972	.021	.021	.082	.002	.001	126.79	114	.19
School type													
Configural invariance	4225	2136	1.98	.966	.961	.015	.017	.098					
Metric invariance	4092	2224	1.84	.970	.967	.014	.018	.103	.004	–.001	57.91	88	.99
Scalar invariance	4436	2282	1.94	.965	.963	.015	.018	.103	.005	.001	53.40	58	.66
Strict invariance	4377	2396	1.83	.968	.967	.014	.019	.104	.003	–.001	123.73	114	.25
Grade level													
Configural invariance	3148	2136	1.47	.979	.976	.010	.012	.065					
Metric invariance	3232	2224	1.45	.979	.977	.010	.012	.071	< .001		93.93	88	.31
Scalar invariance	3298	2282	1.45	.979	.977	.011	.013	.071	< .001	.001	64.82	58	.25
Strict invariance	3435	2396	1.43	.978	.978	.011	.014	.078	.001	.001	144.10	114	.08

Notes. TRd, Satorra-Bentler-scaled χ^2-difference test. Subject types: humanities and languages (n = 4,209), math, sciences, and IT (n = 4,671), and social studies (n = 2,559). School types: university preparatory high schools (n = 6,005), intermediate secondary schools (n = 6,212), vocational schools (n = 1,192). Grade levels: grades 5–7 (n = 5,901), grades 8–10 (n = 4,647), grades 11–12 (n = 3,445).

Table 4
Estimated student-teacher correlations of the latent variables on the between level.

	ρ	S.E.	*p*
Care	.54	.12	.01
Control	.68	.10	< .01
Conferment	.58	.11	< .01
Clarity	.78	.11	< .01
Challenge	.49	.15	< .01
Consolidation	.56	.11	< .01
Captivation	.60	.08	< .01

Notes. ρ represents the correlations of students' perceptions with teachers' self-perceptions on the latent factors.

population in comparison to college and university students (Marsh et al., 2019), and the large sample. Our findings demonstrated that a 7-factor model with the three underlying basic dimensions of instructional quality (Klieme et al., 2006; Praetorius et al., 2018) can be applied to student perceptions of teaching, with classroom-sized samples being large enough to obtain reliable assessments. This worked (a) on both levels, (b) for students and teachers, and (c) invariantly across different contexts, while students' perceptions converged substantially with their teachers' self-perceptions.

Our findings on internal consistencies, reliabilities of the aggregated class means (ICC2), and interrater agreement suggest adequate reliability. In terms of ICC2 values, our results align very well with previous findings (Gaertner & Brunner, 2018; Lüdtke et al., 2006; Praetorius et al., 2018). In contrast, the proportions of individual-level variance explained by class (in terms of ICC1, ranging from .07 to .09) were on the lower bound of variance proportions found in previous research (with ICC1s ranging between .09 and .42 for different dimensions of instructional quality; Gaertner & Brunner, 2018; Lüdtke et al., 2006; Praetorius et al., 2018). Interestingly, we observed the lowest proportions for “conferment”, “clarity”, and “captivation” (dimensions that partly include high-inference items, e.g., on student performance feedback, personal student learning progress, student pre-knowledge) and highest for “control” (which includes rather low-inference items). Nonetheless, the ICCs were not consistently higher for low-inference items than for high-inference items. This might be traced back to different students interpreting items differently (Gitomer, 2019)—and therefore has to be perceived as a hint towards a necessary revision of the used questionnaire in the future, that should not be implemented without caution in the current version. For example, for the item “during the lesson there were plenty of opportunities to practice the new content” a higher ICC should be expected than for “the teacher gave me helpful feedback on my performance”. Not having found differences in their ICC1s might be due to “opportunities to practice new content” being related to different occasions among different students, whereas helpful feedback on individual students’ performances may be perceived in a more similar way by different students when the teacher either gives highly differentiated or little differentiated feedback. In a possible revision of the questionnaire, this issue could for example be addressed by replacing interpretable expressions by more distinct expressions (for example “helpful” by “individual and differentiated”). Beyond this, the interrater agreement that we found implies that the used items have a similar meaning on both levels of analysis. One exception to this was the dimension “captivation” for which interrater agreement was slightly below the critical cut-off. This might be attributed to this category containing more student-dependent aspects (perceived personal learning progress, perceived requirement level, perceived learning pace). Finally, calculating intraclass correlations for each item, we provided information on how many students are needed to acquire reliable perceptions of teaching. Even for small classes (n = 19), a good level of reliability can be expected (confirmation of Hypothesis 1), which points to evaluation questionnaires, such as the one used, being practicable in real life classrooms.

Hypothesis 2 was also confirmed, with MCFA providing construct validity evidence for the two hypothesized models that were shown to operate at the individual and the classroom level of analysis. This is in line with previous findings that concluded that a multidimensional structure is more adequate for representing instructional quality than a single factor. The results therefore provide evidence for the validity of the internal structure of the employed measure, but do not allow different models in general to be compared. Regarding the measure at hand, our results do not support the assumption of a bifactor structure with one general and seven group factors as found by Wallace, Kelcey, and Ruzek (2016) for the related Tripod questionnaire. Such a bifactor structure follows the logic of apprehending instructional quality as a general factor related to teacher responsivity that influences all teacher-student interactions on the one hand and a set of domain-specific dimensions on the other hand (Hamre et al., 2014; Wallace et al., 2016). Conversely, we present evidence for seven highly intercorrelated, but distinguishable dimensions that can be apprehended as subdimensions of three superordinate factors that are theoretically incorporated in the model of three basic dimensions of instructional quality (Klieme et al., 2006; Praetorius et al., 2018). That we found generally relatively high intercorrelations between the factors is consistent with existing findings on the factorial structure of student perceptions of teaching (Kane et al., 2014). However, the magnitudes of the loadings also imply that each factor contains a significant proportion of unique variance. Interestingly, intercorrelations were lower on both levels for the dimensions “captivate” and “challenge” which focus on motivational characteristics. Conversely, there were particularly high intercorrelations between “control” and “care” and between “consolidation” and “captivation” on the between level, but not on the student level. This implies that while individually students perceive these categories as different constructs, students taught by one specific teacher perceive them as strongly interconnected (yet, separable) characteristics of teaching. This could be traced back to interpretations of different teaching styles. For example, Walker (2009) defined control and nurturance (which is similar to the “care” factor) as primary dimensions of an authoritative classroom management style. As such, it is plausible that when observing a specific teacher, students perceive control and care quite similarly, as an aspect of classroom management style. Opposed to this, the strong link between “consolidation” and “captivation” is less apparent but might point to the intertwined nature of teaching in which ensuring that students practice and make progress is related to lessons being interesting, varied, and fitting their preconditions.

Considering the 3 s-order factors, a very high correlation between student support and cognitive activation was found, which seems unexpected at first sight. However, this is in line with prior research (Atlay, Tieben, Hillmert, & Fauth, 2019; Fauth, Decristan, Rieser, Klieme, & Büttner, 2014; Schiepe-Tiska, 2019; Wagner et al., 2013; for an overview see; Praetorius et al., 2018), where close associations of a similar size were identified between these two basic dimensions of instructional quality.

All in all, the 7-factor-model describes our data well, but, as the model was originally developed based primarily on statistical considerations (namely EFA), the embedding in a distinguished theoretical framework is ongoing (Wisniewski & Zierer, 2020). A first outline of the theoretical justification of the separability of the considered facets has been presented in this article but opens up a field of action for future research.

In line with our expectations, we further found that a large part of the variation of SPIQ is explained between teachers (confirmation of Hypothesis 3). Particularly, the seven factors explained more variation of the item responses at the classroom level, or in other words for students taught by a specific teacher in a specific classroom, than for the complete sample, resulting in significantly higher factor loadings on the classroom level. This is in line with previous findings on student perceptions being substantially related to instructional characteristics of different teachers and student-teacher relationships (Ditton & Arnoldt, 2004).

Hypothesis 4, namely measurement invariance across the context variables of school type, subject type, and grade level, was also verified. We supplement previous findings on configural and metric invariance (Feldman, 2007; Göllner et al., 2016), by showing scalar and even strict measurement invariance of the employed measure across humanities/languages, math and natural sciences, and social sciences. This supports the idea of a set of basic aspects of instructional quality that are universal and independent of subject-specific characteristics. We demonstrated the generalizability of the 7-factor structure not only across different subjects, but also different school types and grade levels, indicating a high robustness against contextual distortions (although the underlying measure contained—amongst other aspects—motivation-orientated dimensions that are often considered to be less suitable for comparisons across different contexts; Göllner et al., 2016; Wagner et al., 2013). A major concern held by researchers and practitioners is that student perceptions are influenced to a large extent by the attractiveness of the respective subject than by individual instructional characteristics (Aleamoni, 1987). Our findings open up the possibility of comparing SPIQ with a model that works across these contexts. To the best of our knowledge, this is the first study to analyze measurement invariance of SPIQ with a two-level model in the secondary school context. Therefore, we provide a methodological basis for future investigations in this line of research.

Finally, our findings also provide evidence for validity beyond the internal structure. The substantial convergence between students and teachers' perceptions of instructional quality (both in regard to the instructional quality exhibited by particular teachers, as well as the associations between different aspects of instructional quality) attests that students and teachers generally agree in how they perceive instructional quality (confirmation of Hypothesis 5). This is a strong hint towards convergent validity and stands in contrast to previous research that showed that there is often only little concordance between the two perspectives (Clausen, 2002; Van der Scheer et al., 2019; Wagner et al., 2013). Furthermore, these findings partly confirm Kunter and Baumert's (2007) findings that agreement is higher for low-inference characteristics: In the present study, we found the highest for the dimensions “clarity” ($r = .78$) and “control” ($r = .68$) on the classroom level, dimensions that contain items clearly relating to observable teacher behavior, whereas “challenge,” with the lowest correlation on the classroom level ($r = .49$), contains items that are partly influenced by students' preconditions.

When interpreting these findings, certain limitations of the study at hand need to be borne in mind and open up directions for future research. First of all, we could not include bias variables such as gender and age of students and teachers. As previous research indicated that these can affect students’ evaluations of teaching quality in higher education (Boring, 2017; MacNell et al., 2015), future research should specifically incorporate these aspects into statistical models. This becomes particularly important for the interpretation of student perceptions and their associations with other variables (whereas the structural aspects investigated in the current work should be rather independent from that). Following up on such bias factors is an important avenue for future research. Another limitation is the restriction of our data to German schools within one German federal state. While our findings on the measurement invariance across different school types can be considered a first indication for the generalizability of our findings, they also need to be interpreted cautiously as there was only a limited number of schools from each school type. Finally, we could not observe the stability of perceptions over time (e.g., one school year). In order to inspect the professional development of teachers by comparing their results at multiple points of time with measures such as the one at hand, such evidence on the sensitivity of the measure to individual increases or decreases would still be required.

Despite these limitations, practical implications can already be drawn: First, teachers can acquire reliable and valid information on how students assess their way of teaching and can use this information as a basis for personal development concerning relevant teaching dimensions. Second, our results encourage secondary education teachers to trust in their students' ability to assess effective teaching. We were able to point out that a well definable construct of generic instructional quality precisely differentiates between different classes and is independent of context variables like subject, school type, or grade level. Third, it is helpful for teachers to know that relatively small sample sizes are sufficient for reliable assessments of students’ perceptions of teaching.

5. Conclusions

In brief, secondary school students’ perceptions of teaching offer important value to the diagnosis of instructional quality. Our findings suggest that systematically obtained perceptions are a reliable and valid information source. SPIQ provide valuable information for educational research and they may also provide teachers with valuable information that they can use to reflect on their work—an important aspect of successful teaching that is often overlooked.

Author note

Benedikt Wisniewski, Department of School Pedagogy, University of Augsburg; Klaus Zierer, Department of School Pedagogy, University of Augsburg; Markus Dresel, Department of Psychology, University of Augsburg; Martin Daumiller, Department of Psychology, University of Augsburg.

Declaration of competing interest

Benedikt Wisniewski and Klaus Zierer are consultants to softwarea GmbH, which is a software company that provides solutions for online student feedback assessments. The other authors declare no conflict of interest.

Acknowledgements

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

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.learninstruc.2020.101303.

Appendix

Formulations of the Items Used to Assess Instructional Quality and Descriptive Statistics.

Item	Item formulation	*M*	*SD*	γ_1	γ_2	r_{itc}	ICC1	ICC2
Care								
car 1	The teacher met me in a friendly and appreciative way [Die Lehrperson begegnete mir freundlich und wertschätzend].	3.15	0.91	−.83	−.17	.61	.11	.75
car 2	The teacher created an atmosphere free of fear [Die Lehrperson sorgte für eine angstfreie Atmosphäre].	3.07	0.92	−.70	−.38	.53	.10	.74
car 3	The teacher was interested in whether I really learned something [Die Lehrperson interessierte sich dafür. ob ich wirklich etwas gelernt habe].	3.15	0.89	−.80	−.23	.60	.11	.75
Control								
cont 1	During the lesson, clear rules were discernible, which the teacher set and enforced [In der Stunde waren klare Regeln erkennbar, die die Lehrperson vorgab und durchsetzte].	3.18	0.86	−.82	−.09	.56	.11	.76
cont 2	The teacher did not waste time due to delays or idling [Die Lehrperson verschwendete keine Zeit durch Verzögerungen oder Leerlauf].	3.22	0.85	−.88	.05	.52	.15	.81
cont 3	The teacher provided a trouble–free working atmosphere [Die Lehrperson hat für eine störungsfreie Arbeitsatmosphäre gesorgt].	3.23	0.84	−.89	.06	.54	.14	.80
cont 4	The teacher had a good overview of what was happening in the class [Die Lehrperson hatte einen guten Überblick über das Geschehen in der Klasse].	3.24	0.86	−.95	.14	.57	.14	.80
cont 5	When students violated the rules, the teacher intervened quickly and consistently [Bei Regelübertretungen durch Schüler griff die Lehrperson schnell und konsequent ein].	3.24	0.88	−.97	.09	.59	.15	.81
cont 6	The course of instruction was smooth [Die Übergänge zwischen den Phasen waren reibugslos].	3.20	0.87	−.89	.00	.64	.13	.79
Conferment								
conf 1	The teacher assessed my performance fairly [Die Lehrperson beurteilte meine Leistungen fair].	3.09	0.88	−.68	−.31	.62	.11	.76
conf 2	The teacher gave me helpful feedback on my performance [Die Lehrperson gab mir zu meinen Leistungen ein hilfreiches Feedback].	3.14	0.88	−.76	−.22	.60	.09	.76
conf 3	The teacher was fair and unbiased towards me and my classmates [Die Lehrperson hat sich mir gegenüber fair und unvoreingenommen gezeigt].	3.07	0.88	−.64	−.35	.61	.09	.71
conf 4	The teacher gave me meaningful feedback on my contributions [Die Lehrperson hat mir sinnvolle Rückmeldungen zu meinen Beiträgen in der Stunde gegeben].	3.08	0.87	−.66	−.33	.62	.08	.71
Clarity								
clar 1	The lesson had a clearly recognizable thread [Die Stunde hatte einen klar erkennbaren roten Faden].	3.08	0.85	−.61	−.35	.50	.11	.75
clar 2	The teacher showed me what the new content is related to [Die Lehrperson hat mir gezeigt, womit die neuen Inhalte zusammenhängen].	3.10	0.85	−.67	−.23	.47	.11	.76
clar 3	The teacher showed me what I could use the new content for [Die Lehrperson hat mir gezeigt, wofür ich die neuen Inhalte brauchen kann].	3.10	0.87	−.69	−.25	.47	.10	.74
clar 4	The teacher has tied in content that was already known to me [Die Lehrperson hat an Inhalte angeknüpft, die mir schon bekannt waren].	3.10	0.85	−.68	−.23	.49	.12	.77
Challenge								
chal1	The tasks in the lesson were challenging for me [Die Aufgabenstellungen in der Stunde waren für mich herausfordernd].	3.17	0.87	−.79	−.18	.52	.12	.77
chal2	The teacher had high expectations of me [Die Lehrperson hat hohe Erwartungen an mich gestellt].	3.19	0.87	−.84	−.11	.52	.14	.80
Consolidation								
cons 1	During the lesson, learning and practice phases alternated [In der Stunde wechselten sich Lern– und Übungsphasen ab].	2.99	0.90	−.58	−.46	.62	.08	.69
cons 2	During the lesson, the teacher showed me exactly how I could solve certain tasks [Die Lehrperson hat mir genau gezeigt, wie ich eine bestimmte Aufgabenstellung lösen kann].	3.01	0.89	−.60	−.44	.61	.08	.67
cons 3	I had enough time to concentrate on the content of the lesson [Ich hatte genügend Zeit, mich intensiv mit den Inhalten der Stunde zu beschäftigen.]	3.07	0.90	−.69	−.35	.58	.10	.72
cons 4	During the lesson there were plenty of opportunities to practice the new content [In der Stunde gab es ausreichend Gelegenheiten, die neuen Inhalte zu üben].	3.00	0.90	−.57	−.49	.61	.09	.70
Captivation								

capt1	The content of the lesson was taught by the teacher in an interesting way [Die Inhalte der Stunde wurden durch die Lehrperson auf interessante Art vermittelt].	3.03	0.88	–.61	–.37	.57	.07	.66
capt2	The course of the lesson was varied [Der Ablauf der Stunde war abwechslungsreich].	3.06	0.87	–.62	–.34	.60	.11	.75
capt3	I was able to see personal learning progress through the lessons [Ich konnte während der Stunde einen persönlichen Lernfortschritt feststellen].	3.05	0.87	–.60	–.39	.57	.10	.74
capt4	The requirement level in the lesson was appropriate for me [Das Anforderungsniveau der Stunde war für mich angemessen].	3.05	0.86	–.60	–.37	.59	.10	.74
capt5	The learning pace in the class was appropriate for me [Das Lerntempo in der Stunde war für mich angemessen].	3.06	0.85	–.62	–.31	.58	.09	.73
capt6	During the lesson I was able to apply strategies that are also useful for other problems/topics/areas [Im Unterricht konnte ich Strategien anwenden, die auch für andere Probleme/Themen/Gebiete nützlich sind].	3.05	0.85	–.62	–.24	.58	.11	.75

Notes. Presented are translations of the original German items that are not yet validated in the English language. *SD* represents the standard deviation, γ1 the skewness and γ2 the kurtosis. r_{itc} describes the corrected item total correlation on the item level and ICC1, and ICC2 the intra–class correlations on the item level.

References

AERA, APA, & NCME (2005). *Standards for educational and psychological testing.* Washington, DC: AERA.

Aleamoni, L. (1987). Typical faculty concerns about student evaluation of teaching. *New Directions for Teaching and Learning, 1987*, 25–31. https://doi.org/10.1002/tl.37219873105.

Atlay, C., Tieben, N., Hillmert, S., & Fauth, B. (2019). Instructional quality and achievement inequality: How effective is teaching in closing the social achievement gap? *Learning and Instruction, 63*. https://doi.org/10.1016/j.learninstruc.2019.05.008.

Benning, K., Praetorius, A.-K., Janke, S., Dickhäuser, O., & Dresel, M. (2019). Das Lernen als Ziel: Zur unterrichtlichen Umsetzung einer Lernzielstruktur [learning as a goal]. *Unterrichtswissenschaft, 47*, 523–545. https://doi.org/10.1007/s42010-019-00054-7.

Bliese, P. (2000). Within-group agreement, non-independence, and reliability. In K. Klein, & S. Kozlowski (Eds.). *Multilevel theory, research, and methods in organizations* (pp. 349–381). San Francisco, CA: Jossey-Bass.

Bliese, P. (2016). *Multilevel package for R.* Retrieved from https://cran.r-project.org/web/packages/multilevel/multilevel.pdf.

Boring, A. (2017). Gender biases in student evaluations of teaching. *Journal of Public Economics, 145*, 27–41. https://doi.org/10.1016/j.jpubeco.2016.11.006.

Brophy, J. (2000). *Teaching. Brussels: Iae.* Retrieved from http://www.ibe.unesco.org/fileadmin/user_upload/archive/Publications/educationalpracticesseriespdf/prac01e.pdf.

Burke, M. J., & Dunlap, W. P. (2002). Estimating interrater agreement with the average deviation index. *Organizational Research Methods, 5*, 159–172. https://doi.org/10.1177/1094428102005002002.

Chen, F. (2007). Sensitivity of goodness of fit indexes to lack of measurement invariance. *Structural Equation Modeling, 14*, 464–504. https://doi.org/10.1080/10705510701301834.

Chen, G., Mathieu, J., & Bliese, P. (2004). A framework for conducting multilevel construct validation. In F. Yammarino, & F Dansereau (Vol. Eds.), *Research in multilevel issues: Vol. 3*, (pp. 273–303). Oxford, UK: Elsevier.

Cicchetti, D. (1994). Guidelines, criteria, and rules of thumb for evaluating normed and standardized assessment instruments in psychology. *Psychological Assessment, 6*, 284–290. https://doi.org/10.1037/1040-3590.6.4.284.

Clausen, M. (2002). *Unterrichtsqualität [instructional quality].* Münster, Germany: Waxmann.

Clausen, M., Reusser, K., & Klieme, E. (2003). Unterrichtsqualität auf der Basis hochinferenter Unterrichtsbeurteilungen [Instructional quality on the basis of highly-inferent teaching assessments]. *Unterrichtswissenschaft, 31*(2), 122–141.

Clayson, D. (2008). Student evaluations of teaching. *Journal of Marketing Education, 31*, 16–30. https://doi.org/10.1177/0273475308324086.

Coffey, M., & Gibbs, G. (2001). The evaluation of the Student Evaluation of Educational Quality questionnaire (SEEQ) in UK higher education. *Assessment & Evaluation in Higher Education, 26*, 89–93. https://doi.org/10.1080/02602930020022318.

Cohen, P. (1980). Effectiveness of student-rating feedback for improving college instruction. *Research in Higher Education, 13*, 321–341. https://doi.org/10.1007/BF00976252.

Dietrich, J., Dicke, A.-L., Kracke, B., & Noack, P. (2015). Teacher support and its influence on students' intrinsic value and effort. *Learning and Instruction, 39*, 45–54. https://doi.org/10.1016/j.learninstruc.2015.05.007.

Ditton, H., & Arnoldt, B. (2004). Schülerbefragungen zum Fachunterricht [Student surveys on teaching]. *Empirische Pädagogik, 18*, 115–139.

Dyer, N., Hanges, P., & Hall, R. (2005). Applying multilevel confirmatory factor analysis techniques to the study of leadership. *The Leadership Quarterly, 16*, 149–167. https://doi.org/10.1016/j.leaqua.2004.09.009.

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

Feldman, K. (2007). Identifying exemplary teachers and teaching. In R. Perry, & J. Smart (Eds.). *The scholarship of teaching and learning in higher education* (pp. 93–143). Amsterdam, Netherlands: Springer.

Ferguson, R. (2012). Can student surveys measure teaching quality? *Phi Delta Kappan, 94*, 24–28. https://doi.org/10.2307/41763671.

Ferguson, R., & Danielson, C. (2014). 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). San Francisco, CA: Jossey-Bass.

Follman, J. (1992). Secondary school students' ratings of teacher effectiveness. *High School Journal, 75*, 168–178.

Follman, J. (1995). Elementary public-school pupil rating of teacher effectiveness. *Child Study Journal, 25*, 57–78.

Gaertner, H., & Brunner, M. (2018). Once good teaching, always good teaching? *Educational Assessment, Evaluation and Accountability, 30*, 159–182. https://doi.org/10.1007/s11092-018-9277.

Gates Foundation (2012). *Gathering feedback for teaching.* Retrieved from http://www.metproject.org/downloads/MET_Gathering_Feedback_Practioner_Brief.pdf.

Gates Foundation (2013). *Ensuring Fair and Reliable Measures of Effective Teaching. Culminating Findings from MET Project's Three Year Study.* http://www.metproject.org/downloads/MET_Ensuring_Fair_and_Reliable_Measures_Practitioner_Brief.pdf.

Gitomer, D. (2019). Evaluating instructional quality. *School Effectiveness and School Improvement, 30*, 68–78. https://doi.org/10.1080/09243453.2018.1539016.

Goe, L., Bell, C., & Little, O. (2008). *Approaches to evaluating teacher effectiveness.* Washington, DC: NCCTQ.

Göllner, R., Wagner, W., Klieme, E., Lüdtke, O., Nagengast, B., & Trautwein, U. (2016). Erfassung der Unterrichtsqualität mithilfe von Schülerurteilen [Capturing the quality of teaching through student ratings]. *Forschungsvorhaben in ankopplung an large-scale-assessments. Vol. 44. Forschungsvorhaben in ankopplung an large-scale-assessments* (pp. 63–82). Berlin, Germany: BMBF.

Hamre, B., Hatfield, B., Pianta, R., & Jamil, F. (2014). Evidence for general and domain-specific elements of teacher-child interactions. *Child Development, 85*, 1257–1274. https://doi.org/10.1111/cdev.12184.

Hassel, S., & Ridout, N. (2018). An investigation of first-year students' and lecturers' expectations of university education. *Frontiers in Psychology, 8*, 1–13. https://doi.org/10.3389/fpsyg.2017.02218.

Hattie, J., & Timperley, H. (2016). The power of feedback. *Review of Educational Research, 77*, 81–112. https://doi.org/10.3102/003465430298487.

Hedges, L., Hedberg, E., & Kuyper, A. (2012). The variance of intraclass correlations in three-and four-level models. *Educational and Psychological Measurement, 72*, 893–909. https://doi.org/10.1177/0013164412445193.

Helmke, A. (2012). *Unterrichtsqualität und Lehrerprofessionalität [Quality of teaching and professionalism of teachers* (4th ed.). Seelze, Germany: Klett-Kallmeyer.

Hu, L.-T., & Bentler, P. (1999). Cutoff criteria for fit indexes in covariance structure analysis. *Structural Equation Modeling, 6*, 1–55. https://doi.org/10.1080/10705519909540118.

Jackson, D., Teal, C., Raines, S., Nansel, T., Force, R., & Burdsal, C. (1999). The dimensions of students' perceptions of teaching effectiveness. *Educational and Psychological Measurement, 59*, 580–596. https://doi.org/10.1177/00131649921970035.

James, L., Demaree, R., & Wolf, G. (1984). Estimating within-group interrater reliability with and without response bias. *Journal of Applied Psychology, 69*, 85–98. https://doi.org/10.1037/0021-9010.69.1.85.

Kane, M. (2013). Validating the interpretations and uses of test scores. *Journal of Educational Measurement, 50*, 1–73. https://doi.org/10.1111/jedm.12000.

Kane, T., Kerr, K., & Pianta, R. (2014). *Designing teacher evaluation systems.* New York, NY: Wiley.

Klieme, E., Lipowsky, F., Rakoczy, K., & Ratzka, N. (2006). Qualitätsdimensionen und Wirksamkeit von Mathematikunterricht [Quality dimensions and effectiveness of mathematics teaching]. In M. Prenzel, & L. Allolio-Näcke (Eds.). *Untersuchungen zur Bildungsqualität von Schule* (pp. 127–146). Münster, Germany: Waxmann.

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

Kuger, S. (2016). Curriculum and learning time in international school achievement studies. In S. Kuger, E. Klieme, N. Jude, & D. Kaplan (Eds.). *Assessing contexts of learning* (pp. 395–422). Berlin, Germany: Springer.

Kuhfeld, M. (2016). *Multilevel item factor analysis and student perceptions of teacher effectiveness.* UCLA. Retrieved from https://escholarship.org/uc/item/076175k5.

Künsting, J., Neuber, V., & Lipowsky, F. (2016). Teacher self-efficacy as a long-term predictor of instructional quality in the classroom. *European Journal of Psychology of Education, 31*, 299–322. https://doi.org/10.1007/s10212-015-0272-7.

Kunter, M., & Baumert, J. (2007). Who is the expert? *Learning Environments Research, 9*, 231–251. https://doi.org/10.1007/s10984-006-9015-7.

Kunter, M., & Voss, T. (2013). The model of instructional quality in COACTIV. 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). New York, NY: Springer. https://doi.org/10.1007/978-1-4614-5149-5_6.

Kyriakides, L., Creemers, B., Panayiotou, A., Vanlaar, G., Pfeifer, M., Cankar, G., et al. (2014). Using student ratings to measure quality of teaching in six European countries. *European Journal of Teacher Education, 37*, 125–143. https://doi.org/10.1080/02619768.2014.882311.

Lipowsky, F., Rakoczy, K., Pauli, C., Drollinger-Vetter, B., Klieme, E., & Reusser, K. (2009). Quality of geometry instruction and its short-term impact on students' understanding of the Pythagorean Theorem. *Learning and Instruction, 19*, 527–537. https://doi.org/10.1037/t15601-000.

Lüdtke, O., Marsh, H., Robitzsch, A., & Trautwein, U. (2011). A 2x2 taxonomy of multilevel latent contextual models. *Psychological Methods, 16*, 444–467. https://doi.org/10.1037/a0024376.

Lüdtke, O., Trautwein, U., Kunter, M., & Baumert, J. (2006). Reliability and agreement of student ratings of the classroom environment. *Learning Environments Research, 9*, 215–230. https://doi.org/10.1007/s10984-006-9014-8.

Lüdtke, O., Trautwein, U., Schnyder, I., & Niggli, A. (2007). Simultane Analysen auf Schüler- und Klassenebene [Simultaneous analyses at student and classroom level]. *Zeitschrift für Entwicklungspsychologie und Pädagogische Psychologie, 39*, 1–11. https://doi.org/10.1026/0049-8637.39.1.1.

MacNell, L., Driscoll, A., & Hunt, A. (2015). What's in a name. *Innovative Higher Education, 40*, 291–303. https://doi.org/10.1007/s10755-014-9313-4.

Marsh, H. (1980). The influence of student, course, and instructor characteristics in evaluations of university teaching. *American Educational Research Journal, 17*, 219–237. https://doi.org/10.2307/1162484.

Marsh, H. (1983). Multidimensional ratings of teaching effectiveness by students from different academic settings and their relation to student/course/instructor characteristics. *Journal of Educational Psychology, 75*, 150–166. https://doi.org/10.1037//0022-0663.75.1.150.

Marsh, H. (2007). Students' evaluations of university teaching. In R. Perry, & J. Smart (Eds.). *The scholarship of teaching and learning in higher education* (pp. 319–383). Dordrecht, Netherlands: Springer.

Marsh, H., Dicke, T., & Pfeiffer, M. (2019). A tale of two quests. *Contemporary Educational Psychology, 58*, 1–18. https://doi.org/10.1016/j.cedpsych.2019.01.011.

Marsh, H., Lüdtke, O., Nagengast, B., Trautwein, U., Morin, A., Abduljabbar, A., et al. (2012). Classroom climate and contextual effects. *Educational Psychologist, 47*, 106–124. https://doi.org/10.1080/00461520.2012.670488.

Muthén, B. (1994). Multilevel covariance structure analysis. *Sociological Methods & Research, 22*, 376–398. https://doi.org/10.1177/0049124194022003006.

Muthén, B., & Muthén, L. (2018). *Mplus (version 8.1) [computer software]. Los Angeles, CA*.

Ory, J., & Ryan, K. (2001). How do student ratings measure up to a new validity framework? *New Directions for Institutional Research, 109*, 27–44. https://doi.org/10.1002/ir.2.

Patrick, H., Kaplan, A., & Ryan, A. (2011). Positive classroom motivational environments. *Journal of Educational Psychology, 103*, 367–382. https://doi.org/10.1037/a0023311.

Praetorius, A.-K., Klieme, E., Herbert, B., & Pinger, P. (2018). Generic dimensions of teaching quality. *ZDM, 50*, 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. *Learning and Instruction, 31*, 2–12. https://doi.org/10.1016/j.learninstruc.2013.12.002.

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

R Core Team (2017). *R: A language and environment for statistical computing.* Vienna, Austria: R Foundation. Retrieved from http://www.R-project.org [10.07.2018] .

Rakoczy, K. (2008). *Motivationsunterstützung im Mathematikunterricht [Motivational support in mathematics teaching].* Münster, Germany: Waxmann.

Richardson, J. (2005). Instruments for obtaining student feedback. *Assessment & Evaluation in Higher Education, 30*, 387–415.

Rubie-Davies, C. (2010). Teacher expectations and perceptions of student attributes. *British Journal of Educational Psychology, 80*, 121–135. https://doi.org/10.1348/000709909x466334.

Ryan, R., & Deci, E. (2000). Self-determination theory and the facilitation of intrinsic motivation, social development, and well-being. *American Psychologist, 55*, 68–78. https://doi.org/10.1037/0003-066x.55.1.68.

Scherer, R., Nilsen, T., & Jansen, M. (2016). Evaluating individual students' perceptions of instructional quality. *Frontiers in Psychology, 7*, 1–16. https://doi.org/10.3389/fpsyg.2016.00110.

Schiepe-Tiska, A. (2019). School tracks as differential learning environments moderate the relationship between teaching quality and multidimensional learning goals in mathematics. *Frontiers in Education, 4*. https://doi.org/10.3389/feduc.2019.00004.

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.

Schwab, S., & Helm, C. (2015). Überprüfung von Messinvarianz mittels CFA und DIF-Analysen [Verification of measurement invariance using CFA and DIF analyses]. *Empirische Sonderpädagogik, 3*, 175–193.

Schweig, J. (2014). Cross-level measurement invariance in school and classroom environment surveys. *Educational Evaluation and Policy Analysis, 36*, 259–280. https://doi.org/10.3102/0162373713509880.

Seidel, T., & Shavelson, R. (2007). Teaching effectiveness research in the past decade. *Review of Educational Research, 77*, 454–499. https://doi.org/10.3102/0034654307310317.

Shevlin, M., Banyard, P., Davies, M., & Griffiths, M. (2000). The validity of student evaluation of teaching in higher education. *Assessment & Evaluation in Higher Education, 25*, 397–405. https://doi.org/10.1080/713611436.

Shieh, G. (2016). Choosing the best index for the average score intraclass correlation coefficient. *Behavior Research Methods, 48*, 994–1003. https://doi.org/10.3758/s13428-015-0623-y.

Slavin, R. (1987). A theory of school and classroom organization. *Educational Psychologist, 22*, 89–108. https://doi.org/10.1207/s15326985ep2202_1.

Slavin, R. (1994). Quality, appropriateness, incentive, and time. *International Journal of Educational Research, 21*, 141–157. https://doi.org/10.1016/0883-0355(94)90029-9.

Taut, S., & Rakoczy, K. (2016). Observing instructional quality in the context of school evaluation. *Learning and Instruction, 46*, 45–60. https://doi.org/10.1016/j.learninstruc.2016.08.003.

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

Wagner, W., Göllner, R., Helmke, A., Trautwein, U., & Lüdtke, O. (2013). Construct validity of student perceptions of instructional quality is high, but not perfect. *Learning and Instruction, 28*, 1–11. https://doi.org/10.1016/j.learninstruc.2013.03.003.

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

Walker, J. (2009). Authoritative classroom management. *Theory and Practice, 48*, 122–129. https://doi.org/10.1080/00405840902776392.

Wallace, T., Kelcey, B., & Ruzek, E. (2016). What can student perception surveys tell us about teaching. *American Educational Research Journal, 53*, 1834–1868. https://doi.org/10.3102/0002831216671864.

Wisniewski, B., & Zierer, K. (2020). Entwicklung eines Online-Fragebogens zur Erhebung von Unterrichtsqualität durch Lernendenfeedback und erste Validierungsschritte [development of an online questionnaire to assess teaching quality through learner feedback and initial validation steps]. *Psychologie in Erziehung und Unterricht* In press.