Learning and Instruction 29 (2014) 1–9

Contents lists available at ScienceDirect

Learning and Instruction

journal homepage: www.elsevier.com/locate/learninstruc

Student ratings of teaching quality in primary school: Dimensions and prediction of student outcomes

Benjamin Fauth[a,b,*], Jasmin Decristan[a,c], Svenja Rieser[a,b], Eckhard Klieme[a,c], Gerhard Büttner[a,b]

[a] IDeA Research Center, Frankfurt/Main, Germany
[b] Department of Psychology, Goethe-University, Frankfurt/Main, Germany
[c] German Institute for International Educational Research (DIPF), Frankfurt/Main, Germany

ARTICLE INFO

Article history:
Received 6 February 2013
Received in revised form
3 July 2013
Accepted 9 July 2013

Keywords:
Teaching quality
Student ratings
Primary school
Validity
Teacher popularity

ABSTRACT

The contribution examines theoretical foundations, factorial structure, and predictive power of student ratings of teaching quality. Three basic dimensions of teaching quality have previously been described: classroom management, cognitive activation, and supportive climate. However, student ratings, especially those provided by primary school students, have been criticised for being biased by factors such as teacher popularity. The present study examines ratings of teaching quality and science learning among third graders. Results of multilevel confirmatory factor analyses ($N = 1556$ students, 89 classes) indicate that the three-dimensional model of teaching quality can be replicated in ratings of third graders. In a longitudinal study ($N = 1070$ students, 54 classes), we found ratings of classroom management to predict student achievement, and ratings of cognitive activation and supportive climate to predict students' development of subject-related interest after teacher popularity is controlled for. The analyses show that student ratings can be useful measures of teaching quality in primary school.

1. Theoretical framework

While student evaluations and student feedback are very common in higher education research and practice (Marsh, 2007), ratings of students in primary school are often neglected. It is an open question whether ratings of teaching quality by primary school students are reliable and valid measures (De Jong & Westerhof, 2001). Consistently, most of the previous studies of student ratings have considered only secondary school or college students. Furthermore, existing studies that do include younger students often lack methodologically sound designs. Nevertheless, we suggest that even in primary schools, student ratings can provide unique insight into classroom processes. The present research examines the theoretical foundations, factorial structure, and predictive power of student ratings.

In the following section, we introduce the multidimensional model of teaching quality upon which we based our study. Afterwards, we survey current research on student ratings to assess teaching quality and their connections with educational outcomes.

1.1. Teaching quality

Research on educational effectiveness has shown that classroom processes are an important source of variation in students' learning (Creemers & Kyriakides, 2008). Modern conceptualisations of teaching and learning address both cognitive and motivational learning processes. Additionally, domain-specific and domain-independent aspects of learning and instruction are taken into account (Seidel & Shavelson, 2007).

Klieme, Pauli, and Reusser (2009) present a theoretical framework for teaching quality that has been elaborated in the context of the 1995 TIMSS video study (Klieme, Schümer, & Knoll, 2001) and extended in the video intervention study "Quality of Instruction, Learning, and Mathematical Understanding" (Klieme et al., 2009). This model assumes that the three basic dimensions of teaching quality, namely, supportive climate, effective classroom management, and cognitive activation, are critical for student learning and motivation. These three basic dimensions are in accordance with other international theoretical models and empirical findings (Baumert et al., 2010; Pianta & Hamre, 2009).

Supportive climate covers specific aspects of the teacher–student relationship such as positive and constructive teacher feedback, a positive approach to student errors and misconceptions,

* Corresponding author. IDeA Research Center, Solmsstraße 73, D-60486 Frankfurt/Main, Germany. Tel.: +49 6924708814.
E-mail address: fauth@dipf.de (B. Fauth).

0959-4752/$ – see front matter © 2013 Elsevier Ltd. All rights reserved.
http://dx.doi.org/10.1016/j.learninstruc.2013.07.001

and caring teacher behaviour (Brophy, 2000; Klieme et al., 2009). The impact of positive student–teacher relationships on student motivation and learning has been confirmed empirically (Goodenow, 1992; Pianta, Nimetz, & Bennett, 1997). It has also been conceptualised by different theoretical approaches (Davis, 2003). We focus on a concept of supportive climate that is based on self-determination theory (Ryan & Deci, 2000). It assumes three basic intrinsic needs to be associated with human motivation: social relatedness, autonomy, and competence. Classrooms that are able to fulfil these needs should have positive effects on student outcomes, especially on students' intrinsic motivation and subject-related interest (Kunter, Baumert, & Köller, 2007).

Classroom management is a well-known concept in educational research (e.g., Kounin, 1970) that focusses on classroom rules and procedures, coping with disruptions, and smooth transitions. These classroom features can be seen as preconditions for time on task that is, in turn, crucial for students' learning gains (Seidel & Shavelson, 2007). Meta-analyses consistently show substantial effects of classroom management on student achievement (Seidel & Shavelson, 2007; Wang, Haertel, & Walberg, 1993).

Cognitive activation integrates challenging tasks, the exploration of concepts, ideas, and prior knowledge, and Socratic Dialogue practice as key features (Lipowsky et al., 2009). These classroom practices should foster students' cognitive engagement, which should, in turn, lead to elaborated knowledge (Klieme et al., 2009). Cognitive activation is closely connected to the subject matter. This concept has been predominantly developed in studies of mathematics classrooms (e.g., Baumert et al., 2010). However, research has shown that this concept can successfully be applied to other domains in primary school (Hamre, Pianta, Mashburn, & Downer, 2007).

1.2. *Student ratings of teaching quality*

In addition to video-based observations, teaching quality is frequently measured by student ratings. In student ratings, two sources of variance can be considered: the individual (idiosyncratic) students' perceptions and the (mutually shared) perceptions of the students in the class. The former is reflected by variance within classes (differences between students) and the latter by variance between classes (differences between learning environments; Lüdtke, Robitzsch, Trautwein, & Kunter, 2009). The choice of the level of analysis depends upon the research question addressed (Marsh et al., 2012).

Regarding the reliability and validity of student ratings, discriminant validity is one of the most important concerns about student ratings of instruction (Greenwald, 1997). According to Greenwald (1997), we can distinguish two types of discriminant validity in terms of ratings of instruction. The first is the multidimensionality of the ratings, which refers to the discrimination between components of the same construct (e.g., teaching quality). The second refers to the discrimination of teaching quality from other influences on ratings, such as teacher popularity.

1.2.1. *Dimensionality*

The question of dimensionality is closely related to the discussion of the Halo-effect as a well-known rater error. The "inadequate discrimination model" explains the Halo-effect as the insufficient capability of raters to discriminate between different aspects (Lance, La Pointe, & Stewart, 1994). Attempts to examine dimensionality have drawn on data from secondary schools or universities to perform multilevel confirmatory factor analyses (Dubberke, Kunter, McElvany, Brunner, & Baumert, 2008; Kunter et al., 2008; Marsh, 2007; Wagner, Göllner, Helmke, Trautwein, & Lüdtke, 2013). Taking classroom and individual levels of analyses into account, their findings showed that the factorial structure can differ between levels and that factor correlations between classes tend to be higher than within classes. Thus, the multilevel data structure of student ratings should also be considered statistically.

In primary schools, Doll, Spies, LeClair, Kurien, and Foley (2010) examined the factorial properties of their ClassMaps Survey with students from grades three to five. However, confirmatory analyses were not used, and multilevel data structure was not considered, which makes the results difficult to interpret (Marsh et al., 2012). The dimensionality of primary school students' ratings remains a largely unresolved issue. Attempts to examine the factorial structure of student ratings in primary school appear promising, but they must be extended by applying state of the art methodological approaches (Allen & Fraser, 2007; Doll et al., 2010; see Research Question 1).

1.2.2. *Teacher popularity*

Teacher popularity is generally believed to confound student ratings of teaching quality. Aleamoni (1999) summarises the concerns typically expressed by researchers: "Most student rating schemes are nothing more than a popularity contest with the warm, friendly, humorous instructor emerging as the winner every time" (p. 154). In the present paper, we regard teacher popularity as the affectively coloured general impression of the teacher. A simple operationalisation is the item "I like my teacher". Wagner (2008) found significant correlations between this item and measures of teaching quality (within and between classes) in secondary school. In his study, teacher popularity was also correlated with measures of achievement. It is reasonable that the affective relationship between the teacher and students (teacher popularity) is especially relevant in the earlier grades of primary school (e.g., Doll et al., 2010; La Rocque, 2008). However, teacher popularity must be distinguished theoretically from the concept of teaching quality. Therefore, researchers should determine whether teaching quality can predict student outcomes after the effect of teacher popularity is controlled for. This is one of the main points of the present study (see Research Questions 2 and 3).

1.3. *Connection of student ratings with learning outcomes and subject-related interest*

According to our theoretical framework, teaching quality should not only foster students' achievement but also affect motivational processes (De Jong & Westerhof, 2001; Rieser, Fauth, Decristan, Klieme, & Büttner, 2013). Aspects of intrinsic motivation in the classroom have convincingly been described within the construct of subject-related interest (Pintrich, 2003). Research on interest often defines the construct within the framework of self-determination theory (Deci & Ryan, 2000; Krapp, 2007). Kunter et al. (2007) stated that experiences of social relatedness, autonomy and competence are associated with higher degrees of intrinsic motivation, engagement and subject-related interest. The following section briefly summarises empirical effects of student ratings on learning outcomes and interest.

Ratings of cognitive activation and similar constructs (e.g., task difficulty, see Fraser & Fisher, 1982) have especially been found to predict student achievement (Dubberke et al., 2008, grades nine to ten; Fraser & Fisher, 1982, grade seven). These effects are more pronounced when considering classroom aggregated ratings and less pronounced for ratings of younger students (Haertel, Walberg, & Haertel, 1981). Cognitive activation also affects students' interest via feelings of competence within classes (Kunter, 2005) and between classes (Fraser & Fisher, 1982).

Supportive climate in particular has been found to predict students' motivation and interest (Reeve, 2002; Ryan & Deci, 2000). Ryan and Grolnick (1986) and Allen and Fraser (2007) confirmed effects of supportive climate on motivational variables in primary school (grades four to six), although only in single-level analyses. The authors found no connection between teacher support and achievement in science education.

Effects of student ratings of classroom management on learning outcomes were found primarily between classes in secondary school (Haertel et al., 1981; Kunter & Baumert, 2006). Kunter et al. (2007) show that certain aspects of perceived classroom management are also related to the development of subject-related interest (only within classes).

Research on the predictive power of student ratings usually draws upon the ratings of students aged 12 years or older (grades six or higher). Some studies consider students at the end of primary school (grades four or higher). However, our current knowledge of the predictive power of student ratings in the earlier primary school grades is limited (De Jong & Westerhof, 2001; Fraser, 1998).

2. Research questions and hypotheses

We investigate the application of the three basic dimensions of teaching quality to student ratings in primary school. The theoretical insights outlined above lead to the following research questions:

(1) Can teaching quality in grade three be assessed according to a three-dimensional structure of cognitive activation, supportive climate, and classroom management? We expect that a latent factor model with three dimensions at each level (within and between classes) will best fit the data (Hypothesis 1).
(2) What are the effects of student-rated teaching quality on the development of subject-related interest after teacher popularity is controlled for? We expect supportive climate and cognitive activation to predict the development of subject-related interest at both the individual and classroom level. We expect classroom management to have an effect at the individual level. We expect that the association between student ratings of teaching quality and subject-related interest will be significant over and above teacher popularity (Hypothesis 2).
(3) What are the effects of student-rated teaching quality on student achievement after teacher popularity is controlled for? We expect classroom management and cognitive activation to predict student achievement. However, we expect these effects to be significant only at the classroom level. We expect that these effects will be significant over and above teacher popularity (Hypothesis 3).

3. Methods

3.1. Sample

All analyses of Research Question 1 draw on data from a total of 1556 German third grade students in 89 classes. The average student age was 8.8 years (SD = .51). Analyses of Research Questions 2 and 3 (prediction of outcomes) draw on data from a subsample of 1070 third grade students in 54 classes. These students participated in an intervention study and could be assessed in the longitudinal design described in Section 3.2. The average student age in the subsample was 8.8 years (SD = .50). In both samples, 49% of the students were female. The target population of the study were public primary schools in a German state. Teachers and headmasters were contacted via telephone and invited to attend informational sessions. Participating schools were located in both urban (61% of classes) and rural areas. Participation in the study was voluntary for both teachers and students. The average participation rate for each classroom was 96%.

3.2. Design

The longitudinal study was part of a larger design evaluating different teaching approaches in science education in German primary schools. Here, teachers conducted two pre-designed teaching units on floating and sinking (each consisting of nine lessons of 45 min each). The teaching units were adapted from an empirically evaluated science curriculum for teaching floating and sinking. The curriculum is modelled on the principles of inquiry-based science education (Hardy, Jonen, Möller, & Stern, 2006). The first unit covered the concept of density; the second unit focused on the concepts of buoyancy force and displacement.

Students were assessed two times before the units started, once between the units, and once afterwards. Fig. 1 shows the measurement design. Ratings of teaching quality from measurement point A were used for multilevel factor analyses (Research Question 1), which required a large sample size on the classroom level. Research Questions 2 and 3 (prediction of outcomes) are addressed in the longitudinal design entailing all measurement points from A to D (Fig. 1). Ratings of teaching quality from measurement point C allowed insight specifically into teaching quality in the units on floating and sinking. Students' corresponding knowledge and interest concerning these units were assessed at measurement point D.

3.3. Measures

3.3.1. Teaching quality

We measured perceived teaching quality in science education with a questionnaire that included 21 items on three subscales: supportive climate (nine items), classroom management (five items), and cognitive activation (seven items; see appendix for wording). Cognitive activation comprised exploration of prior knowledge, exploration of students' way of thinking, and challenging tasks. Supportive climate was related to teachers' warmth and friendliness, encouragement, and constructive feedback. Classroom management

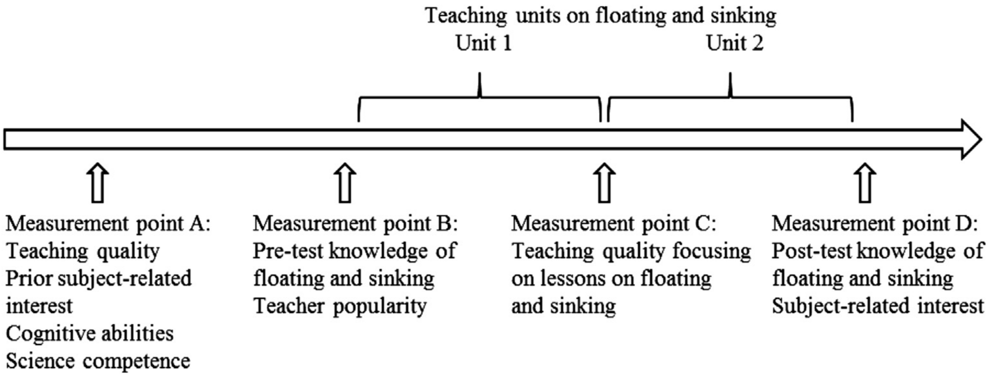
Teaching units on floating and sinking
Unit 1
Unit 2
Measurement point A:
Teaching quality
Prior subject-related interest
Cognitive abilities
Science competence
Measurement point B:
Pre-test knowledge of floating and sinking
Teacher popularity
Measurement point C:
Teaching quality focusing on lessons on floating and sinking
Measurement point D:
Post-test knowledge of floating and sinking
Subject-related interest

Fig. 1. Design of the study, points of measurement, and constructs assessed.

Table 1
Reliability of subscales of teaching quality focussing on science education and lessons on floating and sinking.

Dimension	Focus on science education (measures at point A)			Focus on specific unit (measures at point C)		
	Cronbach's α	ICC1	ICC2	Cronbach's α	ICC1	ICC2
Supportive climate	.73	.28	.87	.88	.16	.78
Classroom management	.82	.25	.85	.87	.25	.86
Cognitive activation	.82	.17	.77	.78	.13	.73

was operationalised through a lack of disciplinary problems and disruptions in the classroom. Items were adapted from Diel and Höhner (2008) and Rakoczy, Buff, and Lipowsky (2005) and reworded for application in primary school classrooms. Pre-versions of the items were tested in a pilot-study with 159 second and third graders (six classes). Items that were not understandable at grades two and three were revised for a more suitable wording. We avoided negative formulations, inverted items, and unfamiliar expressions. All of the survey items were rated on a four-point scale ranging from 1 = *strongly disagree* to 4 = *strongly agree*.

The longitudinal analyses of Research Questions 2 and 3 were carried out using similar items assessing teaching quality. The only difference was that these items focused on specific lessons within the science topic of floating and sinking (e.g., "During the lessons on floating and sinking nobody disrupted the lesson.").

Intra-class correlations (ICC1) of these scales indicated a substantial amount of variance between classes, ranging from 13% to 28% (Table 1). Regarding the multilevel data structure, an aggregation of students' ratings at the classroom level is only feasible when a sufficient interrater agreement within classes can be assured. We computed ICC2-indices as indicators for the accuracy of class-mean ratings (Lüdtke et al., 2009). ICC2 scores were sufficient (>.70), as was the internal consistency (Cronbach's α; Table 1).

3.3.2. Outcome measures and covariates

Teacher popularity was measured using a three-item-scale based on Wagner (2008; e.g., "I like my science teacher very much"; Cronbach's $\alpha = .92$, ICC1 = .15, ICC2 = .74). To measure students' prior interest in science education we used a four-item scale (e.g., "I put effort into science education class because it was fun"; Cronbach's $\alpha = .89$, ICC1 = .20) based on Blumberg (2008). Post-interest was measured with a similar scale that was formulated to focus on students' interest on the teaching unit (e.g., "I put effort into the topic of floating and sinking because it was fun"; Cronbach's $\alpha = .91$, ICC1 = .16).

We assessed students' prior scientific literacy using an adapted version of the TIMSS-test (Martin, Mullis, & Foy, 2008) that fitted the 1PL-Rasch Model (13 items, EAP/PV reliability = .70). Cognitive abilities were assessed using the CFT 20-R (56 items, Cronbach's $\alpha = .72$; Weiß, 2006), a German version of the Culture Fair Intelligence Tests. Students' knowledge of floating and sinking was assessed using standardised tests. The tests were adapted from Hardy et al. (2006). The pre-test comprised 16 items (EAP/PV reliability = .52), and the post-test comprised 13 items (EAP/PV reliability = .76). Items were scored dichotomously or politomously, and both tests were scaled separately using the Partial Credit Model each time. Student parameters were estimated using weighted likelihood estimates (Warm, 1989).

3.4. Procedure

Data were collected during classroom-wide assessments by trained staff using standardised instructions. Students were instructed how to handle the test items. The items were read aloud to the class to account for language and reading difficulties. After each item, students were given time to respond.

3.5. Data analyses

Multilevel analyses have been successfully applied to student ratings of classroom environments (Lüdtke et al., 2009). To examine the factorial structure of student ratings, we conducted multilevel confirmatory factor analyses (MCFA). These models were estimated as doubly-latent models according to the framework proposed by Marsh et al. (2009).

Concerning goodness of fit indices, common thump rules for cut-off criteria in single-level analyses (RMSEA near .06, SRMR near .08; Hu & Bentler, 1999) remain a controversial issue (Marsh, Hau, & Wen, 2004). Furthermore, there is only limited research on the interpretation of global fit indices in multilevel models. We compared different models using the Wald Chi-square test, the Akaike Information Criterion (AIC), and Bayes Information Criterion (BIC), preferring models with lower values (Raftery, 1993).

We conducted multilevel regression analyses to examine the predictive power of ratings of teaching quality on the development of subject-related interest and achievement. We used group mean centring (Lüdtke et al., 2009) for individual-level (level 1) variables that assessed learning environments (supportive climate, classroom management, cognitive activation, and teacher popularity). This means that the within-class effect only represents variance explained within classes. The between-class effect represents variance explained by class-mean scores. Covariates of student achievement (cognitive abilities, science knowledge and scientific literacy) and prior interest were introduced as grand-mean centred level 1 predictors representing variance within and between classes (Lüdtke et al., 2009). Note that all classroom-level beta-weights are standardised in relation to variance between classes of the dependent variables. Regression analyses were estimated as doubly-manifest models according to the framework proposed by Marsh et al. (2009), with single manifest indicators for the scales and manifest aggregation of individual ratings at the classroom level.

The issue of missing values requires careful consideration (Enders, 2010). In our study, the amount of missing data per scale was relatively small (average 8.2%, range 7.3–10.2%). There were no missing data at level 2 (class-aggregated ratings). Missing values were generated when students did not attend school on the day the measurements were taken. There was no indication of a systematic accumulation of missing data patterns across scales or measurement points. A full information maximum likelihood algorithm (FIML; Arbuckle, 1996) could be used to deal with missing data when performing the factor analyses. For multilevel regression analyses, cases with missing data on the manifest predictor variables were not included in the analyses. All analyses were conducted in MPlus 7 (Muthén & Muthén, 1998–2012) using robust maximum likelihood estimation (MLR; Yuan & Bentler, 2000).

4. Results

4.1. Descriptive results

The means and standard deviations of the teaching quality scale scores were comparable (Table 2). Note that items focussing on

Table 2
Descriptive results of subscales on teaching quality focussing on science education ($N = 1556$) and on specific teaching unit ($N = 1070$).

Dimension	Focus on science education (measures at point A)		Focus on specific unit (measures at point C)	
	M	SD	*M*	SD
Supportive climate	3.50	.50	3.30	.67
Classroom management	2.74	.73	2.56	.85
Cognitive activation	3.27	.55	3.19	.66

Note. Items were rated on a four-point scale ranging from 1 = strongly disagree to 4 = strongly agree.

Table 3
Fit indices of multilevel confirmatory factor analyses.

Index	3/3-Factor model	1/1-Factor model	3/1-Factor model
χ^2 (df)	869.78 (370)	2795.63 (416)	1061.41 (373)
$p(\chi^2)$	<.001	<.001	<.001
CFI	.92	.63	.90
TLI	.91	.59	.88
RMSEA	.03	.06	.03
SRMR (within)	.04	.09	.04
SRMR (between)	.11	.20	.21
AIC	73,047	74,992	73,202
BIC	73,651	75,564	73,791
Comparison with the 3/3-factor model	–	$\chi^2 = 20{,}417$ (df = 6) $p < .001$	$\chi^2 = 6922$ (df = 3) $p < .001$

science education were used to examine the factorial structure (Section 4.2) and that items focussing on the specific teaching unit were used to predict student outcomes (Section 4.3).

4.2. *Factorial structure (Research Question 1)*

We specified a three-dimensional model of teaching quality with the indicator variables of each factor outlined above. Items were assumed to load on the same factor at both levels (3/3-factor model). The global fit of this model was acceptable (Table 3).[1] Factor loadings and inter-factor correlations between the latent variables of the 3/3-factor model were all significant at both the within-class and between-class levels (Fig. 2).

To investigate Research Question 1, we compared this model to a one-dimensional model with all items loaded on one factor at the individual level and one factor at the classroom level (1/1-factor model representing a global impression of teaching quality). The global fit of this model was not acceptable (Table 3).

Multilevel models often have a simpler factorial structure at the class level compared with the individual level (Wagner, 2008). Therefore, we conducted comparisons with one additional model. Three factors were loaded on this model at level 1 and a single factor at level 2 (3/1-factor model). This model met the cut-off criteria for some indices (RMSEA, $\text{SRMR}_{\text{within}}$) but not for others ($\text{SRMR}_{\text{between}}$, which specifically refers to a misfit at level 2; see Table 3). The 1/1-factor model and the 3/1-factor model fit the data significantly worse than the 3/3-factor model when assessed by comparative fit indices and the Wald Chi-square test (Table 3). All models were estimated with one free estimated correlation between residuals of two items of cognitive activation (Fig. 2). These residuals were expected to be correlated from a theoretical perspective due to the common wording stem of these two items (see Appendix).

4.3. *Ratings of teaching quality predicting student outcomes (Research Questions 2 and 3)*

We first computed the correlations between ratings of teaching quality in the teaching units and teacher popularity to explore the relationships between the two constructs. Correlations were substantial within and between classes and tended to be higher between classes (Table 4). Teacher popularity was controlled for in all subsequent regression models.

The following results answer Research Question 2: Do student ratings of teaching quality predict the development of subject-related interest when teacher popularity is controlled for?

We first tested the covariates of prior interest and teacher popularity in a model that predicts subject-related interest (Table 5, Model 1). In the next models (Models 2–4), the basic dimensions of teaching quality were included in a stepwise fashion to examine the predictive power of each dimension (Lipowsky et al., 2009). Cognitive activation and supportive climate were significant predictors at both levels, even when teacher popularity was controlled for (Models 2 and 3). A significant effect of classroom management was observed within classes but not between classes (Model 4).

In Model 5 (Table 5), each predictor is evaluated in terms of its unique contribution, and shared variance components are not assigned to any single basic dimension. Therefore, we examined the unique contribution of each basic dimension to the prediction of the outcomes (Tabachnick & Fidell, 2007). In this model, cognitive activation had no significant predictive power over and above supportive climate at the individual level. In contrast, at the classroom level, supportive climate had no significant predictive power over and above cognitive activation (Table 5). Remarkably, teacher popularity was a significant predictor of the development of subject-related interest in every model at both levels (Models 1–5). This finding indicates that teacher popularity has a predictive power beyond what can be explained by teaching quality.

The following section investigates Research Question 3: Are student ratings of teaching quality predictive of student achievement when teacher popularity is controlled for? In these regression models (Table 6), we controlled for pre-test scores, prior science competence, and general cognitive abilities in addition to teacher popularity because we assumed these variables to also be relevant to post-test scores (Model 1, Table 6). The dimensions of teaching quality were added in a stepwise fashion in the following regression models (Models 2–4). Only classroom management was a significant predictor of post-test scores (Model 4). As expected, this effect was observed only at the classroom level. If the specific contribution of each predictor is examined (controlling for each of the other basic dimensions), classroom management remains significant at the classroom level (Model 5). Teacher popularity was a significant predictor at the classroom level in the Models 1–3 and in Model 5 (Table 6).

[1] The student reports of teacher popularity could also be separated from the three dimensions of teaching quality. Like the three-factor model, the model with four factors had an adequate fit to the data ($\chi^2(490) = 1086.54$, $p < .05$, CFI = .94, TLI = .93, RMSEA = .03, $\text{SRMR}_{\text{within}} = .04$, $\text{SRMR}_{\text{between}} = .11$).

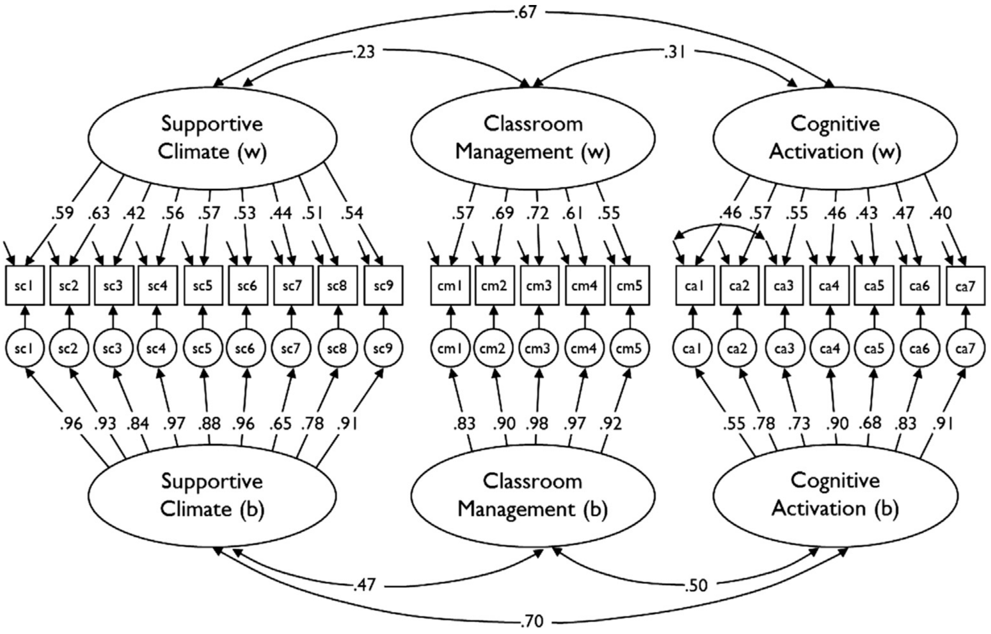

Fig. 2. 3/3-Factor model of student ratings of teaching quality.

5. Discussion

This article presents assessments of the factorial validity and predictive power of primary school students' ratings of teaching quality. We methodologically investigated the dimensional structure of these ratings and their partial independence from teacher popularity as a potential confounder. Substantively, we were able to confirm classroom management to be predictive of student learning, whereas supportive climate and cognitive activation are predictive of students' interest. Theoretically, our results provide evidence for the usefulness of the three-dimensional framework of teaching quality in primary school. In the next sections we discuss our results in detail, followed by a discussion of the general strengths, limitations, and educational implications of the study.

5.1. *Factorial structure*

Hypothesis 1 addressed the factorial structure of primary school students' ratings of teaching quality: supportive climate, classroom management, and cognitive activation. These basic dimensions were previously identified in studies of secondary school students' ratings (Kunter et al., 2008) and video-based observations (Klieme et al., 2009; Pianta & Hamre, 2009). Our study extends this knowledge with evidence from primary schools. We succeeded in identifying the three-dimensional structure along the basic dimensions in our younger student sample. Even third graders are able to distinguish certain aspects of teaching quality in survey assessments. Using multilevel confirmatory factor analyses, we confirmed that the distinction between the basic dimensions is not only valid at the students' individual perception (level 1); instead, whole classes also differ in their judgements regarding the basic dimensions (level 2). To our knowledge, these techniques have not yet been applied to primary school students' ratings of instruction. Taking into account the two-level structure is of special importance because teaching quality is conceptually a classroom-level construct that is individually perceived by students.

While the RMSEA and $\text{SRMR}_{\text{within}}$ of the 3/3-factor model meet Hu and Bentler's (1999) criteria, the model fit worse at level 2 as indicated by the $\text{SRMR}_{\text{between}}$. However, general cut-off criteria have been criticised to be inadequate for the investigation of multifactor rating instruments with more than two or three items loading onto one factor. Therefore, decisions should also be based on model comparisons rather than only on a priori cut-off values (Marsh et al., 2004). Considering the interpretability of parameter estimates and the fact that this model fit significantly better than the global factor models, we accepted the 3/3-factor model.

Table 4
Correlations between dimensions of teaching quality and teacher popularity.

Dimension	Teacher popularity	
	Within classes	Between classes
Supportive climate	.45*	.62*
Classroom management	.21*	.42*
Cognitive activation	.36*	.40*

Note. *$p < .05$; one-tailed test.

Table 5
Multilevel regression analyses. Dependent variable: students' subject-related interest. Dimensions of teaching quality focussing on the lessons on floating and sinking.

Predictor	Model 1	Model 2	Model 3	Model 4	Model 5
Individual level					
Prior interest	.16 (.05)*	.14 (.05)*	.14 (.05)*	.15 (.05)*	.12 (.05)*
Teacher popularity	.18 (.03)*	.12 (.04)*	.12 (.04)*	.19 (.04)*	.11 (.04)*
Cognitive activation	–	.18 (.04)*	–	–	.08 (.06)
Supportive climate	–	–	.19 (.05)*	–	.14 (.07)*
Classroom management	–	–	–	.11 (.04)*	.04 (.04)
Classroom level					
Teacher popularity	.58 (.12)*	.43 (.13)*	.36 (.15)*	.54 (.13)*	.40 (.15)*
Cognitive activation	–	.41 (.12)*	–	–	.38 (.23)*
Supportive climate	–	–	.40 (.13)*	–	.02 (.26)
Classroom Management	–	–	–	.15 (.16)	.10 (.16)
R^2 (within)	.07	.10	.10	.09	.11
R^2 (between)	.33	.49	.47	.38	.52

Note. Standardised regression weights; standard errors are in parentheses.
*$p < .05$; one-tailed test.

Table 6
Multilevel regression analyses. Dependent variable: students' post-test scores knowledge on floating and sinking. Dimensions of teaching quality focussing on the lessons on floating and sinking.

Predictor	Model 1	Model 2	Model 3	Model 4	Model 5
Individual level					
Pre-test	.21 (.03)*	.21 (.03)*	.21 (.03)*	.21 (.03)*	.21 (.03)*
Scientific literacy	.27 (.03)*	.28 (.03)*	.28 (.03)*	.29 (.03)*	.28 (.03)*
Cognitive abilities	.21 (.03)*	.23 (.03)*	.23 (.03)*	.22 (.03)*	.22 (.03)*
Teacher popularity	.03 (.03)	.03 (.03)	.02 (.03)	.03 (.03)	.03 (.03)
Cognitive activation	–	−.02 (.03)	–	–	−.05 (.04)
Supportive climate	–	–	.02 (.03)	–	.06 (.04)
Classroom management	–	–	–	−.01 (.03)	−.02 (.04)
Classroom level					
Teacher popularity	.31 (.11)*	.39 (.10)*	.43 (.14)*	.17 (.13)	.30 (.15)*
Cognitive activation	–	−.15 (.14)	–	–	−.11 (.23)
Supportive climate	–	–	−.16 (.18)	–	−.15 (.30)
Classroom management	–	–	–	.37 (.13)*	.41 (.15)*
R^2 (within)	.27	.28	.27	.28	.27
R^2 (between)	.10	.13	.12	.22	.27

Note. Standardised regression weights; standard errors are in parentheses.
*$p < .05$; one-tailed test.

Level 2 factor correlations of this model were moderate compared with those reported in Wagner et al. (2013) and similar to those of Kunter et al. (2008), who applied the three-dimensional framework to student ratings in secondary school. Thus there might be some halo bias in the ratings, but this does not overwhelm the dimensional structure. Taken together, our findings provide evidence for the discriminant validity of primary school students' ratings of teaching quality (Greenwald, 1997).

5.2. Teacher popularity, subject-related interest, and achievement

Our study confirmed the empirical relationship between student ratings of teaching quality and teacher popularity among primary school students. Referring to classroom level, teachers with high teaching quality might be more popular as a result. At the individual level, we can assume that a student who feels more affiliated with his or her teacher will tend to judge the teacher's teaching quality in a more positive manner.

The covariate regression models confirmed that students were more interested in the subject and learned more in classes that were instructed by popular teachers (level 2). Teacher popularity was significantly related to both outcomes at the classroom level, even when teaching quality was considered in the analyses. Additionally, students who liked their teacher more (compared with their classmates) were more likely to become interested in a certain teaching unit (level 1) which might be explained through more positive student–teacher relationships (Pianta et al., 1997). Therefore, teacher popularity should be considered when student ratings of teaching quality are examined.

As proposed in Hypothesis 2, a student who feels more supported and more cognitively activated tends to be more interested in the teaching units (level 1). This also holds true for class-aggregated ratings, indicating that students tend to be more interested if they are instructed by teachers who are generally regarded as more supportive and cognitively activating (level 2). Supportive climate is theoretically the most obvious dimension to support student motivation and interest (Klieme et al., 2009; Ryan & Deci, 2000). The empirical results of Ryan and Grolnick (1986) support this assumption in cross-sectional, single-level analyses in grades 4–6. Our findings confirm and extend the results of earlier studies: We found a connection between supportive climate and students' interest in longitudinal multilevel analyses in grade three. Traditionally, research on cognitive activation has focused on students' learning and conceptual understanding (Baumert et al., 2010; Lipowsky et al., 2009). It is assumed theoretically that cognitively activating classroom activities lead to more cognitive engagement of students, which should in turn lead to a deeper understanding of learning content (Lipowsky et al., 2009). However, cognitive activation was also found to predict student interest and motivation (Kunter, 2005). This link is seen as a second consequence of high cognitive engagement and partly mediated by feelings of competence and autonomy. In our study, we confirmed the effect of cognitive activation on the development of interest within and between classes (Hypothesis 2). However, we could not confirm its effect on achievement between classes (Hypothesis 3). Previous studies that measured cognitive activation predominantly drew upon data of students from grade five or higher (e.g., Fraser & Fisher, 1982). Possibly, the link between cognitive engagement and learning is not that pronounced in the early grades of primary school (third graders in our study) whereas the other link (between engagement and interest) is already established. Another explanation could be our choice of outcome measures. Our tests focused on knowledge and conceptual understanding of a specific teaching unit (short-term effects), whereas other studies have focused on one-year effects on broader constructs, such as mathematics literacy (Baumert et al., 2010; Dubberke et al., 2008) or the understanding of the nature of science (Fraser & Fisher, 1982). Apparently, further research on the construct of cognitive activation in different contexts is needed to prove these assumptions.

The regression model that considers the three basic dimensions of teaching quality simultaneously revealed that to a large extend the shared variance of cognitive activation and supportive climate was predictive for the development of students' interest rather than the specific predictive value of each dimension. Substantively, we suggest that the impact of teaching quality on student interest can mostly be attributed to features that cognitive activation and supportive climate have in common.

As expected, an effect of classroom management on the development of interest could only be confirmed at the individual level (Hypothesis 2). Kunter et al. (2007) assumed that it is the specific individual experience of teachers' classroom management strategies (level 1) that can be connected to students' feelings of competence and autonomy, which in turn supports students' subject-related interest. These assumptions underline individual effects and the lack of a classroom-level effect. In contrast, the effect of classroom management on student learning (Hypothesis 3) was limited to the classroom level, which is also in line with our expectations. More organised and less disrupted lessons provide more time on task (De Jong & Westerhof, 2001). Thus, students have better opportunities to engage with learning content. This mechanism is relevant on the classroom level: student achievement is promoted when the entire class spends more time on a task and is not dependent on the individual perception of classroom management. From a methodological point of view, these findings further emphasise the need to distinguish between different levels (within and between classes) when student surveys are analysed (Marsh et al., 2012).

5.3. Further limitations, strengths, and educational implications

In the longitudinal part of this study, we used student ratings that focused on specific teaching units to predict corresponding outcome measures: students' knowledge and interest concerning these units. Additionally, our lessons were standardised with regard to sequence and materials. This makes it easier to attribute effects to teaching quality rather than other influences which is an important strength of our study. However, this strength must be considered along with two further limitations of the study. First, we could only determine short-term effects. Further research should confirm our findings over longer-term studies. Second, the complex design of the longitudinal part of the study brought with it constraints regarding sample size. A larger sample would have been necessary to apply structural equation modelling to the longitudinal data, which is also desirable for future research.

The educational implications of our findings are relevant to research and have bearing on practical and policy issues. In educational research, most studies concerned with learning and instruction seek for valid indicators of teaching quality. This study provides evidence that the individual value of student ratings depends on the expected outcome, the aspect of teaching measured, and the level of analysis. For example, predictive validity regarding student achievement can only be assumed for aggregated ratings of classroom management. In educational practice, there is growing interest among teachers in using surveys to collect feedback from their students. Teachers are increasingly using this evidence-based method to promote their professional development. Many self-developed questionnaires are available for this purpose, but their psychometric properties often remain unproven or unreported. Finally, a discussion of the usefulness of student ratings is highly relevant in the current educational policy context because student data are increasingly used to monitor teaching practices (Gates Foundation, 2012). Based on our results, basing judgements or high-stakes decisions about individual teachers on the ratings of third grade students is not warranted. However, student ratings can be valuable to describe the teaching quality of groups of teachers.

Taken together, our study demonstrates the usefulness of student ratings, even in primary school. Our three-dimensional approach provides evidence for the impact of specific dimensions of teaching quality on specific educational outcomes.

Acknowledgement

We would like to thank three anonymous reviewers for their very helpful and constructive feedback. Earlier versions of this article benefited much from discussions with José-Felipe Martinez, Jan Hochweber, and Tanja Könen. This research was funded by the Hessian initiative for the development of scientific and economic excellence (LOEWE).

Appendix

Items on teaching quality

Classroom management

In our science class…
none of the students disturb the lesson
students are quiet when the teacher speaks
everybody listens and students are quiet
nobody interrupts with talking
everybody follows the teacher

Cognitive activation

In our science class…
we are working on tasks that I have to think about very thoroughly
Our science teacher…
asks me what I have understood and what I haven't
asks questions that I have to think about very thoroughly
gives us tasks that seem to be difficult at a first glance
asks what we know about a new topic
gives us tasks I like to think about
wants me to be able to explain my answers

Supportive climate

Our science teacher…
is nice to me even when I make a mistake
cares about me
encourages me when I find a task difficult
tells me how to do better when I make a mistake
likes me
tells me what I'm already good at and what I still have to learn
is friendly to me
compliments me when I did something good
believes that I can solve difficult tasks

References

Aleamoni, L. M. (1999). Student rating myths versus research facts from 1924 to 1998. *Journal of Personnel Evaluation in Education, 13*, 153–166. http://dx.doi.org/10.1023/A:1008168421283.

Allen, D., & Fraser, B. J. (2007). Parent and student perceptions of classroom learning environment and its association with student outcomes. *Learning Environments Research, 10*, 67–82. http://dx.doi.org/10.1007/s10984-007-9018-z.

Arbuckle, J. L. (1996). Full information estimation in the presence of incomplete data. In G. A. Marcoulides, & R. E. Schumacker (Eds.), *Advanced structural equation modeling* (pp. 243–277). Mahwah, NJ: Lawrence Erlbaum.

Baumert, J., Kunter, M., Blum, W., Brunner, M., Voss, T., Jordan, A., et al. (2010). Teachers' mathematical knowledge, cognitive activation in the classroom, and student progress. *American Educational Research Journal, 47*(1), 133–180. http://dx.doi.org/10.3102/0002831209345157.

Blumberg, E. (2008). *Multikriteriale Zielerreichung im naturwissenschaftsbezogenen Sachunterricht der Grundschule* [*Multi-criterial goal attainment in science education in primary school*] (Doctoral dissertation). Germany: University of Münster.

Brophy, J. (2000). *Teaching*. Brussels, Belgium: International Academy of Education.

Creemers, B., & Kyriakides, L. (2008). *The dynamics of educational effectiveness*. London: Routledge.

Davis, H. A. (2003). Conceptualizing the role and influence of student–teacher relationships on children's social and cognitive development. *Educational Psychologist, 38*(4), 207–234. http://dx.doi.org/10.1207/S15326985EP3804_2.

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

Deci, E. L., & Ryan, R. M. (2000). The "what" and "why" of goal pursuits: human needs and the self-determination of behavior. *Psychological Inquiry, 11*(4), 227–268. http://dx.doi.org/10.1207/S15327965PLI1104_01.

Diel, E., & Höhner, W. (2008). *Fragebögen zur Unterrichtqualität* [*Questionnaires on teaching quality*]. Wiesbaden, Germany: Institut für Qualitätsentwicklung.

Doll, B., Spies, R. A., LeClair, C. M., Kurien, S. A., & Foley, B. P. (2010). Student perceptions of classroom learning environments: development of the ClassMaps survey. *School Psychology Review, 39*(2), 203–218.

Dubberke, T., Kunter, M., McElvany, N., Brunner, M., & Baumert, J. (2008). Lerntheoretische Überzeugungen von Mathematiklehrkräften [Mathematics teachers' beliefs about learning]. *Zeitschrift für Pädagogische Psychologie, 22*(3), 193–206. http://dx.doi.org/10.1024/1010-0652.22.34.193.

Enders, C. K. (2010). *Applied missing data analysis*. New York: Guilford.

Fraser, B. J. (1998). Classroom environment instruments: development, validity and applications. *Learning Environments Research, 1*, 7–33. http://dx.doi.org/10.1023/A:1009932514731.

Fraser, B. J., & Fisher, D. L. (1982). Predicting students' outcomes from their perceptions of classroom psychosocial environment. *American Educational Research Journal, 19*, 498–518. http://dx.doi.org/10.3102/00028312019004498.

Gates Foundation (Ed.). (2012). *Gathering feedback for teaching*. Seattle: Gates Foundation.

Goodenow, C. (1992). Strengthening the links between educational psychology and the study of social contexts. *Educational Psychologist, 27*, 177–196. http://dx.doi.org/10.1207/s15326985ep2702_4.

Greenwald, A. G. (1997). Validity concerns and usefulness of student ratings of instruction. *American Psychologist, 52*(11), 1182–1186. http://dx.doi.org/10.1037/0003-066X.52.11.1182.

Haertel, G. D., Walberg, H. J., & Haertel, E. H. (1981). Socio-psychological environments and learning: a quantitative synthesis. *British Educational Research Journal, 7*(1), 27–36. http://dx.doi.org/10.1080/0141192810070103.

Hamre, B. K., Pianta, R. C., Mashburn, A. J., & Downer, J. T. (2007). *Building a science of classrooms: Application of the CLASS framework in over 4,000 U.S. early childhood and elementary classrooms*. New York: Foundation for Child Development.

Hardy, I., Jonen, A., Möller, K., & Stern, E. (2006). Effects of instructional support within constructivist learning environments for elementary school students' understanding of "floating and sinking". *Journal of Educational Psychology, 98*, 307–326. http://dx.doi.org/10.1037/0022-0663.98.2.307.

Hu, L., & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: conventional criteria versus new alternatives. *Structural Equation Modelling, 6*(1), 1–55. http://dx.doi.org/10.1080/10705519909540118.

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

Klieme, E., Schümer, G., & Knoll, S. (2001). Mathematikunterricht in der Sekundarstufe I: Aufgabenkultur und Unterrichtsgestaltung [Mathematics instruction at secondary level. Task culture and instructional design]. In Bundesministerium für Bildung und Forschung (BMBF) (Ed.), *TIMSS – Impulse für Schule und Unterricht* (pp. 43–57). München, Germany: Mediahaus Biering.

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

Krapp, A. (2007). An educational–psychological conceptualisation of interest. *International Journal for Educational and Vocational Guidance, 7*, 5–21. http://dx.doi.org/10.1007/s10775-007-9113-9.

Kunter, M. (2005). *Multiple Ziele im Mathematikunterricht* [*Multiple goals in mathematics instruction*]. Münster, Germany: Waxmann.

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. http://dx.doi.org/10.1007/s10984-006-9015-7.

Kunter, M., Baumert, J., & Köller, O. (2007). Effective classroom management and the development of subject-related interest. *Learning and Instruction, 17*(5), 494–509. http://dx.doi.org/10.1016/j.learninstruc.2007.09.002.

Kunter, M., Tsai, Y.-M., Klusmann, U., Brunner, M., Krauss, S., & Baumert, J. (2008). Students' and mathematics teachers' perceptions of teacher enthusiasm and instruction: motivation for teaching. *Learning and Instruction, 18*(5), 468–482. http://dx.doi.org/10.1016/j.learninstruc.2008.06.008.

La Rocque, M. (2008). Assessing perceptions of the environment in elementary classrooms: the link with achievement. *Educational Psychology in Practice, 24*(4), 289–305. http://dx.doi.org/10.1080/02667360802488732.

Lance, C. L., LaPointe, J. A., & Stewart, A. M. (1994). A test of the context dependency of three causal models of halo rater error. *Journal of Applied Psychology, 79*(3), 332–340. http://dx.doi.org/10.1037/0021-9010.79.3.332.

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

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. http://dx.doi.org/10.1016/j.cedpsych.2008.12.001.

Marsh, H. W. (2007). Students' evaluations of university teaching: a multidimensional perspective. In R. P. Perry, & J. C. Smart (Eds.), *The scholarship of teaching and learning in higher education* (pp. 319–384). New York: Springer.

Marsh, H. W., Hau, K. T., & Wen, Z. (2004). In search of golden rules: comment on hypothesis-testing approaches to setting cutoff values for fit indexes and dangers in overgeneralizing Hu & Bentler's findings. *Structural Equation Modeling, 11*, 320–341. http://dx.doi.org/10.1207/s15328007sem1103_2.

Marsh, H. W., Lüdtke, O., Nagengast, B., Trautwein, U., Morin, A. J. S., Abduljabbar, A. S., et al. (2012). Classroom climate and contextual effects: conceptual and methodological issues in the evaluation of group-level effects. *Educational Psychologist, 47*, 106–124. http://dx.doi.org/10.1080/00461520.2012.670488.

Marsh, H. W., Lüdtke, O., Robitzsch, A., Trautwein, U., Asparouhov, T., Muthén, B., et al. (2009). Doubly-latent models of school contextual effects: integrating multilevel and structural equation approaches to control measurement and sampling error. *Multivariate Behavioral Research, 44*, 764–802. http://dx.doi.org/10.1080/00273170903333665.

Martin, M. O., Mullis, I. V. S., & Foy, P. (2008). *TIMSS 2007 international science report*. Chestnut Hill, MA: TIMSS & PIRLS International Study Center.

Muthén, L. K., & Muthén, B. O. (1998–2012). *Mplus user's guide* (7th ed.). Los Angeles, CA: Muthén & Muthén.

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. http://dx.doi.org/10.3102/0013189X09332374.

Pianta, R. C., Nimetz, S. L., & Bennett, E. (1997). Mother–child relationships, teacher–child relationships, and school outcomes in preschool and kindergarten. *Early Childhood Research Quarterly, 12*, 263–280. http://dx.doi.org/10.1016/S0885-2006(97)90003-X.

Pintrich, P. R. (2003). A motivational science perspective on the role of student motivation in learning and teaching contexts. *Journal of Educational Psychology, 95*(4), 667–686. http://dx.doi.org/10.1037/0022-0663.95.4.667.

Raftery, A. E. (1993). Bayesian model selection in structural equation models. In K. A. Bollen, & J. S. Long (Eds.), *Testing structural equation models* (pp. 163–180). Newbury Park: Sage.

Rakoczy, K., Buff, A., & Lipowsky, F. (2005). *Dokumentation der Erhebungs- und Auswertungsinstrumente zur schweizerisch-deutschen Videostudie. Befragungsinstrumente* [Technical report of the German–Swiss video study]. Frankfurt, Germany: GFPF.

Reeve, J. (2002). Self-determination theory applied to educational settings. In E. L. Deci, & R. M. Ryan (Eds.), *Handbook of self-determination research* (pp. 183–203). Rochester: University of Rochester Press.

Rieser, S., Fauth, B., Decristan, J., Klieme, E., & Büttner, G. (2013). The connection between primary school students' self-regulation in learning and perceived teaching quality. *Journal of Cognitive Education and Psychology, 12*(2), 138–156. http://dx.doi.org/10.1891/1945-8959.12.2.1.

Ryan, R. M., & Deci, E. L. (2000). Self-determination theory and the facilitation of intrinsic motivation, social development, and well-being. *American Psychologist, 55*(1), 68–78. http://dx.doi.org/10.1037/0003-066X.55.1.68.

Ryan, R. M., & Grolnick, W. S. (1986). Origins and pawns in the classroom: a self-report and projective assessment of children's perceptions. *Journal of Personality and Social Psychology, 50*, 550–558. http://dx.doi.org/10.1037/0022-3514.50.3.550.

Seidel, T., & Shavelson, R. (2007). Teaching effectiveness research in the past decade. *Review of Educational Research, 77*(4), 454–499. http://dx.doi.org/10.3102/0034654307310317.

Tabachnick, B. G., & Fidell, L. S. (2007). *Using multivariate statistics*. Boston: Pearson.

Wagner, W. (2008). *Methodenprobleme bei der Analyse von Unterrichtswahrnehmung aus Schülersicht* [*Methodological issues in analysing students' perceptions of teaching*] (Doctoral dissertation). Germany: University of Koblenz-Landau. Retrieved from http://d-nb.info/987591800/34.

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

Wang, M. C., Haertel, G. D., & Walberg, H. J. (1993). Toward a knowledge base for school learning. *Review of Educational Research, 63*, 249–294. http://dx.doi.org/10.3102/00346543063003249.

Warm, T. A. (1989). Weighted likelihood estimation of ability in item response theory. *Psychometrika, 54*, 427–450. http://dx.doi.org/10.1007/BF02294627.

Weiß, R. H. (2006). *CFT 20-R. Grundintelligenztest Skala 2 – Revision* [Culture fair test]. Göttingen, Germany: Hogrefe.

Yuan, K. H., & Bentler, P. M. (2000). Three likelihood based methods for mean and covariance structure analysis with nonnormal missing data. *Sociological Methodology, 30*, 165–200. http://dx.doi.org/10.1111/0081-1750.00078.