Z Erziehungswiss
DOI 10.1007/s11618-015-0660-4

SCHWERPUNKT

ZfE

The same in German as in English? Investigating the subject-specificity of teaching quality

Anna-Katharina Praetorius · Svenja Vieluf · Steffani Saß · Andrea Bernholt · Eckhard Klieme

Abstract Teaching quality often is assumed to be a personal and stable characteristic of teachers. Whether this is true has scarcely been investigated empirically. In this study the extent to which value-added scores of teachers teaching German and English as a foreign language (EFL) to the same class remain consistent across subjects was investigated. Then, the consistency of two teaching quality dimensions—classroom management and motivational support—across subjects was explored. A sample consisting of 25 classes with 548 students to whom German and EFL were taught by the same teacher was analyzed using multivariate multilevel models and generalizability theory. The results showed that the value-added scores were highly correlated across subjects. While there was hardly any subject-dependent variance in classroom management, there was substantial subject-dependent variance in motivational support. The results indicate that it is important to conduct further studies

Dr. A.-K. Praetorius (✉) · Dr. S. Vieluf · Prof. Dr. E. Klieme
German Institute for International Educational Research (DIPF),
Schlossstraße 29,
60486 Frankfurt a. Main, Germany
e-mail: praetorius@dipf.de

Dr. S. Vieluf
e-mail: vieluf@dipf.de

Prof. Dr. E. Klieme
e-mail: klieme@dipf.de

Dr. S. Saß · Dr. A. Bernholt
Leibniz Institute for Science and Mathematics Education,
Olshausenstraße 62,
24118 Kiel, Germany
e-mail: sass@ipn.uni-kiel.de

Dr. A. Bernholt
e-mail: abernholt@ipn.uni-kiel.de

Published online: 10 November 2015

on the situational and contextual factors that might influence teaching quality to gain a more comprehensive picture regarding the consistency of teaching quality across various conditions.

Keywords Generalizability theory · Teacher effectiveness · Teaching quality · Subject specificity · Value-added model

Wie in Deutsch, so in Englisch? Zur Fachspezifität von Unterrichtsqualität

Zusammenfassung Unterrichtsqualität wird oft als stabiles Personenmerkmal von Lehrkräften interpretiert. Inwiefern dies angemessen ist, wurde bislang jedoch kaum empirisch überprüft. Im vorliegenden Beitrag steht die Stabilität von Unterrichtsqualität über Unterrichtsfächer im Fokus. Zum einen wurde untersucht, inwiefern Value-added-Werte für Lehrkräfte, die die Fächer Deutsch und Englisch in einer Klasse unterrichten, über diese Fächer stabil ausgeprägt sind und zum anderen, wie stabil zwei Basisdimensionen von Unterrichtsqualität (Klassenführung und unterstützendes Unterrichtsklima) über Fächer hinweg ausgeprägt sind. Es wurden Daten von 25 Klassen mit ihren 548 Schülerinnen und Schülern analysiert, die in den Fächern Deutsch und Englisch von derselben Lehrkraft unterrichtet wurden. Ausgewertet wurden die Daten mittels multivariaten Mehrebenenanalysen sowie Generalisierbarkeitsanalysen. Die Valued-added-Werte wiesen einen hohen Zusammenhang auf. Bei beiden untersuchten Unterrichtsqualitätsdimensionen überwog der fachübergreifende Anteil an Unterrichtsqualität, wobei für Klassenführung kaum fachspezifische Anteile, für das unterstützende Unterrichtsklima hingegen substanzielle fachspezifische Anteile gefunden wurden. Die Befunde weisen auf die Bedeutsamkeit weiterer Studien zum Einfluss situationaler und kontextueller Bedingungen auf Unterrichtsqualität hin, um einen umfassenderen Eindruck bezüglich der Variation von Unterrichtsqualität über verschiedene Bedingungen hinweg zu erhalten.

Schlüsselwörter Unterrichtsqualität · Effektivität · Fachspezifität · Generalisierbarkeitstheorie · Value-added-Modell

1 Theoretical background

Teaching quality often is believed to be strongly dependent on personal characteristics of teachers and thus to remain consistent across classes and school subjects. Many researchers (e.g., Kennedy 2010; Hiebert and Morris 2012) have, however, recognized the importance of taking into consideration the relevance of situational and contextual factors to teaching quality. In this study the extent to which teaching quality is subject-dependent for secondary school teachers teaching two subjects to the same class is investigated.

1.1 Approaches to research on the effectiveness of teaching

A large body of research has shown that teachers' contributions to student learning vary considerably (see e.g., Hanushek and Rivkin 2010). To determine the extent to which student achievement is influenced by teachers' teaching quality or by other factors, statistical value-added models (VAMs) have been developed. The idea behind VAMs is that measures of student achievement at a certain point in time can be adjusted for everything a teacher cannot be made responsible for; therefore, prior achievement and other factors such as students' personal characteristics, family backgrounds, and schools' characteristics are controlled for (for an overview, see McCaffrey et al. 2004). VAMs are widely used in the United States, where VAM scores sometimes have high-stakes consequences, for example when they are used in teacher evaluations and, accordingly, to determine teachers' pay (see e.g., Kersting et al. 2013). Consequently, many studies have been conducted of the validity of VAMs. These studies have shown great inconsistencies in results depending on the conditions under which VAMs are applied (e.g., different estimation methods, formulas, or criterion tests; see e.g., Papay 2011; Loeb and Candelaria 2013; Grossman et al. 2014). The VAM approach has also been criticized for focusing exclusively on student achievement and ignoring other important outcomes of schooling (e.g., motivational characteristics). In addition, it has been argued that student test scores are merely a distal indicator of teaching quality and thus have limited validity (e.g., Patrick and Mantzicopoulos 2014; for an overview, see Haertel 2013).

Another research tradition, namely the process-product paradigm, focuses on characteristics of teaching quality. Process-product models use student outcomes (= product) to identify characteristics indicative of high quality teaching (= process). In addition to student achievement, other student outcomes have been investigated within this paradigm (e.g., Kunter 2005; Rakoczy 2008). However, the focus of process-product studies is not on outcomes but rather on characteristics of teaching quality such as classroom processes, teaching methods, and/or teacher behavior. These characteristics usually are identified and assessed using data from observer ratings or student surveys. Lists of such characteristics have been compiled and several attempts have been made to integrate these lists into an overarching model. Interestingly, researchers from different countries such as Germany and the United States have made similar suggestions (see Klieme et al. 2001; Tschannen-Moran and Woolfolk-Hoy 2001; Klieme and Rakoczy 2003; Creemers and Kyriakides 2008; Lipowsky et al. 2009; Pianta and Hamre 2009; Baumert et al. 2010; Vieluf and Klieme 2011; Reyes et al. 2012; Kunter et al. 2013; Fauth et al. 2014). The overarching model developed contains three general dimensions considered essential to high quality teaching across education systems, school types, grade levels, and school subjects (Klieme and Rakoczy 2003): classroom management (also called organizational support), motivational support (also called emotional support), and potential for cognitive activation (also called instructional support).

Classroom management refers to effective prevention of and intervention in disruptions and disciplinary conflicts with the goal of maximizing students' learning time (see the seminal work of Kounin 1970). This can be achieved by clearly stating rules and routines or through efficient organization. Classroom management has the

most consistent effects on cognitive as well as motivational student outcomes (see e.g., Brophy 1986; Kunter et al. 2013). It is considered a basic precondition for learning in the classroom (Klieme et al. 2001).

Motivational support refers to teachers' sensitivity to the individual needs and situations of students and their ability to appreciate the students' perspectives (Davis 2003; Cornelius-White 2007; Pianta and Hamre 2009). This can be achieved, for example, by dealing with students' errors in an appreciative and constructive way in class. Motivational support has been linked mainly to motivational student outcomes (see e.g., Reeve et al. 2004; Patrick et al. 2007; Wentzel et al. 2010).

The aim of cognitive activation is to encourage students' cognitive engagement in higher-level thinking. This can be achieved by initiating challenging tasks and employing strategies such as activating previous knowledge (see also the concept of teaching for understanding, Cohen 1993; Mayer 2004). Cognitive activation has been most influential on cognitive student outcomes (see e.g., Klieme et al. 2001; Hamre and Pianta 2009; Lipowsky et al. 2009; Baumert et al. 2010).

1.2 Consistency in teaching effectiveness

The value-added approach and the process-product approach are similar in that they implicitly or explicitly assume consistency in teaching effectiveness across various conditions such as time, classes, and subjects. In previous studies, however, teaching quality usually was investigated only cross-sectionally for one class and one subject (Hiebert and Morris 2012; Hill et al. 2012). In these studies, differences in teaching were confounded with differences between classes and subjects. Accordingly, many researchers (e.g., Kennedy 2010; Hiebert and Morris 2012) have criticized that the relevance of situational and contextual factors to teaching quality has not been sufficiently considered. In this study, variation in teaching quality across two subjects taught to the same class is investigated.

1.3 Subject-dependent variation in teaching quality

Teaching quality usually is examined in a single school subject. However, researchers using VAMs or adhering to the process-product paradigm often are interested in general rather than subject-specific teaching quality. They tend to assume that teaching quality is dependent on general personal characteristics of the teacher and therefore does not vary much according to the subjects he or she teaches. There are, however, arguments suggesting that teaching quality depends on the subject being taught. First, teacher training is based on and organized according to school subjects, each of which tends to have its own subculture of norms and values (Klieme and Vieluf 2009; Baumert and Kunter 2013). Second, teachers' pedagogical content knowledge influences the quality of their instruction (Baumert et al. 2010). As this kind of knowledge is subject-specific, the teaching quality of individual teachers could differ according to the subjects they teach. Third, every subject has specific content which can influence the methods used as well as their effectiveness (Prange 2011). For example, it is common to conduct experiments during science classes but rather difficult to imagine many useful experiments conducted in English literature classes. Fourth, students'

knowledge, attitudes, and motivation can vary across subjects. As teaching is a co-constructive process, this variation could influence teaching. Fifth, in meta-analyses of teaching effectiveness, content-related features have shown the largest effect sizes (Seidel and Shavelson 2007).

All empirical studies of the amount of subject-related within-teacher differences in teaching effectiveness have been conducted in the United States and have focused mainly on the variation in VAMs in mathematics and reading. Further, in these studies, the teaching effectiveness of elementary school teachers, opposed to secondary school teachers, has been investigated, as only elementary school teachers in the United States teach several subjects to the same class. Correlations between mathematics and reading ranged from small to medium/large: between 0.35 and 0.64 in a study by Koedel and Betts (2007), between 0.21 and 0.58 in a study by Loeb et al. (2012), between 0.47 and 0.54 in a study by Goldhaber et al. (2012), and between 0.23 and 0.53 in a study by Rowan et al. (2002). Loeb and Candelaria (2013), however, pointed out that the inconsistencies in VAMs across subjects taught by the same teachers found in previous studies might have been due to differences among teachers in their effectiveness or to measurement error. The authors stressed the importance of further studies to provide evidence of the sources of inconsistencies in teaching effectiveness across subjects. VAM approaches are of limited use for such purposes as they cannot provide evidence of what actually is happening in the classroom. Using more proximal indicators of teaching quality identified in process-product research seems more suitable for investigating the subject-specificity of teaching quality. This is the aim of the present study.

1.4 Research questions and hypotheses

To gain a better understanding of the subject-specificity of teaching quality, the present study extends previous VAM findings by providing correlations for value-added scores for secondary school teachers (instead of primary school teachers) and for a different set of subjects, namely German and English as a Foreign Language (EFL) (instead of mathematics and reading). In line with previous studies, correlations between value-added scores in German and EFL are expected to be small to medium-sized (Hypothesis 1).

Second, in this study consistency in two dimensions of teaching quality is investigated in the two subjects being taught by one teacher to the same class. Although it seems reasonable to expect differences in teaching quality between subjects, it is not plausible that all teaching quality dimensions are influenced to the same extent by the subject being taught. Content-related aspects of teaching quality can be assumed to vary more than aspects related to teachers or students, or the relationship between teachers and students. In this study, the subject dependency of teaching quality in terms of classroom management and motivational support—two general teaching quality dimensions hardly related to the content being taught[1]—is investigated. Classroom management depends on teachers' general pedagogical knowledge (see e.g., Baumert et al. 2010). Motivational support refers, among others, to the relationship between teachers and students. Thus, it can be expected that classroom management (Hypothesis 2) as well as motivational support (Hypothesis 3) are rather stable across subjects.

2 Method

In this study, data from the Assessment of Student Achievements in German and English as a Foreign Language (DESI) study (DESI-Konsortium 2008) was reanalyzed. The DESI study was conducted in the 2003–2004 school year and researchers not only assessed the language competencies of 9th-grade students, but also collected information on classroom and school processes as well as details about the students, classes, and schools.

2.1 Sample

The sample drawn for the DESI study was representative of 9th-graders attending secondary school (i.e., academic track, intermediate track, and lower track secondary schools) in Germany. In each of the 219 schools that participated two classes were sampled resulting in 427 teachers (55 % female; 16 % missing information). The teachers had been working in their profession for an average of 20.11 years ($SD = 11.14$). Participation in the study was compulsory for all students in the classes sampled.

To investigate the subject-specificity of VAMs and the teaching quality dimensions, only a subsample from the DESI study was used; we selected teachers who taught their students both German and EFL. This was possible because in Germany secondary school teachers usually teach at least two subjects in academic track secondary schools and possibly more subjects in lower track secondary schools. Our subsample consisted of 25 teachers (68 % female) and their 548 students. These teachers had been working in their profession for 20.48 years on average ($SD = 10.37$); 80 % of these teachers had studied German and 56 % EFL at university.

2.2 Instruments

2.2.1 Achievement tests

Student achievement in German To assess student achievement in German, a language awareness test was administered at the beginning and again at the end of 9th grade. It consisted of 34 items of varying difficulty, ranging from knowledge of simple grammatical structures (level 1) to active application of declarative knowledge (level 5). The test measured the extent to which the adolescents were able to use language in a grammatically correct as well as stylistically appropriate way (see Eichler 2008). A rotated booklet design was used to ensure that students worked on different items at the two time points. The data were scaled using a longitudinal multidimensional Rasch model with virtual persons (Rost 2004), estimating five plausible values (PVs) per student (for further details regarding the scaling, see Hartig et al. 2008). The EAP/PV reliability was 0.79 at both time points.

Student achievement in EFL. To measure students' achievement in EFL, a C test comprising 29 items was administered (see Harsch and Schroeder 2008). C tests are written tests that assess general language abilities in a foreign language. The students

worked on four short texts at both time points; again, a rotated booklet design was used. The difficulty of the items was ranging from filling gaps in basic language structures (level 1) to filling gaps in larger stretches of text (level 5). The data were scaled using a longitudinal multidimensional Rasch model with virtual persons; five plausible values (PVs) per student were estimated. The EAP/PV reliability was 0.90 at both time points.

2.2.2 Control variables

Student level Four student characteristics were included in the value-added analyses: (a) gender, (b) basic cognitive abilities, measured with the German version of the Cognitive Ability Test (Thorndike and Hagen 1993), (c) socioeconomic status, measured using the highest International Socio-Economic Index of Occupational Status of both parents (HISEI) (see Ganzeboom et al. 1992), and (d) first language, using two dummy variables (dummy 1: German and another language, dummy 2: language other than German; only German was used as the reference category) (for further information regarding these measures, see DESI-Konsortium 2008).

Class level. We included all four variables measured at the student level also at the class level. Additionally, we included secondary school type, using two dummy variables (academic track, lower track; intermediate track was used as the reference category).

2.2.3 Teaching quality assessment

Teaching quality was assessed with student ratings aggregated to the class level. Four-point Likert-type scales ranging from 1 (= totally disagree) to 4 (= totally agree) were used.

For classroom management, two subscales were combined (classroom management and discipline problems); the higher order factor for this combined scale had an internal consistency of $\alpha_G = 0.82$ for German and $\alpha_E = 0.79$ for EFL. Classroom management was assessed using two items (e.g., "My German/EFL teacher always knows exactly what is happening in the classroom."). The internal consistencies were $\alpha_G = 0.77$ and $\alpha_E = 0.87$. The intraclass correlations (ICC) at level 2 (i.e., indicators of the class-level reliabilities) were $ICC(2)_G = 0.85$ and $ICC(2)_E = 0.88$. Discipline problems were measured using seven items (e.g., "During our German/EFL instruction, many disruptions occur."). The internal consistencies were $\alpha_G = 0.87$ and $\alpha_E = 0.88$, and the ICC(2)s were $ICC(2)_G = 0.85$ and $ICC(2)_E = 0.88$.

Motivational support was operationalized with three scales (student orientation, supportiveness, instructional climate; $\alpha_G = 0.94$ for the higher order factor for German and $\alpha_E = 0.93$ for EFL). Student orientation was measured using six items (e.g., "If someone has a good idea, my German/EFL teacher acknowledges it."). The internal consistencies were $\alpha_G = 0.87$ and $\alpha_E = 0.87$, and the ICC(2)s were $ICC(2)_G = 0.85$ and $ICC(2)_E = 0.86$. To measure teacher support, three items were used (e.g., "My German/EFL teacher provides me with help when I need it."). The internal consistencies were $\alpha_G = 0.84$ and $\alpha_E = 0.86$, and the ICC(2)s were $ICC(2)_G = 0.81$ and

$ICC(2)_E = 0.84$. Classroom climate was measured using three items (e.g., "My German/EFL teacher likes me."). The internal consistencies were $\alpha_G = 0.83$ and $\alpha_E = 0.76$, and the ICC(2)s were $ICC(2)_G = 0.78$ and $ICC(2)_E = 0.79$.

2.3 Analyses

2.3.1 Consistency in VAM scores across subjects

To assess consistency in VAM scores across subjects, multivariate two-level models (students nested in teachers; see Raudenbush and Bryk 2002) were used. The equations for each of the two dependent variables were as follows:

$$\text{Model level 1: } y_{ij} = \beta_{0j} + \beta_{pj} S_{pij} + r_{ij}$$

$$\text{Model level 2: } \beta_{0j} = \gamma_{00} + \gamma_{0p} C_{pj} + u_{0j}$$

$$\beta_{pj} = \gamma_{10} + u_{pj}$$

At level 1, the achievement of student i of teacher j was a function of the intercept (β_{0j}), student variables (S_{pij}; predictor p for student i for teacher j), as well as random error (r_{ij}). The residuals for the intercept as well as the slopes were modeled as random at level 2 (u_{0j}; u_{pj}), thus allowing variation among teachers. At level 2, all class variables (C_{pj}; predictor p for teacher j) were included as predictors. The correlations between the two dependent variables at level 2 could be interpreted as consistent teaching effectiveness across subjects.

The models were estimated in Mplus 7.1 (Muthén and Muthén 1998–2012) using robust maximum likelihood estimation. To estimate the models, all five plausible values were used.

2.3.2 Consistency in teaching quality dimensions across subjects

The data on the two teaching quality dimensions, classroom management and motivational support, were analyzed according to generalizability theory (G theory). G theory enables multiple sources of errors (called facets) to be separated via variance component analysis (Shavelson and Webb 1991; Brennan 2001a). To identify consistencies across the two subjects, a fully crossed two-facet design (teacher $t \times$ subject $s \times$ subscale su) was applied (for discussion on whether to treat subjects as a random facet, see Sect. 4.4). Analyses were conducted separately for both dimensions. Using the G theory framework allowed differentiation of (1) between-teacher differences across subjects (teacher variance component σ^2_t; variance component teacher by subscale σ^2_{tsu}), (2) subject-specific differences (variance component subject σ^2_s; variance component teacher by subject σ^2_{ts}; variance component subject by subscale σ^2_{ssu}), and (3) further variance components and error variance (variance component subscale σ^2_{su}; variance component teacher by subject by subscale $\sigma^2_{tssu,e}$).

To infer the reliability of the data, relative as well as absolute G coefficients were calculated. Both can be interpreted in a way similar to classic reliability coefficients.

Table 1 Descriptive statistics for the included variables in the value-added analyses

Variable	*M*	*SD*
Student variables		
Student achievement at T1 German	457.36	96.90
Student achievement at T2 German	490.26	102.04
Student achievement at T1 EFL	476.23	100.91
Student achievement at T2 EFL	499.14	106.65
IQ	49.40	11.16
Female	0.59	0.49
HISEI	48.42	16.53
Language mixed	0.04	0.20
Language other than German	0.10	0.30
Class variables		
Mean student achievement at T1 German	438.40	78.48
Mean student achievement at T2 German	465.02	91.04
Mean student achievement at T1 EFL	460.69	82.83
Mean student achievement at T2 EFL	477.22	94.23
Mean IQ	47.28	8.35
Proportion of females	0.49	0.14
Mean HISEI	48.10	9.37
Proportion of language mixed	0.05	0.06
Proportion of language other than German	0.14	0.15
Proportion of academic track	0.20	0.40
Proportion of lower academic track	0.56	0.50

For all analyses, the interaction between teachers and subscales was counted as universe score variance (see also Praetorius et al. 2012).

Small negative variance estimates can occur occasionally due to sampling errors. As suggested by Brennan (2001a), these negative variances were used to calculate the remaining variance components and subsequently were set to zero.

The G analyses were conducted with version 2.1 of the urGENOVA program (Brennan 2001b). The implemented estimator in this program is the ANOVA procedure. A large advantage of this estimator is that normality assumptions are not required (Brennan 2001a; see also the simulation study by Shumate et al. 2007).

Results

Descriptive statistics

In Table 1, the mean scores and standard deviations for all variables included in the value-added analyses can be found.

Table 2 contains descriptive statistics regarding the two teaching quality dimensions investigated, which have been separated for the subsample of teachers teaching German and EFL to the same class and for the rest of the sample, that is, those teachers teaching either only German or only EFL to a class.

Table 2 Descriptive statistics for the scales of the three teaching quality dimensions, separately for the subjects German and EFL as well as the two samples

Dimension	*M*	*SD*
Sample with different teachers in German and EFL (n = 402)		
Classroom management		
German	2.73/2.62	0.38/0.44
EFL	2.73/2.62	0.34/0.40
Motivational support		
German	2.73/2.78/2.67	0.34/0.35/0.33
EFL	2.64/2.79/2.64	0.33/0.33/0.33
Sample with the same teacher in German and EFL (n = 25)		
Classroom management		
German	2.83/2.74	0.39/0.44
EFL	2.82/2.70	0.40/0.44
Motivational support		
German	2.80/2.96/2.87	0.29/0.29/0.28
EFL	2.69/2.90/2.84	0.24/0.27/0.28

The different values between the slashes refer to the subscales used for measuring classroom management and motivational support

Variation in VAM scores between subjects

The results for all included control variables in the model can be found in Table 3. Estimating the multivariate multilevel model revealed a correlation of $r = 0.76$ ($SE = 0.13$; $p < .001$) between the VAM scores for teachers teaching both German and EFL to the same class. Hypothesis 1, assuming a small to medium correlation between the VAM scores, thus must be rejected.

The central assumption of VAMs is that all variance in student achievement that cannot be attributed to the teacher is partialled out by adjusting the scores for prior student achievement and for control variables at the student and class levels. To test whether this assumption was true the same model for the remaining sample of students who were taught German and EFL by different teachers was employed. The correlation between VAM scores was $r = 0.27$ ($SE = 0.06$; $p < .001$). This indicates that the adjustment of the data was not completely successful.

Variation in teaching quality between subjects

To investigate the extent to which teaching quality of a specific teacher varied between the two subjects taught, a teacher × subject × subscale variance component analysis was conducted. The G analyses revealed that, in addition to residual and subscale variance, 72 % of the total variance in classroom management was due to subject-independent teacher differences (*t, tsu*) in classroom management, whereas 7 % was due to subject-specific differences (*s, ssu, ts*) in classroom management (see Table 4). Thus, the ratio of variance in subject-independent to subject-dependent teacher differences was approximately 10:1. Therefore, finding only small subject-dependent variation in classroom management, Hypothesis 2 can be confirmed.

Table 3 Multivariate multi-level model predicting the achievement scores at T2 in German and EFL using student and class control variables

Variables	German		EFL	
	β	*SE*	*β*	*SE*
Student variables				
Student achievement at time 1	0.65***	0.06	0.81***	0.03
IQ	0.28***	0.05	0.14***	0.03
Female	0.07*	0.03	0.06*	0.03
HISEI	0.05	0.04	0.01	0.02
Language mixed	−0.03	0.03	−0.02	0.03
Language other than German	−0.04	0.03	−0.02	0.02
Class variables				
Mean student achievement at time 1	0.64***	0.17	0.75***	0.07
Mean IQ	0.12	0.22	0.18*	0.09
Percentage of females	0.13	0.10	0.19**	0.06
Mean HISEI	0.03	0.10	0.11	0.07
Percentage of language mixed	0.10	0.06	0.01	0.03
Percentage of language other than German	0.01	0.07	−0.07	0.05
Academic track	0.02	0.10	−0.14*	0.06
Lower academic track	−0.14	0.10	0.06	0.06
R^2				
Level 1	0.81		0.87	
Level 2	0.94		0.98	

***$p < .001$; **$p < .01$; *$p < .05$ (two-tailed)

Table 4 G Analyses for the teacher × subject × subscale design

Variance component	Classroom management		Motivational support	
	VC	%	VC	%
Stable components				
t	0.084	29	0.031	28
tsu	0.124	43	0.021	9
Subject-specific components				
s	0.000	0	0[a]	0
ts	0.020	7	0.021	19
ssu	0[a]	0	0.000	0
Subscale-specific component				
su	0[a]	0	0.003	3
Error				
tssu,e	0.063	21	0.046	41
Total variance	0.29		0.11	
$E\rho^2$	0.90		0.72	
Φ	0.86		0.70	

t teacher, *s* subject, *su* subscale, *e* residual, *VC* absolute variance component, *%* relative variance component, $E\rho^2$ relative G coefficient, *Φ* absolute G coefficient. The interaction *t* × *su* was set as universe score-variance for computing the G coefficients

[a]A small negative variance component was estimated and thus set to zero

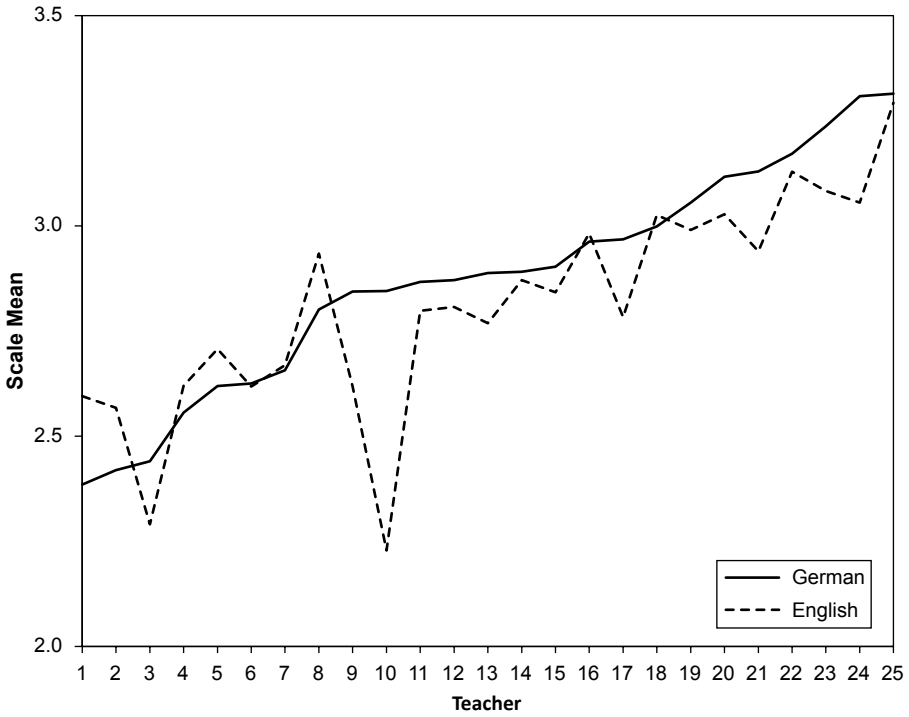

Fig. 1 Scale means for the dimension motivational support for the subjects German and EFL. Scale range: 1–4. The teachers are arranged according to their scale mean in German

Of the total variance in motivational support, 37% was due to subject-independent teacher differences (*t, tsu*) in motivational support, whereas 19% was due to subject-specific differences (*s, ssu, ts*) in motivational support. Hence, the ratio of variance in subject-independent to subject-dependent teacher differences was approximately 2:1. The subject-specific variance was due mainly to an interaction between teacher and subject (see Fig. 1). Thus, finding substantial subject-dependent variation in motivational support, Hypothesis 3 must be rejected.

As teachers and classes were confounded in the data, further analyses of the remaining classes of the DESI study were conducted. In these classes, German and EFL were taught by different teachers, which allowed the amount of variance attributable to the classes across subjects to be determined. Analysis with the class × subject × subscale design revealed the following: For both teaching quality dimensions, the majority of the systematic variance was due to a subject-by-class interaction (49% for classroom management, 54% for motivational support), indicating large differences in teaching quality between subjects taught to one class (confounded with teachers) (see Table 5). The stable, class-specific amount of variance (*c, csu*) was 10% for classroom management and 3% for motivational support. The ratio of variance in subject-independent to subject-dependent class differences in teaching quality was 1:5 for classroom management and 1:18 for motivational support.

Table 5 G analyses for the class × subject × item design

Variance component	Classroom management		Motivational support	
	VC	%	VC	%
Stable components				
c	0[a]	0	0.002	1
csu	0.027	10	0.004	2
Subject-specific components				
s	0[a]	0	0.001	0
cs	0.130	49	0.106	54
ssu	0[a]	0	0.001	1
Item-specific component				
i	0.002	1	0.004	2
Error				
cssu,e	0.107	40	0.078	40
Total variance	0.265		0.195	
$E\rho^2$	0.72		0.74	
Φ	0.70		0.71	

c class, *s* subject, *su* subscale, *e* residual, *VC* absolute variance component, *%* relative variance component, $E\rho^2$ relative G coefficient; Φ absolute G coefficient. The interaction $c \times su$ was set as universe score-variance for computing the G coefficients

[a]A small negative variance component was estimated and thus set to zero

Discussion

It often is assumed that teaching quality remains consistent across various conditions such as classes and subjects. However, many researchers have failed to consider the influence of situational and contextual factors on teaching quality. In this study the teaching quality of secondary school teachers teaching German and EFL to the same class was investigated to determine the extent to which teaching quality is subject-dependent.

Consistency in achievement gains across subjects

A high correlation between the German and EFL VAM scores was found for teachers teaching German and ELF to the same class. There are several possible reasons for this correlation to be higher than those found in previous studies. First, the content of the two subjects investigated and the corresponding requirements for students in this study were much more similar than those of the subjects examined in previous studies (i.e., mathematics and reading). Teachers might therefore be more likely to be equally effective in teaching German and EFL. Second, in this study secondary school teachers were investigated rather than elementary school teachers. One of the main differences between these two groups is that elementary school teachers have to teach main subjects regardless of their preferences or dislikes for certain subjects. Secondary school teachers, on the contrary, choose the subjects they teach, which might lead to similar efficacy in teaching different subjects. Third, all previous studies were conducted in the United States whereas this study was conducted in Germany; therefore, the differences in the findings might have a cultural basis.

The results indicate a need to consider carefully whether generalizing findings from a study of the quality of teaching a subject to a particular sample in a specific school type and country is appropriate.

The results of the present study point towards another critical issue regarding VAM analyses. The central assumption in VAM analyses is that VAM scores represent the part of the variance in student achievement scores attributable solely to the teacher. Results of the analyses in this study indicate that this is not the case, as the VAM scores for classes to which German and EFL were taught by different teachers were correlated to a significant degree as well. Although variables usually explored in VAM studies were included in this study, not all determinants of teacher-unrelated variance were controlled for in the data. This finding supports the caution raised by several researchers regarding the validity of VAMs (see e.g., Haertel 2013).

Consistency in dimensions of teaching quality across subjects

Results of the analyses in this study indicate that for both dimensions of teaching quality investigated, subject-independent variance was greater than subject-dependent variance. However, the two dimensions differed in the amount of subject-dependency: The ratio of subject-independent to subject-dependent variance for classroom management was 10:1 and for motivational support 2:1. This indicates that classroom management was less dependent on the subject being taught than motivational support. This was in accordance with assumptions in previous literature on classroom management. The results of this study provide an explanation for the consistent effects of classroom management on student achievement in previous studies (see e.g., Emmer and Stough 2001): If classroom management is a characteristic of teaching quality that is consistent across situations and contexts, effects of classroom management are not as dependent as other teaching quality dimensions on characteristics of individual studies. Further, teacher training programs usually do not distinguish between subjects when covering classroom management issues (see e.g., Borg and Ascione 1982; Piwowar et al. 2013). The present study showed evidence that this practice is justified.

In contrast, subject-dependency for motivational support was much greater. Having a closer look at the three subscales used to measure motivational support—student orientation, supportiveness, and instructional climate—one can easily imagine differences among subjects in student orientation. Approving of students' ideas and opinions might be easier in a subject in which students can express themselves in their first language. However, supportiveness (e.g., helping students) and instructional climate (e.g., giving students the impression they are liked) seem to be rather independent of the subject being taught. Whether or not subject-to-subject variation occurred only for some subscales can be inferred from the G analyses: If there are such differences among subscales, the subject-by-subscale interaction would be large. In the present study, however, this two-way interaction equaled zero. Thus, the differences in motivational support between the two subjects could be identified similarly across all subscales used. For the majority of teachers, motivational support seemed greater in German than in EFL; however, there were some teachers for whom it was the other way around (see Fig. 1). Explanations for these differences could be

the varying achievement levels of the students in the two subjects or differing content and pedagogical content knowledge of teachers in the two subjects. Independent of the reasons for the occurring differences, one conclusion to be drawn from the present analyses is that because there is considerable variation in motivational support between subjects, subjects should be held constant or be varied systematically in studies of teaching quality to gain comparable scores for different teachers.

The following comment was made by an anonymous reader of Kennedy's (2010) paper: "while situations may vary from teacher to teacher and these differences in situations may make effects on students variable, teachers must somehow accommodate these variations in situation *and still be effective*". Kennedy (2010) replied to this comment arguing that measuring the quality of teaching requires taking into account such possibly aggravating circumstances for teaching. The results of the present study indicate that subjects are one source of such variations in teaching quality, but their influence is not large. Based on the present findings, the conclusion might be drawn that it is the teachers that are most relevant (see Barber and Mourshed 2007; Hanushek and Rivkin 2012; Chetty et al. 2013). However, this is, for several reasons, not necessarily the case: (a) The two general dimensions investigated are conceptualized as subject-independent. The third dimension, cognitive activation, on the contrary, is related to the content being taught. It thus can be expected that cognitive activation is more subject-dependent than classroom management or motivational support. (b) Several other situational and contextual characteristics might influence teaching quality such as the composition of students in a class (as, e.g., Rjosk et al. 2014, have shown with DESI data for the socioeconomic composition). In our study, class and teacher effects were confounded. To determine the extent to which the teacher effect found really was a teacher effect and not a class effect we conducted further analyses of the remaining sample. We found that the ratio of variance in teaching quality attributable to the class vs. the subjects was 1:5 for classroom management and 1:18 for motivational support. In this case subjects were confounded with teachers, as the subjects were taught by different teachers. This indicates that at least for classroom management, class characteristics might play a significant role in teaching quality (see Rjosk et al. 2014). To understand better the importance of teacher vs. class effects, it would be useful to combine both effects in one design to disentangle them properly.

Different approaches, coherent picture?

The VAM analyses indicate that the effectiveness of teaching can be generalized across subjects. Thus, characteristics of teachers and/or their teaching practices that are subject-independent obviously play an important role in teaching effectiveness. According to the G analyses, classroom management is one possible explanatory variable for consistency across subjects. Motivational support, in contrast, might help explain why value-added scores are not entirely consistent across subjects. The analyses therefore do not contradict each other but rather supplement each other. They also indicate that analyzing dimensions of teaching quality allows investigation into possible reasons for consistency or inconsistency across conditions; these reasons cannot be identified analyzing VAM scores alone.

Limitations and further directions

In the present study, only two general dimensions of teaching quality were investigated; the third dimension, cognitive activation, was not. This is especially misfortunate as variation between subjects would be expected for this dimension in particular. To investigate the subject-specificity of cognitive activation in the future, new questionnaires or other measures will need to be developed, as cognitive activation as yet has been mainly investigated with regard to mathematics (for an exception see Lotz 2016).

The results of the present study rely—as all empirical studies do—on the sample and the instruments used. Although the total sample was representative of the German context, the subsample used for the G analyses was not. The teachers in the sample teaching German and EFL seemed to be a positive selection of the total sample (see the descriptive statistics in Table 1). Also, the sample size was small for VAM analyses, leading to high standard errors in the estimations. Moreover, the results might have been specific for the German context. As conceptualizations of teaching quality vary among studies (see also Praetorius et al. 2014), our results might additionally not be generalizable beyond the instruments used. Replication of the results therefore seems important.

In the present investigation, subjects were treated as random facets within the G analyses although they are indeed fixed facets. This procedure was taken as suggested by Shavelson and Webb (1991) to get an impression of the influence of a fixed facet. An alternative would be to analyze the data separately for both subjects. However, as the comparison of both subjects was the main interest of the present paper, this was not an option. Another alternative would be to use a multivariate design (see Brennan 2001a). With such a design, covariation between subjects could be computed for the different variance components. However, the results cannot be interpreted as easily as in the univariate design.

Another limitation is related to the small total variance for both teaching quality dimensions. One reason could be selectivity of the sample; teachers of both German and EFL might be more similar to each other with respect to teaching quality than teachers with different combinations of subjects. However, compared to the total variance for the rest of the sample (see the class × subject × subscale design), only the total variance for motivational support was small whereas the total variance for classroom management was somewhat larger for the subsample. Thus, another possible reason for the small total variance is more reasonable: the subscales on the questionnaire did not differentiate enough between teachers/classes, as only a few easy and difficult items were included. Developing new instruments enabling such differentiation, thus, seems to be an important task for future research.

Conclusions

Overall, teaching effectiveness as measured by value-added scores seems to be rather consistent in German and EFL. Nevertheless, the subject being taught seems to influence significantly teachers' motivational support. The results of the present study indicate that it is important to conduct further studies on the situational and contextual factors that might influence teaching quality to gain a more comprehensive picture regarding the consistency of teaching quality across various conditions.

Notes

1 The third dimension, potential for cognitive activation, was not investigated, as the data analyzed in the present study did not contain such information.

References

Barber, M., & Mourshed, M. (2007). *How the world's best-performing school systems came out on top*. London: McKinsey.

Baumert, J., & Kunter, M. (2013). The COACTIV model of teachers' professional competence. 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. Results from the COACTIV project* (pp. 25–48). New York: Springer.

Baumert, J., Kunter, M., Blum, W., Brunner, M., Voss, T., & Jordan, A. (2010). Teachers' mathematical knowledge, cognitive activation in the classroom, and student progress. *American Educational Research Journal, 47*, 133–180.

Borg, W. R., & Ascione, F. R. (1982). Classroom management in elementary mainstreaming classroom. *Journal of Educational Psychology, 74*, 85–95.

Brennan, R. L. (2001a). *Generalizability theory*. New York: Springer.

Brennan, R. L. (2001b). *Manual for urGenova (Version 2.1)*. Iowa City: Iowa Testing Programs, University of Iowa.

Brophy, J. (1986). Teacher influences on student achievement. *American Psychologist, 41*(10), 1069–1077.

Chetty, R., Friedmann, J. N., & Rockoff, J. E. (2013). *Measuring the impact of teachers I: Evaluating bias in teacher value-added estimates*. NBER Working Paper No. 19423.

Cohen, D. K. (1993). *Teaching for understanding: Challenges for policy and practice*. San Francisco: Jossey-Bass.

Cornelius-White, J. (2007). Learner-centered teacher-student relationships are effective: A meta-analysis. *Review of Educational Research, 77*, 113–143.

Creemers, B. P. M., & Kyriakides, L. (2008). *The dynamics of educational effectiveness: A contribution to policy, practice and theory in contemporary schools*. 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*, 207–234.

DESI-Konsortium (Ed.) (2008). *Unterricht und Kompetenzerwerb in Deutsch und Englisch: Ergebnisse der DESI-Studie*. Weinheim: Beltz.

Eichler, W. (2008). Sprachbewusstheit Deutsch. In DESI-Konsortium (Ed.), *Unterricht und Kompetenzerwerb in Deutsch und Englisch. Ergebnisse der DESI-Studie* (pp. 149–156). Weinheim: Beltz.

Emmer, E. T., & Stough, L. M. (2001). Classroom management: A critical part of educational psychology, with implications for teacher education. *Educational Psychologist, 36*, 103–112.

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

Ganzeboom, H. B. G., De Graaf, P., Treiman, D. J., & De Leeuw, J. (1992). A standard international socio-economic index of occupational status. *Social Science Research, 21*(1), 1–56.

Goldhaber, D., Cowan, J., & Walch, J. (2012). *Is a good elementary teacher always good? Assessing teacher performance estimates across subjects. CEDR working paper 2012—7.2*. Seattle: University of Washington.

Grossman, P., Cohen, J., Ronfeldt, M., & Brown, L. (2014). The test matters: The relationship between classroom observation scores and teacher value added on multiple types of assessment. *Educational Researcher, 43*, 293–303.

Haertel, E. H. (2013). *Reliability and validity of inferences about teachers based on student test scores*. Princeton: Education Testing Service.

Hanushek, E. A., & Rivkin, S. G. (2010). Generalizations about using value-added measures of teacher quality. *American Economic Review, 100*, 267–271.

Hanushek, E. A., & Rivkin, S. G. (2012). The distribution of teacher quality and implications for policy. *Annual Review of Economics, 4*, 131–157.

Harsch, C., & Schröder, K. (2008). Textrekonstruktion Englisch. In DESI-Konsortium (Ed.), *Unterricht und Kompetenzerwerb in Deutsch und Englisch. Ergebnisse der DESI-Studie* (pp. 149–156). Weinheim: Beltz.

Hartig, J., Jude, N., & Wagner, W. (2008). Methodische Grundlagen der Messung und Erklärung sprachlicher Kompetenzen. In DESI-Konsortium (Ed.), *Unterricht und Kompetenzerwerb in Deutsch und Englisch. Ergebnisse der DESI-Studie* (pp. 34–54). Weinheim: Beltz.

Hiebert, J., & Morris, A. K. (2012). Teaching, rather than teachers, as a path toward improving classroom instruction. *Journal of Teacher Education, 63*, 92–102.

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

Kennedy, M. M. (2010). Attribution error and the quest for teacher quality. *Educational Researcher, 39*, 591–598.

Kersting, N. B., Chen, M.-K., & Stigler, J. W. (2013). Value-Added Teacher Estimates as Part of Teacher Evaluations: Exploring the Effects of Data and Model Specifications on the Stability of Teacher-Value Added Scores (Special issue on value-added research for policy). Educational Policy Analysis Archives.

Klieme, E., & Rakoczy, K. (2003). Unterrichtsqualität aus Schülerperspektive: Kulturspezifische Profile, regionale Unterschiede und Zusammenhänge mit Effekten von Unterricht. In J. Baumert, C. Artelt, E. Klieme, M. Neubrand, M. Prenzel, U. Schiefele, W. Schneider, & K.-J. Tillmann (Eds.), *PISA 2000. Ein differenzierter Blick auf die Länder der Bundesrepublik Deutschland* (pp. 334–359). Opladen: Leske + Budrich.

Klieme, E., & Vieluf, S. (2009). Teaching practices, teachers' beliefs and attitudes. In OECD (Ed.), *Creating effective teaching and learning environments. First results from TALIS* (pp. 87–135). Paris: OECD.

Klieme, E., Schümer, G., & Knoll, S. (2001). Mathematikunterricht in der Sekundarstufe I: Aufgabenkultur und Unterrichtsgestaltung. In Bundesministerium für Bildung und Forschung (BMBF) (Ed.), *TIMSS—Impulse für Schule und Unterricht. Forschungsbefunde, Reforminitiativen, Praxisberichte und Video-Dokumente* (pp. 43–57). Munich: Medienhaus Biering.

Koedel, C., & Betts, J. (2007). Re-Examining the Role of Teacher Quality In the Educational Production Function (Working Papers 0708, Department of Economics, University of Missouri).

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

Kunter, M. (2005). *Multiple Ziele im Mathematikunterricht*. Münster: Waxmann.

Kunter, M., Klusmann, U., Baumert, J., Richter, D., Voss, T., & Hachfeld, A. (2013). Professional competence of teachers: Effects on quality and student development. *Journal of Educational Psychology, 105*, 805–820.

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.

Loeb, S., & Candelaria, C. A. (2013). How stable are value-added estimates across years, subjects and student groups? http://www.carnegieknowledgenetwork.org/briefs/value-added/value-added-stability/. Accessed 10 Sept 2015.

Loeb, S., Kalogrides, D., & Beteille, T. (2012). Effective schools: Teacher hiring, assignment, development, and retention. *Education Finance and Policy, 7*(3), 269–304.

Lotz, M. (2016). *Kognitive Aktivierung im Leseunterricht der Grundschule. Eine Videostudie zur Gestaltung und Qualität von Leseübungen im ersten Schuljahr*. Wiesbaden: Springer VS.

Mayer, R. E. (2004). Should there be a three-strikes rule against pure discovery learning? *American Psychologist, 59*, 14–19.

McCaffrey, D. F., Lockwood, J. R., Koretz, D., Louis, T. A., & Hamilton, L. (2004). Models for value-added modeling of teacher effects. *Journal of Educational and Behavioral Statistics, 29*, 67–101.

Muthén, B., & Muthén, L. (1998–2012). *Mplus* (Version 7.11). Los Angeles: StatModel.

Papay, J. P. (2011). Different test, different answers: The stability of teacher value-added estimates across outcome measures. *American Educational Research Journal, 48*(1), 163–193.

Patrick, H., & Mantzicopoulos, P. (2014). Is effective teaching stable? *The Journal of Experimental Education, 11*, 1–25.

Patrick, H., Ryan, A. M., & Kaplan, A. (2007). Early adolescents' perceptions of the classroom social environment, motivational beliefs, and engagement. *Journal of Educational Psychology, 99*, 83–98.

Pianta, R. C., & Hamre, B. K. (2009). Conceptualization, measurement, and improvement of classroom processes: Standardized observation can leverage capacity. *Educational Researcher, 38*, 109–119.

Piwowar, V., Thiel, F., & Ophardt, D. (2013). Training inservice teachers' competencies in classroom management—a quasi-experimental study with teachers of secondary schools. *Teaching and Teacher Education, 30*, 1–12.

Praetorius, A.-K., Lenske, G., & Helmke, A. (2012). Observer ratings of instructional quality: Do they fulfill what they promise? *Learning and Instruction, 6*, 387–400.

Praetorius, A.-K., Pauli, C., Reusser, K., Rakoczy, K., & Klieme, E. (2014). One lesson is all you need? Stability of instructional quality across lessons. *Learning and Instruction, 31*, 2–12.

Prange, K. (2011). Didaktik und Methodik. In J. Kade, W. Helsper, C. Lüders, B. Egloff, F.-O. Radtke, & W. Thole (Eds.), *Pädagogisches Wissen. Erziehungswissenschaft in Grundbegriffen* (pp. 183–190). Stuttgart: Kohlhammer.

Rakoczy, K. (2008). *Motivationsunterstützung im Mathematikunterricht – Unterricht aus der Perspektive von Lernenden und Beobachtern*. Münster: Waxmann.

Raudenbush, S. W., & Bryk, A. S. (2002). *Hierarchical linear models* (2nd edition). Thousand Oaks: Sage Publications.

Reeve, J., Jang, H., Carrell, D., Jeon, S., & Barch, J. (2004). Enhancing students' engagement by increasing teachers' autonomy support. *Motivation and Emotion, 28*, 147–169.

Reyes, M. R., Brackett, M. A., Rivers, S. E., White, M., & Salovey, P. (2012). Classroom emotional climate, student engagement, and academic achievement. *Journal of Educational Psychology, 104*, 700–712.

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

Rost, J. (2004*). Lehrbuch Testtheorie – Testkonstruktion* (2nd edn.). Bern: Huber.

Rowan, B., Correnti, R., & Miller, R. (2002). *What large-scale, survey research tells us about teacher effects on student achievement: Insights from the Prospects Study of Elementary Schools (Conceptual Paper)*. Michigan: University of Michigan.

Seidel, T., & Shavelson, R. J. (2007). Teaching effectiveness research in the last decade: Role of theory and research design in disentangling meta-analysis results. *Review of Educational Research, 77*, 454–499.

Shavelson, R., & Webb, N. (1991). *Generalizability theory: A primer*. Thousand Oaks: Sage Publications.

Shumate, S. R., Surles, J., Johnson, R. L., & Penny, J. (2007). The effects of the number of scale points and non-normality on the generalizability coefficient: A Monte Carlo study. *Applied Measurement in Education, 20*, 357–376.

Thorndike, R. L., & Hagen, E. P. (1993). *Form 5 cognitive abilities test – norms booklet*. Chicago: Riverside Publishing.

Tschannen-Moran, M., & Woolfolk-Hoy, A. (2001) Teacher efficacy: Capturing an elusive concept. *Teaching and Teacher Education, 17*, 783–805.

Vieluf, S., & Klieme, E. (2011). Mathematics teachers' beliefs about the nature of teaching and learning and their classroom teaching practices in cross-cultural comparison. In G. Kaiser & Y. Li (Eds.), *Expertise in mathematics instruction. An international perspective* (pp. 295–325). New York: Springer.

Wentzel, K. R., Battle, A., Russell, S., & Looney, L. (2010). Social supports from teachers and peers as predictors of academic and social motivation. *Contemporary Educational Psychology, 35*, 193–202.