pedocs
Open Access Erziehungswissenschaften

DIPF

Praetorius, Anna-Katharina; Pauli, Christine; Reusser, Kurt; Rakoczy, Katrin; Klieme, Eckhard

One lesson is all you need? Stability of instructional quality across lessons

formal und inhaltlich überarbeitete Version der Originalveröffentlichung in:
formally and content revised edition of the original source in:
Learning and instruction (2014) 31, S. 2-12

Bitte verwenden Sie in der Quellenangabe folgende URN oder DOI /
Please use the following URN or DOI for reference:
urn:nbn:de:0111-pedocs-216602
10.25656/01:21660

https://nbn-resolving.org/urn:nbn:de:0111-pedocs-216602
https://doi.org/10.25656/01:21660

Nutzungsbedingungen

Dieses Dokument steht unter folgender Creative Commons-Lizenz:
http://creativecommons.org/licenses/by-nc-nd/3.0/de/deed - Sie dürfen das Werk bzw. den Inhalt unter folgenden Bedingungen vervielfältigen, verbreiten und öffentlich zugänglich machen: Sie müssen den Namen des Autors/Rechteinhabers in der von ihm festgelegten Weise nennen. Dieses Werk bzw. dieser Inhalt darf nicht für kommerzielle Zwecke verwendet werden und es darf nicht bearbeitet, abgewandelt oder in anderer Weise verändert werden.
Mit der Verwendung dieses Dokuments erkennen Sie die Nutzungsbedingungen an.

Terms of use

This document is published under following Creative Commons-License:
http://creativecommons.org/licenses/by-nc-nd/3.0/de/deed.en - You may copy, distribute and transmit, adapt or exhibit the work in the public as long as you attribute the work in the manner specified by the author or licensor. You are not allowed to make commercial use of the work or its contents. You are not allowed to alter, transform, or change this work in any other way.

By using this particular document, you accept the above-stated conditions of use.

Kontakt / Contact:

peDOCS
DIPF | Leibniz-Institut für Bildungsforschung und Bildungsinformation
Informationszentrum (IZ) Bildung
E-Mail: pedocs@dipf.de
Internet: www.pedocs.de

Mitglied der Leibniz-Gemeinschaft

One lesson is all you need? Stability of instructional quality across lessons

Anna-Katharina Praetorius
Department of Psychology, University of Augsburg

Christine Pauli & Kurt
Institute of Education, University of Zurich

Katrin Rakoczy
German Institute for International Educational Research, Frankfurt

Author Note

Anna-Katharina Praetorius, Department of Psychology, University of Augsburg, Universitaetsstrasse 10, 86159 Augsburg, Germany

Christine Pauli and Kurt Reusser, University of Zurich, Institute of Education, Freiestrasse 36, 8032 Zurich, Switzerland

Katrin Rakoczy, German Institute for International Educational Research, Schlossstrasse 29, 60486 Frankfurt am Main, Germany

Abstract

Observer ratings are often used to measure instructional quality. They are, however, usually based on observations gathered over short periods of time. Few studies have attempted to determine whether these periods are sufficient to provide reliable measures of instructional quality. Using generalizability theory, this study investigates (a) how three dimensions of instructional quality – classroom management, personal learning support, and cognitive activation of students – vary between the lessons of a specific teacher, and (b) how many lessons per teacher are necessary to establish sufficiently reliable measures of the dimensions. Analyses are based on ratings of five lessons for 38 teachers. Classroom management and personal learning support were stable across lessons, whereas cognitive activation showed high variability. Consequently, one lesson per teacher suffices to measure classroom management and personal learning support, whereas nine lessons would be needed for cognitive activation. The importance of advancing our theoretical understanding of cognitive activation is discussed.

Keywords: Instructional quality; Observer ratings; Number of lessons; Generalizability theory; Cognitive activation

1. Introduction

Researchers regularly use teacher reports, student reports, and/or observer reports when measuring dimensions of instructional quality. Observer ratings often are considered the best option (Clare, Valdés, Pascal, & Steinberg, 2001; Helmke, 2009; Petko, Waldis, Pauli, & Reusser, 2003; Pianta & Hamre, 2009) and sometimes are included as a constitutive component of instructional research (e.g., Helmke, 2009; Klieme, 2006). Recently, the Gates Foundation invested U.S.$50 million into research on teacher effectiveness using classroom observation and analysis of videoed lessons as core measurement instruments (Kane, McCaffrey, Miller, & Staiger, 2013); however, there are some drawbacks to using observer ratings to measure dimensions of instructional quality, for example, these ratings usually are based on observations obtained over a very short period of time (Clausen, 2002; Kunter, 2005; Lüdtke, Robitzsch, Trautwein, & Kunter, 2009; Reyes, Brackett, Rivers, White, & Salovay, 2012; Seidel et al., 2006; Waldis, Grob, Pauli, & Reusser, 2010). The question of whether the quality of the observed lessons is sufficiently indicative of the lessons the teachers generally conduct is crucial. Until now, the stability of instructional quality dimensions across lessons rarely has been investigated (see also Brophy, 2006; Calkins, Borich, Pascone, Kluge, & Marston, 1997; Hill, Charalambous, & Kraft, 2012), particularly high-inference ratings (i.e., ratings which require a certain amount of inference beyond the behavior observed). The aim of this study is to shed light on the topic by applying generalizability theory (G theory) (Brennan, 2001a; Shavelson &Webb, 1991). In addition to deepening our understanding of the variation in features of instruction, this research has practical relevance, given the interest in monitoring teacher performance through lesson observation.

After an introduction to the concept as well as to the measurement of instructional quality, when and why short periods of observation can be problematic for measuring specific dimensions of instructional quality will be addressed. Afterward, the results of empirical studies concerning variations in instructional quality dimensions across lessons will be considered. Finally, research questions and hypotheses will be derived.

1.1. Conceptualizing instructional quality: three basic dimensions

Instructional quality has been investigated in diverse research traditions differing in approach, focus, and definition. One of the most influential of these is teacher effectiveness

research (for an overview, see Seidel & Shavelson, 2007) in which several attempts have been made to conceptualize instructional quality. Opinions in the field are beginning to converge on the belief that instructional quality can be described via three basic dimensions (e.g., Baumert et al., 2010; Creemers & Kyriakides, 2008; Klieme, Schümer, & Knoll, 2001; Kunter & Baumert, 2006; Lipowsky et al., 2009; Pianta & Hamre, 2009; Reyes et al., 2012; Tschannen-Moran & Woolfolk Hoy, 2001; Vieluf & Klieme, 2011). Rather than describing surface-level characteristics of instruction, such as social forms, instructional methods, and the use of teaching materials, this model refers to the deep structure of teaching which is assessed through broader ratings conducted by observers, by teachers, or by students. The dimensions identified are classroom management, personal learning support, and cognitive activation. In the CLASS observation system developed by Pianta and Hamre (see, e.g., Pianta & Hamre, 2009), these dimensions have been labeled organizational, emotional, and instructional support. Several studies have demonstrated the predictive validity of these dimensions on student outcomes (e.g., Baumert et al., 2010; Kane & Staiger, 2012; Klieme et al., 2001; Klieme, Pauli, & Reusser, 2009; Kunter et al., 2013; Lipowsky et al., 2009; Reyes et al., 2012).

Since the initial work by Kounin (1970), many studies on teacher effectiveness have focused on the first dimension, classroom management, which deals with providing students with quality learning time by preventing or dealing effectively with disruptions and disciplinary conflicts (for an overview see Kunter, Baumert, & Köller, 2007). The most important aspects of good classroom management have proven to be clearly formulated compulsory rules and routines, efficient organization, and well-structured lessons.

The second dimension, personal learning support, refers to efforts to enhance student motivation to learn by, among others things, creating a positive learning climate. Fostering a positive teacher–student relationship and providing constructive feedback are some of the aspects summarized in this dimension, the importance of which often is based on self-determination theory (Deci & Ryan, 1985).

The third dimension, cognitive activation, focuses on teacher assistance regarding student engagement in higher-level thinking (Klieme et al., 2009; Lipowsky et al., 2009; see also Brophy, 2000; Hiebert & Grouws, 2007; Mayer, 2004; Reusser, 2006), based on the concept of teaching for understanding (Cohen, 1993; Pauli, Reusser, & Grob, 2007). Examples of fostering higher-level thinking include providing challenging tasks in zones of proximal development, activating previous knowledge, building on students' ideas and experiences, and posing stimulating questions.

1.2. Measuring instructional quality

While the general idea of distinguishing three dimensions is supported by several researchers, the operationalization of the dimensions differs considerably between studies, especially for personal learning support and cognitive activation.

Some models assume personal learning support to be comprised mainly of climate variables (e.g., student–teacher relationship) (e.g., Klieme et al., 2001; Pianta & Hamre, 2009), while others (e.g., Baumert et al., 2010) view it as a combination of climate variables and content-related support activities (e.g., adaptive explanations in mathematics). These different assumptions have considerable implications: Whereas personal learning support should be relatively independent of the subject and the content taught in the climate-focused operationalization (see also Klieme et al., 2009), subject and content are both important components of this dimension in a content-focused operationalization (see also Baumert et al., 2010).

Obvious differences regarding operationalization also exist for the third dimension, cognitive activation. For example, Baumert et al. (2010) focused on task quality in measuring cognitive activation: They collected all tests, examinations, homework assignments, and tasks related to two selected topics in mathematics instruction in grade 10 and coded them based on certain criteria (e.g., required level of mathematical argumentation). Following a different line, Lipowsky et al. (2009) examined cognitive activation related to one topic in mathematics instruction (a three-lesson unit focusing on the Pythagorean Theorem) in grades 8 and 9 via external observer ratings. Cognitive activation is conceptualized as pedagogical practices used by teachers to promote student engagement in higher-level thinking (e.g., asking students to explain how they arrived at their answers). What is common to both examples is that measurements were restricted to specific topics, as cognitive activation is tied closely to the content taught and how it is implemented in tasks, materials and discourse (Baumert et al., 2010; Klieme et al., 2009; Lipowsky et al., 2009).

1.3. Short periods of observation: a problem?

Studies using external observer ratings, such as Baumert et al. (2010) and Lipowsky et al. (2009), usually take small samples of tasks or lessons per teacher as indicators of instructional quality due to the high cost of such investigations. Table 1 provides an overview of the most significant recent video studies using observer ratings (see Helmke, 2009; Janík, Seidel, & Najvar, 2009).

One goal of research using videos is to describe general or content-specific dimensions

of instructional quality on an aggregate level (e.g., entire countries). In this case, single lessons conducted by individual teachers are used to estimate the instructional quality dimensions with regard to this aggregate variable. A second goal is to describe the specific quality of a teacher's videoed lessons. In this case, researchers are interested in making specific statements about the lessons being videoed. A third goal is to analyze the

Table 1
Overview of the number of lessons per teacher used in video studies.

Study name	Reference	Subject	Number of Lessons
CES	Anderson and Burns (1989)	Mathematics	6-10
Co^2Ca	Bürgermeister et al. (2011)	Mathematics	2
CPV video study	Janik et al. (2006)	Physics	4-8
QuiP	Neumann et al. (2009)	Physics	2
DESI	Helmke, T. et al. (2008)	English	2
IPN video study	Seidel et al. (2009)	Physics	2
LPS	Clarke et al. (2006)	Mathematics	10
PERLE	Lotz et al. (2013)	Mathematics	2
		German	2
		Art	2
Pythagoras	Klieme et al. (2009)	Mathematics	5
Bern video study	Dalehefte et al. (2009)	Physics	2
SINUS at primary schools	Kobarg et al. (in prep.)		3
TIMSS	Baumert et al. (1997) Stigler et al. (2000)	Mathematics	1(-3[a])
TIMSS 1999 Video Study	Hiebert et al., (2003) /Roth et al. (2006)	Mathematics & Science	1
VERA	Helmke, A. et al. (2008)	German	0-2[b]
		Mathematics	0-2[b]
		Further Subjects	0-1[b]

[a]In the German part of the TIMS study, some of the teachers were videotaped three times (Kunter, 2005).
[b]In this study, the subject as well as the number of lessons per teacher were not set, so not all teachers were filmed in every subject.

instructional behavior of individual teachers so a general profile of competence and instructional efficiency can be formulated. In this case, researchers make statements about the teacher beyond the videoed lessons (e.g., when investigating the effect of instructional quality on student achievement). If researchers are interested in making general statements about the dimensions of instructional efficiency of individual teachers, they act on the following fundamental assumption: The number of lessons videoed in the study in question is sufficient to provide a reliable estimate of instructional quality dimensions (see also Doyle, 1977; Shavelson & Dempsey-Atwood, 1976). If just a few lessons are videoed, this assumption is only justified if the dimensions under investigation vary only to a small degree between

lessons.

Researchers using video studies typically argue that a few lessons are sufficient to measure diverse dimensions of instructional quality at the individual teacher level arguing the following: (a) It is assumed that instructional competence and, thus, related patterns of instructional behavior cannot be modified in the short-term and should be observable in every lesson (e.g., Klieme et al., 2001; Kunter, 2005; Stigler, Gonzales, Kawanaka, Knoll, & Serrano, 1999); (b) Self-reports by teachers support the representativeness of the videoed lessons for instruction in general (e.g., Clausen, 2002; Janík, Miková, Najvar, & Najvarová, 2006; Seidel et al., 2006); and (c) former studies using low-inference ratings (e.g., the use of calculators, Mayer, 1999; see also Seidel & Prenzel, 2006; Seidel et al., 2002) showed sufficient stability in instructional patterns (e.g., Kunter, 2005; Meyer, Seidel, & Prenzel, 2006).

However, some researchers argue that short observation periods are inadequate (e.g., Berliner, 2005; Brophy, 2006; Helmke, 2009; Medley & Mitzel, 1963). Brophy (2006), for example, stated that instead of one or two lessons per teacher, 20–30 lessons would be necessary. The main argument against using short observation periods to draw conclusions about the quality of instruction of individual teachers is that teaching practices are supposed to vary between lessons (e.g., Clarke et al., 2007; Pauli & Reusser, 2011; Staub, 2007; Stigler et al., 1999).

1.4. Empirical investigations concerning the variation of dimensions of instructional quality between lessons

Statements regarding variations in the dimensions of instructional quality between lessons are not based on ample evidence. They either use plausibility arguments, teacher self-reports, or low-inference ratings. Plausibility arguments do not suffice to ensure that instructional quality dimensions vary little between lessons. By the mid-1970s, Shavelson and Dempsey-Atwood (1976) already had raised concern that the generalizability of teaching behavior across lessons is much more an empirical question than a theoretical one. Teacher self-reports as well as low-inference studies provide empirical evidence regarding this question, but they also do not suffice. The usefulness of teacher self-reports is restricted because they are subject to distortion (Porter, 2002). Studies using low-inference ratings do not suffice as indicators of the variability of instructional quality dimensions across lessons measured with high-inference ratings. This is because low-inference ratings cannot adequately assess the characteristics of the deep structure of instruction (e.g., cognitive

activation). Few empirical investigations into the variability of instructional quality dimensions measured with high-inference ratings exist; they either use correlational designs or G theory. The results of these studies are described in the following.

1.4.1. Studies using correlational designs: results and limitations.

One of the first reviews of the variability of instructional quality measures was conducted by Shavelson and Dempsey-Atwood (1976). The authors summarized the results of 11 correlational studies based on $5 \leq n \leq 84$ teachers and $2 \leq t \leq 10$ measurement points, respectively lessons. Shavelson and Dempsey-Atwood (1976) found large variations among the 11 studies: The correlations varied between $-.90 \leq r \leq 0.92$. According to the authors, four of the 16 content clusters investigated were moderately stable (e.g., positive and negative feedback), six were unstable (e.g., teacher questions), and six were inconsistent regarding their stability (e.g., presentation of content). The exact criteria for distinguishing stable and unstable patterns were not mentioned in the review.

More recent studies also found large differences in the stability of instructional quality measures across measurement points. Kunter (2005), for example, focused on how instruction can be arranged to enable positive interest as well as achievement development simultaneously. To show evidence regarding the validity of the observer ratings used, Kunter analyzed the correlations among three randomly chosen consecutive lessons given by 28 teachers in the German longitudinal TIMS video study. The lessons were rated via high-inference ratings for the dimensions active construction, relevance, and autonomy. The stability across lessons ranged between $0.07 \leq r \leq 0.65$ for these dimensions. Rakoczy (2008) investigated the fostering of motivation in classrooms. One of her research questions focused on whether the aspects of instruction students perceive as important for fostering their motivation are observable in a stable way across topics. To answer this question, Rakoczy aggregated five mathematics lessons of 40 teachers into two-lesson units (one unit consisted of two lessons, the other three; some of the data in this study is used in the present investigation). The content of the lesson units was standardized across teachers. The 40 teachers were rated with regard to organizational scope for development, classroom management, appreciation of relationship and feedback, cognitive level, and everyday life relevance. The correlations between the two-lesson units ranged between $0.05 \leq r \leq 0.69$, depending on the dimension investigated.

A large disadvantage in the use of correlations to investigate stability is that they confound stability and other factors, therefore they cannot be interpreted unambiguously in

terms of high or low stability (see also Shavelson & Dempsey-Atwood, 1976). One option to effectively separate stability from further parameters is G theory.

1.4.2. Studies using generalizability theory.

G theory enables multiple sources of error (called facets) to be separated via variance component analysis (Brennan, 2001a; Shavelson & Webb, 1991). To estimate the dependability of the measurement in question, two G coefficients can be used. Both can be interpreted analogous to classical reliability coefficients. For studies focusing on the relative position of variable values (instead of absolute values), the relative G coefficient is suitable (Shavelson & Webb, 1991). Based on the estimated variance components, decision studies (D studies) can be conducted. These studies enable researchers to estimate reliability under multiple measurement conditions (e.g., differing numbers of lessons per teacher/class) and, thus, provide evidence regarding how many conditions of variables (e.g., raters, teachers/classes, items, lessons) would be necessary to obtain sufficient reliability (Brennan, 2001a).

According to Shavelson and Dempsey-Atwood (1976), prior to the mid-1970s, G theory was used to investigate the stability of instructional patterns in only two studies, both of which showed insufficient generalizability. Further studies showed, as well, that the number of raters and measurement points chosen in the respective studies did not suffice. Calkins et al. (1997), for example, focused on the generalizability of teacher behaviors across several classroom observation systems. They examined the lessons of 12 teachers teaching the same lesson unit in social studies at three measurement points using a completely crossed, three-facet design (teachers/classes × measurement points × raters). They deemed a measure sufficiently reliable if a G coefficient of 0.70 was reached using a maximum of eight raters and eight measurement points. Their results showed that 69% of the high-inference items were highly unstable. No further information was provided about the investigated dimensions. Kane and Staiger (2012) investigated the reliability of five existing classroom observation systems. They analyzed a subset of 1333 teachers regarding the variability of instructional quality measures in four to eight randomly chosen lessons per teacher. Only limited information was provided regarding the design used, thus, one only can hypothesize which variance components are included in the reported percentages. The dimensions investigated differed with regard to the amount of variability between lessons, for example, regarding the CLASS instrument, the ratio of stable to unstable variance was 2:1 for the class organization dimension, and 1:1 for the instructional support and emotional support dimensions. Analyzing

up to four lessons per teacher, none of the investigated dimensions reached reliability greater than 0.70. However, Kane and Staiger based their reliability estimates on different raters rating each lesson, thus, rater effects cannot be separated from other variance components in these estimates.

Findings varied considerably among studies investigating exactly how many measurement points are necessary to obtain a sufficiently reliable evaluation of the dimensions under investigation. Hill et al. (2012) identified a comparatively small number. The authors focused on how teacher observation systems should be constructed to enable reliable and valid interpretations and investigated eight teachers with different levels of mathematical pedagogical content knowledge at three randomly chosen measurement points. The design of the study was partially nested: Lessons were nested in teachers/classes and crossed with raters. Hill et al. included three dimensions: richness of mathematics, errors and imprecision, and student participation in meaning-making and reasoning. Between 1% and 40% of the variance had to be attributed to the individual lessons instead of to the teacher/classes, depending on the item. The ratio of stable to unstable variance varied between 0:24 (single item, developing generalizations) and 15:1 (holistic rating of the dimension, richness of the mathematics). Hill et al. concluded that for the three dimensions under investigation, three lessons per teacher rated by two raters were necessary for a reliable estimation of teaching quality. In a study by Newton (2010), the number of measurement points was higher. Newton developed an instrument to evaluate a mathematics reform initiative. Within the study, 32 teachers were assessed at three randomly chosen measurement points by four raters. Newton used a completely crossed design (teachers/classes × measurement points × raters). The variance proportion attributed to measurement point variance varied considerably between the dimensions (0% for content elicitation; 50% for clarity regarding learning content). To reach a G coefficient of 0.80, four raters and six measuring points were necessary for the elementary school teachers/classes, and four raters and five measuring points were needed for the secondary school teachers/classes. Erlich and Shavelson (1978) identified the need for a considerably larger number of measurement points. They investigated the generalizability of teacher behavior measures over measurement facets by investigating 10 teachers in two subjects (reading and mathematics) at three measurement points. The content was not standardized in this study. The design was partially nested: Lessons were nested in teachers/classes and crossed with raters. The authors concluded that most of the high-inference items investigated (e.g., teacher warmth, teacher motivational skills, and need for intervention) would be sufficiently generalizable using four raters and ten

measurement points.

1.5. The Present Study

The investigations described above reveal that there are large differences in the stability of instructional quality dimensions over time; however, most of the studies did not discuss possible reasons for this. The theoretical arguments for and against whether small sample sizes are sufficient to measure instructional quality, do not distinguish between different dimensions as well.

One possible explanation for differences between the dimensions can be derived from the instructional triangle (see e.g., Pauli & Reusser, 2011; Reusser, 2006). According to the triarchic structure of instruction, three main entities can be distinguished in classroom teaching: teachers, students, and content. If these entities change, instruction will change as well. For a certain class taught by a certain teacher, the main entity of change is the content. It seems reasonable to assume that dimensions of instructional quality which are content-dependent will vary to a larger degree than those which are content-independent (see also Rakoczy, 2008). Thus, the problem that small samples of tasks or lessons might not be sufficiently indicative of instructional quality dimensions in general only should apply to content-dependent dimensions.

Applying the instructional triangle to the three basic dimensions of instructional quality – classroom management, personal learning support, and cognitive activation (Klieme et al., 2001; Kunter & Baumert, 2006; Pianta & Hamre, 2009; see also section 1.1) – the following can be expected: Consistent with the often met assumption that classroom management is a stable characteristic of instruction (see e.g., Brophy, 2000) as well as the fact that it is based on general pedagogical knowledge (see e.g., Baumert et al., 2010), classroom management can be assumed to be content-independent and, thus, to vary only slightly among the lessons of a specific teacher (*Hypothesis 1*). In the present investigation, the dimension of personal learning support was operationalized as a climate-focused construct so this dimension is expected to be highly stable too (*Hypothesis 2*). In contrast, cognitive activation should vary to a large degree among lessons depending on the content and the kind of lesson (e.g., introduction vs. consolidation focus; see also Lipowsky et al., 2009; Pauli & Reusser, 2011; *Hypothesis 3*). Regarding the dimension of cognitive activation, the additional assumption was made that ratings of lessons focusing on the same content will be correlated to a higher degree than ratings of lessons which focus on different content (*Hypothesis 4*).

As far back as the 1970s, Rosenshine and Furst (1973) already were critical of the fact

that decisions concerning the number of lessons to be observed in such studies rarely were made on an empirical basis. This has changed little and is problematic for two reasons: First, interpretations based on empirical data may not be trustworthy; and second, video studies are very expensive so researchers should only video as many lessons as is absolutely necessary. Therefore, the following was added as a *research question*: How many lessons per teacher/class are necessary to ascertain sufficiently reliable measures of the three basic dimensions of instructional quality?

2. Method

2.1 Database and sample

2.1.1 Instructional lessons

The videos that served as targets in the present study were collected in the German–Swiss video study regarding the quality of instruction, learning, and understanding of mathematics (see Klieme et al., 2009) which was carried out in 40 grade eight and nine classrooms at intermediate and academic-track schools in Germany and Switzerland. Participation was voluntary. Over the course of 2002 and 2003, every teacher/class was videoed five times during mathematics lessons. The five lessons (45 min. each) were to occur in conjunction with two specific units: A three-lesson unit on the Pythagorean Theorem, and a two-lesson unit on word problems (for more detailed information concerning these lesson units see Rakoczy, 2008). Apart from the default content, teachers were directed to conduct routine instruction. The analyses for the present study considered only those teachers (n = 38) for whom ratings of all five lessons were available.

The overall time interval was not standardized between the two lesson units, thus, they varied considerably (M = 16 weeks, SD = 6 weeks). The time intervals needed to complete the lesson units were small because consecutive lessons were videoed: For the word problems unit, both lessons were videoed within one day in 92% of the cases and within two days in the remaining cases; and for the Pythagorean Theorem unit, 79% of the three lessons were videoed within two days, 13% within three days and 8% within four days.

2.1.2 Raters

The four teachers who conducted the high-inference ratings all previously had completed additional university-level psychology studies. They were given 40 h of additional training in June 2004 after working through the rating manual. At the beginning, the raters were introduced to the project, to high-inference ratings and to the unit of analysis (i.e.,

complete lessons). Subsequently, they were given theoretical input as well as examples concerning the rating dimensions and the rating instrument to attain a joint theoretical understanding of the dimensions they would be rating. Any comprehension questions were resolved in this context as well. The raters then watched some videoed lessons, rated them, and afterwards discussed their ratings. As the ratings following the training did not attain sufficient reliability for the lesson unit on the Pythagorean Theorem, a two-day follow-up training session was conducted.

Three of the four raters executed ratings for the lesson unit on the Pythagorean Theorem; two rated the word problems unit. Thus, one rater rated both lesson units, the other raters only rated one of the two lesson units.

2.1.3 Rating instrument

The rating instrument measured the three basic dimensions of instructional quality: classroom management, personal learning support, and cognitive activation (see also sections 1.1 and 1.2). To assess these ratings, the main aspects of each dimension were translated into one item. In total, eight high-inference items were used. Each item described its basic idea as well as some observable behavioral patterns to guide the raters in their decision as to whether, and to what degree, a certain feature occurred in a certain lesson. The ratings comprised an overall impression concerning these indicators. An overview of the items can be found in Table 2 (see also Rakoczy & Pauli, 2006). All items used a 4-point Likert-type scale ranging from 1 to 4 (for 5 items: 1 = strongly disagree; 4 = strongly agree; for 3 items: item-specific labels, see Table 2).

Rakoczy (2008) could prove the factorial validity of the three dimensional instrument and Lipowsky et al. (2009) showed that it is possible to predict the development of student achievement with the three dimensions of the rating instrument used in the Pythagoras study.

2.2 Procedure for the observational ratings

The rating sessions were conducted between August 2004 and January 2005. Each rater assessed all lessons independently. The raters were free to organize their time for the ratings on their own. On average, they took about six weeks to finish. The order of the lessons was fixed; the single lessons given by a teacher relating to one unit were rated consecutively. The procedure was the same for each video: The raters viewed the entire lesson and rated it immediately after. They were free to take notes and to view segments of the video again if they were unsure. It was not possible to skip items; thus, there was no missing data.

Table 2
Number of items and item content concerning the three basic dimensions of instructional quality in the Pythagoras project.

Dimension	Number of Items	Item Content	Example Item Indicator	Answer Category Labeling
Classroom management	2	Discipline problems/disruptions	The teacher repeatedly has to remind students to be silent or request them to work.	1= No disruptions; 4 = many discipline problems
		Classroom management	No time is.	1 = Strongly disagree; 4 = strongly agree
Personal learning support	3	Student approval	Students speak to their teacher politely.	1 = No one shows approval; 4 = all show approval
		Factual, constructive feedback	Feedback is formulated benevolently, even in response to errors.	1 = No feedback is factual and constructive; 4 = all feedback is factual and constructive
		Learning community	Students pay attention to each other and not only to the teacher.	1 = Strongly disagree; 4 = strongly agree
Cognitive activation	3	Exploration of the students' ways of thinking	The teacher tries to understand the students' ways of thinking by asking how they came to certain answers.	1 = Strongly disagree; 4 = strongly agree
		Challenging activities at a high cognitive level	The teacher poses open questions which stimulate contemplation.	1 = Strongly disagree; 4 = strongly agree
		Receptive understanding of learning through the teacher (recoded)	The teachers' questions are structured in small steps.	1 = Strongly disagree; 4 = strongly agree

2.3 Overview of the generalizability study designs

The ratings used in the present study were analyzed using G theory. Several lessons (*l*) per teacher/class (*t*) were observed and rated. Lessons were nested within teachers/classes because the lessons were not videoed at the same time (between teachers), and the time intervals between lessons (within teachers) were in some cases smaller than the intervals for a specific lesson (between teachers) (Shavelson, Webb, & Burstein, 1986). Each lesson was assessed by several raters (*r*). The analyses were conducted on an item level as previous research has shown that analyzing ratings of instructional quality on an item level instead of a scale level is useful for detecting problematic interaction effects (see Praetorius, Lenske, & Helmke, 2012). Items (*i*), therefore, were added as a facet.

The present study had a two-facet, partly nested random effects design ((*l*:*t*) × *r* × *i* design) which allowed the separating out of variance in the rating data resulting from 11 sources: (a) the teachers/classes (σ^2_t); (b) the lessons, which were nested within the teachers/classes ($\sigma^2_{l:t}$); (c) the raters (σ^2_r); (d) the items (σ^2_i); (e) the interaction between teachers/classes and raters (σ^2_{tr}); (f) the interaction between teachers/classes and items (σ^2_{ti}); (g) the interaction between raters and items (σ^2_{ri}); (h) the interaction between raters and lessons, which were nested within teachers/classes ($\sigma^2_{(l:t)r}$); (i) the interaction between items and lessons, which were nested within teachers/classes ($\sigma^2_{(l:t)i}$); (j) the interaction between teachers/classes, raters and items (σ^2_{tri}); and (k) the interaction between raters and items and lessons, which were nested within teachers/classes ($\sigma^2_{(l:t)ri,e}$). The latter interaction was confounded with an unspecific error component as it was the highest order interaction.

The subdividing method described by Chiu and Wolfe (2002) was used because the raters were not the same across both lesson units, thus, the lesson units were analyzed separately and subsequently the variance components were averaged. One rater assessed both lesson units and is part of both data sets.

In the following, G coefficients are reported for relative decisions only (G coefficient ρ^2) because the focus of the present study mainly is relevant for research purposes where they are of major interest. For all analyses, the interaction between teachers and items were 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 analyses regarding the D studies were made using the GENOVA software (Crick & Brennan, 1983). The implemented estimators are the ANOVA procedure for GENOVA and the analogous ANOVA procedure for urGENOVA. A large advantage of these estimators is that normality assumptions are not required (Brennan, 2001a; see also the simulation study of Shumate, Surles, Johnson, & Penny, 2007).

3 Results

3.1 Descriptive statistics

Table 3 contains the mean scores and standard deviations for all three instructional quality dimensions and the internal consistencies, averaged over all lessons as well as separately for both lesson units. The means for classroom management were very high, in contrast to cognitive activation where they were rather low. The means did not vary greatly in the two lesson units on any of the dimensions. For the Pythagorean Theorem lesson unit, the internal consistencies were admissible; this was not true for the word problems lesson unit ($\alpha_{min} = 0.62$). This is not problematic for the present study as the generalizability analyses conducted take these restrictions into account explicitly by estimating diverse item effects. For classroom management, the internal consistencies partly take a value of 1.00. Rakoczy (2008) explains this fact through the construction of mutually dependent subscales.

Table 3
Descriptive statistics for the basic dimensions scales, aggregated over lesson units, as well as separately for each lesson unit.

Dimension	*M*	*SD*	α
Classroom management	3.64	0.54	1.00
Pythagorean Theorem lesson unit	3.64	0.54	1.00
Word problems lesson unit	3.65	0.64	1.00
Personal learning support	2.60	0.43	0.83
Pythagorean Theorem lesson unit	2.58	0.46	0.82
Word problems lesson unit	2.66	0.48	0.62
Cognitive activation	1.93	0.40	0.74
Pythagorean Theorem lesson unit	1.96	0.47	0.81
Word problems lesson unit	1.86	0.60	0.67

Table 4
G analyses for the three basic dimensions of instructional quality.

	Classroom Management			Personal Learning Support			Cognitive Activation		
Variance Component	VC	%	*SD* VC	VC	%	*SD* VC	VC	%	*SD* VC
Stable components									
t	0.29	63	0.10	0.14	21	0.00	0.13	12	0.02
ti	0	0	0.00	0.11	16	0.07	0.02[b]	2	0.03
Lesson-specific components									
l:t	0.06	13	0.01	0.01	2	0.01	0.17	16	0.11
(l:t)i	0[a]	0	0.00	0.02	3	0.01	0.33	30	0.46
Rater bias components									
r	0.01	2	0.01	0.02	3	0.02	0.01[a]	1	0.01
tr	0.04	8	0.04	0.06	9	0.06	0.05	4	0.02
(l:t)r	0.05	11	0.04	0[a]	0	0.00	0.03	3	0.03
ir	0	0	0.00	0.02	3	0.02	0.06	5	0.07
tir	0	1	0.00	0.08	12	0.02	0.06	5	0.07
Item-specific component									
i	0[a]	0	0.00	0.08	12	0.07	0[a]	0	0.00
Error									
(l:t)ir,e	0.01	2	0.01	0.12	18	0.11	0.24	22	0.17
Total variance	0.47			0.66			1.10		
$E\rho^2$	0.92			0.94			0.63		
Φ	0.91			0.83			0.60		

3. *Note*. *t* = teacher; *l* = lesson; *r* = rater; *i* = item; *e* = residual; VC = absolute variance component; % = relative variance component; *SD* VC = standard deviation of the variance components between both lesson units; $E\rho^2$ = relative G coefficient; Φ = absolute G coefficient. The interaction $t \times i$ was set as universe score-variance for computing the G coefficients.
[a]A small negative variance component was estimated for one of the lesson units and thus set to zero.
4. [b]A high negative variance component was estimated for the word problems lesson unit. For the analyses, this estimate was set to zero.

3.2. The variation of instructional quality dimensions between lessons

For classroom management, a large amount of the universe score variance in both of the lesson units (Pythagorean Theorem and word problems) was stable across the lessons (see also Table 4). The stable amount of variance (*t, ti*) was very high (63%) compared to the lesson-specific amount (*l:t*, (*l:t*)*i*) (13%). Thus, Hypothesis 1 can be seen as confirmed. Variance concerning rater bias (*r*, *tr*, (*l:t)r*, *ir*, *tir*) comprised 22% of the entire variance. The amount of unexplained variance ((*l:t)ir,e*) was very small. Thus, for classroom management, nearly all variance could be explained through the present facets.

The high stability of classroom management across lessons primarily could be traced back to the 22 teachers/classes with a very high total mean for classroom management. Of these teachers/classes, 14 showed no variability between lessons and eight showed a very small variation. On the contrary, for the seven teachers/classes with the lowest total means for classroom management, a high intra-individual variation was found between lessons.

For personal learning support, a high amount of variance also was stable across lessons (see Table 4). The stable amount of variance accounted for 37%, whereas the lesson-specific amount accounted for 5%. Thus, Hypothesis 2 can be maintained as well. Variance concerning rater bias comprised 27% of the entire variance. The amount of unexplained variance accounted for 18% of the entire data set.

For cognitive activation, only a small amount of the universe score variance was stable over lessons (see Table 4). The stable amount of variance was 14%, the lesson-specific proportion was 46%. Thus, Hypothesis 3 also can be maintained. Rater bias accounted for 18% of the variance for the entire data set. The amount of unexplained variance was 22%.

The results concerning the stability of cognitive activation, however, were subject to limited interpretation: The interaction between teachers and items revealed a highly negative variance estimate for the word problems lesson unit (see also Table 4). This suggested a misspecification in the underlying model (for a discussion of possible reasons behind this problem see 4.4). Seeking to clarify the degree to which this misspecification distorted the results, the word problems lesson unit was also analyzed using a different design in which teachers/classes, points of measurement, raters, and items were crossed ($t \times m \times r \times i$ design). This design presupposed that points of measurement are interchangeable (i.e., points of measurement are the same for all teachers/classes; or the time intervals of one point of measurement between teachers/classes is smaller than the time intervals between points of measurement within teachers/classes). This presupposition was not provided for the study at

hand; thus, this design also was not entirely appropriate for the data at hand. Nevertheless, a comparison of the two designs is helpful in estimating the proportion of distortion which can be traced to the high negative variance in the original design. The analyses of the $t \times m \times r \times i$ design revealed the stable proportion of variance (*t*, *ti*) to be 11% and the lesson-specific proportion (*m*, *mi*, *tmi*) to be 76%. These figures were almost identical to those for the partially nested design (12% compared to 77%, see also Table 4).

For cognitive activation, correlations among the five lessons were computed (see Table 5). Supporting Hypothesis 4, the ratings of lessons which were similar regarding the content (i.e., those within one lesson unit) correlated to a higher degree than ratings of lessons focusing on different content. These differences were significant (proof of significance of correlations in dependent samples; see Williams, 1959) with one exception: The correlations between the ratings of the first lesson of the word problems unit and the ratings of the Pythagorean Theorem lessons did not differ significantly.

Table 5
Bivariate correlations of the amount of cognitive activation between lessons.

	Pythagoras L2	Pythagoras L3	Word p. L1	Word p. L2
Pythagoras L1	0.52**	0.47**	0.36*	0.13
Pythagoras L2		0.44**	0.28	0.03
Pythagoras L3			0.29	-0.07
Word p. L1				0.40*

Note. L = lesson; p. = problems.
* p < .05. ** p < .01.

3.3 The number of lessons necessary to measure dimensions of instructional quality

To determine how many lessons per teacher are required to measure instructional quality dimensions with sufficient reliability (*research question 1*), D analyses were conducted with 1–50 lessons per teacher/class. The number of raters and items was fixed to the actual number in the study at hand. Fig. 1 illustrates the results for the D studies for all three basic dimensions: To obtain a G coefficient greater than 0.70 for classroom management and personal learning support, only one lesson per teacher/class was needed. In contrast, nine lessons were needed for cognitive activation.

One must keep in mind that the results regarding the word problems lesson unit are subject to limited interpretation for cognitive activation because a high negative-variance estimate was found for the interaction between teachers/classes and items. However, since the alternative $t \times m \times r \times i$ design revealed nearly identical results, one can reasonably expect that the associated distortions were not large.

4. Discussion

As a rule, observer ratings concerning instructional quality are based on one to a few lessons per teacher/class. Few previous investigations have focused on the question of whether a small number of lessons per teacher/class are sufficient to make reliable statements about specific dimensions of instructional quality. The present study has shed some light on this topic by investigating variability in measures of instructional quality across several lessons of an individual teacher/class, as well as by investigating how many lessons per teacher/class are needed to measure dimensions of the instructional quality of teachers/classes reliably.

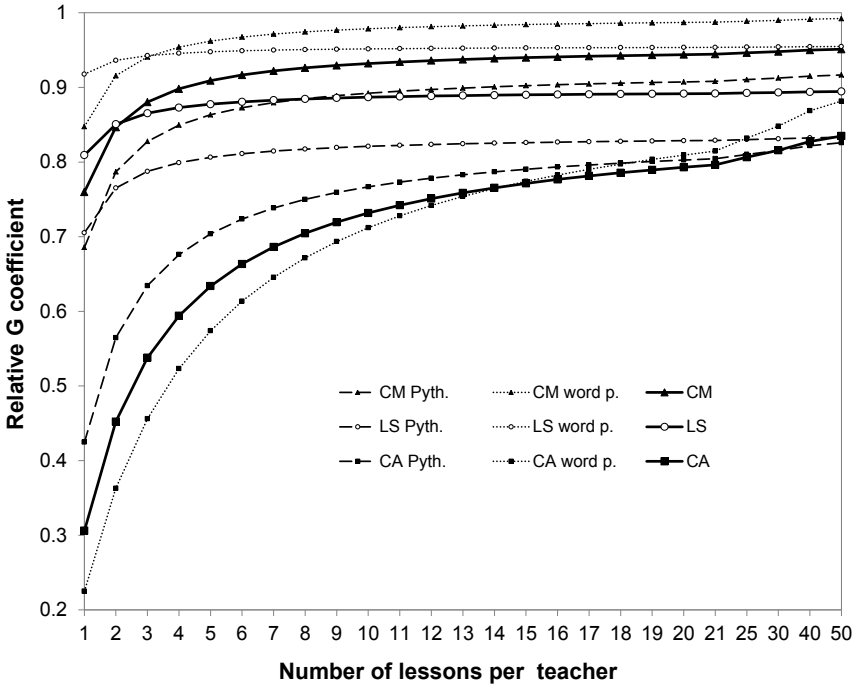

Fig. 1. Relative G coefficients for D studies, with 1–20 lessons per teacher, for all three basic dimensions. CM = classroom management (total data set); LS = personal learning support (total data set); CA = cognitive activation (total data set); CM/LS/CA Pyth. = classroom management/personal learning support/cognitive activation for the Pythagorean Theorem data set; CM/LS/CA word p. = classroom management/personal learning support/cognitive activation for the word problem data set.

4.1 The number of lessons necessary to measure dimensions of instructional quality reliably

Some authors point to the fact that the individual lessons of a teacher/class can differ in quality, therefore, when measuring instructional quality for a teacher/class, it may be inadequate to consider only one or a few lessons (e.g., Brophy, 2006; Calkins et al., 1997; Clarke et al., 2007; Erlich & Shavelson, 1978; Hill et al., 2012; Shavelson et al., 1986; Staub, 2007). In the present study, these doubts were tested empirically (*research question 1*) in

regards to the three basic dimensions of instructional quality in the sense of Klieme et al. (2009). Actually, it was found that the critical assumptions mentioned above are pertinent, but only for one of the three dimensions under consideration. For cognitive activation, at least nine lessons per teacher/class are needed to establish a stable estimate of instructional quality with a reliability level of .70. In contrast, for classroom management and personal learning support, a mere one lesson per teacher/class is sufficient to reach a reliability level of .70.

At least two things can be learned from these results: First, as instructional variables show considerable differences regarding their temporal stability (see also Calkins et al., 1997; Erlich & Shavelson, 1978; Hill et al., 2012; Kunter, 2005; Rakoczy, 2008), the stability of instructional quality dimensions has to be discussed with respect to each dimension. Second, the comparatively large number of lessons necessary to measure cognitive activation emphasizes the need to focus more strongly on the unit of analysis in video studies (see also Staub, 2007). If too few lessons per teacher are included in a study, actual teacher practice cannot be depicted adequately (see also Kane & Staiger, 2012; Lakes & Hoyt, 2008; see 4.2.2. for a challenge concerning this point). To avoid such misleading conclusions regarding teacher practice, researchers need to include the key conditions about which they want to generalize in their studies (Calkins et al., 1997; Lakes & Hoyt, 2008). If several lessons per teacher cannot be measured, it seems necessary either to restrict investigations to variables which are stable enough to be measured with one single lesson per teacher or to avoid generalizations beyond the processes and outcomes of short teaching units.

4.2 The variation in instructional quality dimensions between lessons: explanation and consequences

Confirming our hypotheses (see Hypotheses 1 to 3), the content-independent dimensions of classroom management and personal learning support varied only to a small degree between lessons, whereas the content-dependent dimension cognitive activation showed a large variability between lessons.

4.2.1 Measuring content-dependent dimensions of instructional quality

In agreement with the results of Kane and Staiger (2012), the variability between lessons for the content-dependent dimension of cognitive activation was very high. The ratio of unstable to stable variance was 4:1. The differing correlations between the amount of cognitive activation in the single lessons (see Hypothesis 4) indicates that the amount of cognitive activation implemented is dependent on the topic at hand (see the significant differences in the correlations between Pythagoras and word problem lessons) as well as other

situational characteristics (see the significant differences in the correlations between the Pythagoras lessons and the first vs. second word problem lessons). The high variability between lessons of a specific teacher/class indicates that a problem exists in measuring cognitive activation; however, the variability itself does not tell us what causes this problem. At least three reasons are conceivable, the first of which is measurement error (see also Kane & Staiger, 2012). It can be assumed that cognitive activation is observable to a different degree depending, among other things, on the content taught, the social form, and the stage in the instructional unit (e.g., introduction to a new topic, or practice phase) (Pauli & Reusser, 2011). In turn, these differences in observability lead to a decrease in data reliability. If the reason for the instability across lessons is measurement error, an increase in the number of lessons observed per teacher solves the underlying problem as it increases reliability.

The second possible cause is the limited suitability of the construct of cognitive activation for instruction in general: In a lesson in which new content is introduced, ideally all students are engaged in higher-level thinking through the teacher's stimulating questions, challenging tasks, etc. But what about a practice lesson in which skill automatization is of foremost interest? One could argue that cognitive activation is an important aspect of instruction – but not in all kinds of lessons. Its variability between lessons then would be neither surprising nor objectionable, but simply a hint that cognitive activation is not a suitable construct for all kinds of lessons and, thus, should be restricted to introduction lessons and the like.

The third possible cause is a mismatch between cognitive activation and its measurement. Based on the theoretical foundation of cognitive activation – teaching for understanding (see e.g. Pauli et al., 2007) – one would attempt to adjust all kinds of instruction in accord with this construct. For example, when new content is introduced, students may work on inquiry-type tasks which can be more or less demanding, or the teacher may start by introducing somewhat complex concepts in a lecture format; similarly, practice sessions may involve tasks requiring different levels of transfer. Therefore, a practice lesson with far-reaching transfer may include more cognitive activation than an introductory lesson based on shallow inquiry. Nevertheless, it is obvious that cognitive activation in an introduction lesson is not exactly the same as cognitive activation in a practice lesson. Measuring only one specific aspect of cognitive activation may be sufficient to predict student learning within a single lesson or a short introductory unit, as in the case of the Pythagoras project (Lipowsky et al., 2009); however, if cognitive activation is to be used as a predictor of student learning in a broader sense or even as an indicator of teacher effectiveness that

generalizes across classrooms and contents, its operationalization should be revisited. The existing instruments are not suited in this case as they only focus on small aspects of instruction (e.g., the tasks in the instrument of Baumert et al., 2010; the observer ratings in the instrument of Lipowsky et al., 2009, with its focus on aspects of introductory lessons). In all likelihood, a combination of measurement methods will be necessary to capture the entire construct in its complexity. Some aspects and forms of cognitive activation, for example, are rarely observable: Finding solutions to certain tasks can cause a high level of cognitive activation, although this may not be apparent directly to an observer.

The immediate consequence for investigations using existing measures of cognitive activation either is to state clearly in future publications that only a part of cognitive activation is captured by the respective measure, or to restrict investigations concerning cognitive activation to introductory lessons. In the long term it is of paramount importance to advance our theoretical understanding of cognitive activation (see also Brophy, 2006; Pauli & Reusser, 2011). According to the concept of teaching for understanding, cognitive activation should be conceptualized as an important aspect of every lesson, independent of the content taught, the social form, or the stage in the instructional unit. This indicates the importance of developing a comprehensive theoretical understanding of cognitive activation and criteria to measure it which are suitable to all lessons.

4.2.2 Measuring content-independent dimensions of instructional quality

The stability of the content-independent dimensions of classroom management and personal learning support was very high: The ratio of unstable variance to stable variance was 1:4 and 1:8, respectively. Therefore, according to the present study one would conclude that aspects of classroom management and climate can be observed accurately within short time periods. The high stability of classroom management is in line with findings regarding the on-task behavior of primary school children in a study conducted by Renkl and Helmke (1992); however, Kane and Staiger (2012) found classroom management to be less stable and personal learning support considerably less stable than the present investigation found. These inconsistencies across the studies point to the importance of comparing different instruments to be used in future studies to identify the aspects of the respective operationalization which lead to differences in time stability.

Besides the high temporal stability of the two content-independent dimensions in the present study, some further aspects were striking. Regarding the variance components for personal learning support, a large amount of universe score variance is not due to the main

effect of the teachers/classes but rather to an interaction between teachers/classes and items. This interaction means that teachers/classes either differ with regard to the items that show low or high item difficulty in their lessons, or in their ranges of item difficulty. These differences are evident consistently across all of the lessons observed, thus, teachers not only differ with respect to how good their personal learning support is, but also in the specific aspects of this support in which they are particularly strong or weak. Such interactions between teachers/classes and items challenge attempts to develop hierarchical models of levels of instructional quality based on IRT approaches (see e.g., Pietsch, 2010). If items change their order considerably between teachers/classes, a fixed assignment of items to specific levels is only of limited value.

According to the results of the present study, classroom management is a stable characteristic of instruction. However, there were two groups of teachers: One group was characterized by high mean classroom management and little, if any, variation between lessons; the other group showed high variability in classroom management across lessons. Two possible causes for the high variability among this segment of teachers/classes have been expanded upon. First, situational variables (e.g., the social form applied in lessons or the situation-specific students' behaviors) do have a larger influence on classroom management than often is assumed. If these situational variables fluctuate to a large degree between lessons, classroom management measures also differ. The second reason is methodological in nature. It is possible that the low or missing variability concerning classroom management observed for many teachers/classes is caused by the fact that the items used cannot adequately depict existing differences between lessons, so part of the distribution was not depicted, or its depiction was cut off (see also Praetorius et al., 2012). It seems worthwhile to use future studies to investigate whether one of these reasons can be proven.

4.3 Limitations and further directions

Crucial to the relevance of the study at hand is the question of whether the items used and the results are generalizable across the sample of the Pythagoras study (consisting of Swiss and German teachers teaching in intermediate and academic track schools). The fact that the results of the present study and those of the study of Kane and Staiger (2012) are not completely in agreement indicates that comparative studies using several samples as well as several instruments are required. This would make it possible to gain insight into how sample and instrument specificities influence results.

The importance of future studies replicating the results of high instability for the

cognitive activation dimension is particularly compelling as interpretations of the analyses in question are rather limited. A high negative variance estimate occurred in the interaction between teachers/classes and items in the word problems lesson unit suggesting an easily explained misspecification in the underlying model (Shavelson & Webb, 1991). For the word problems lesson unit, unstable variance is very high (77%) and stable variance is very low (12%); therefore, the model assumption of nesting lessons in teachers/classes is not correct for the cognitive activation dimension. In addition, an alternative model was specified in which teachers/classes and lessons (as occasions) were specified as crossed facets and as such, independent. The assumptions underlying this alternative model also were not correct for the present study`s data, as the lessons were videoed sequentially. The analyses still were conducted so the results of both models could be compared when estimating the extent to which the results were biased through misspecification. As the results of both models were highly similar, the assumption was made that the results could be interpreted with regard to the content; however, to ensure the results were not distorted, the stability of the cognitive activation dimension should be investigated once again using another sample. If one assumes the variability again will be high, the specification of a nested model (lessons in teachers/classes) would not be suitable. Ideally, the individual lessons should be videoed simultaneously, between teachers/classes, to meet the requirements of a completely crossed model – although this is a huge logistical challenge.

Also striking are the extensive differences regarding the variance estimates of both lesson units. For example, in almost all cases rater bias was considerably higher for the Pythagorean Theorem lesson unit than for the word problems lesson unit. There are at least three possible reasons for these differences. First, that G analyses are often problematic regarding the stability of variance component estimates is especially true for complex designs with two or more facets (Brennan, 2001a; Shavelson & Webb, 1991; Smith, 1978); thus, the differences could simply have a methodological basis. Second, also methodological in nature, is that the raters between the two lesson units were not the same and may well have differed in rating quality. Third, the word problem lessons may have been easier to judge within and between teachers/classes because the implementation of word problems, for example, regarding the social form of instruction, was standardized to a higher degree than the Pythagoras lessons. With regard to further investigations, it would be worthwhile to first test and possibly rule out the methodological explanations and then to focus on the influence of different kinds of lessons and further situational factors on rating quality (see also Kennedy, 2010; Malmberg, Hagger, Burn, Mutton, & Colls, 2010; Shavelson & Dempsey-Atwood,

1976).

Variance component estimates are subject to sampling variability (Brennan, 2001). To obtain an impression of how credible the variance component estimates are, standard errors respectively confidence intervals should be reported (see Hoyt & Melby, 1999); however, for unbalanced designs such as the one used in the present study, the computation of standard errors is not implemented in existing G theory software. To gain insight on the credibility of the results of the present study, a replication study seems of paramount importance.

4.4 Conclusions

In their review, Shavelson and Dempsey-Atwood (1976) stated: "Pessimistically, we entertained the possibility that generalizability may be extremely limited in an educational context" (p. 609). The authors noted that at the time of their review, the state of research only allowed empirical examination of their statement to a very limited degree and that further research was necessary. The results of the study at hand point out that the pessimistic view of Shavelson and Dempsey-Atwood (1976) can be put in perspective: The two content-independent basic dimensions of instructional quality investigated showed high stability across lessons, however, the content-dependent third dimension revealed very low stability. Further research is necessary to identify the factors behind this high instability and to revise existing measures of cognitive activation based on this information. This will not be an easy task and will require much effort – but should be worth it. After all, the goal is to measure a construct underlying consistent high quality instruction and not a single teaching performance (Berliner, 2005).

References

Anderson, L. W., & Burns, R. B. (1989). *Research in classrooms: The study of teachers, teaching and instruction*. Oxford: Pergamon Press.

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,* 133–180. http://dx.doi.org/10.3102/0002831209345157.

Baumert, J., Lehmann, R., Lehrke, M., Schmitz, B., Clausen, M., Hosenfeld I., et al. (1997). *TIMSS - Mathematisch-naturwissenschaftlicher Unterricht im internationalen Vergleich: Deskriptive Befunde [TIMSS – An international comparison of lessons in mathematics and the natural sciences: Descriptive results]*. Opladen, Germany: Leske & Budrich.

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

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

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

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

Brophy, J. (2006). Observational research on generic aspects of classroom teaching. In P. A. Alexander, & P. H. Winne (Eds.), *Handbook of educational psychology* (pp. 755–780). Mahwah, NJ: Erlbaum.

Bürgermeister, A., Klimczak, M., Klieme, E., Rakoczy, K., Blum, W., Leiß, D., et al. (2011). Leistungsbeurteilung im Mathematikunterricht: Eine Darstellung des Projekts „Nutzung und Auswirkungen der Kompetenzmessung in mathematischen Lehr-Lernprozessen" [Performance assessments in mathematics lessons: A presentation of the project "Conditions and consequences of classroom assessment"]. *Schulpädagogik – heute*, *2*, 1–18.

Calkins, D., Borich, G. D., Pascone, M., Kluge, S., & Marston, P. T. (1997). Generalizability of teacher behaviors across classroom observation systems. *Journal of Classroom Interaction*, *13*, 9–22.

Chiu, C. W. T., & Wolfe, E. W. (2002). A method for analyzing sparse data matrices in the generalizability theory framework. *Applied Psychological Measurement*, *26*, 321–338. http://dx.doi.org/10.1177/0146621602026003006.

Clare, L., Valdés, R., Pascal, J., & Steinberg, J. (2001). *Teachers' assignments as indicators of instructional quality in elementary schools* (CSE Technical Report No. 545). Los

Angeles, CA: Center for Research on Evaluation, Standards, and Student Testing, University of California.

Clarke, D., Keitel, C., & Shimizu, Y. (2006). *Mathematics classrooms in twelve countries – The insider's perspective*. Rotterdam, Netherlands: Sense Publishers.

Clarke, D., Mesiti, C., O'Keefe, C., Xu, L. H., Jablonka, E., Mok, I. A. C., et al. (2007). Addressing the challenge of legitimate international comparisons of classroom practice. *International Journal of Educational Research, 46*, 28–293. http://dx.doi.org/10.1016/j.ijer.2007.10.009

Clausen, M. (2002). *Qualität von Unterricht – Eine Frage der Perspektive? [Quality of instruction – A question of perspective?]* Münster, Germany: Waxmann.

Cohen, D. K. (1993). *Teaching for Understanding: Challenges for Policy and Practice*. San Francisco: Jossey-Bass.

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

Crick, J. E., & Brennan, R. L. (1983). *Manual for GENOVA: A generalized analysis of variance system* (American College Testing Technical Bulletin No. 43). Iowa City, IA: ACT, Inc.

Dalehefte, I. M., Rimmele, R., Prenzel, M., Seidel, T., Labudde P., & Herweg, C. (2009). Observing instruction "next-door": a video study about science teaching and learning in Germany and Switzerland. In T. Janik, & T. Seidel (Eds.), *The power of video studies in investigating teaching and learning in the classroom* (pp. 83–99). Münster, Germany: Waxmann.

Deci, E. L., & Ryan, R. M. (1985). *Intrinsic motivation and self-determination in human behavior*. New York: Plenum.

Doyle, W. (1977). Paradigms for research on teacher effectiveness. *Review of Research in Education*, *5*, 163–198. http://dx.doi.org/10.3102/0091732X005001163.

Erlich, O., & Shavelson, R. J. (1978). The search for correlations between measures of teacher behavior and student achievement: measurement problem, conceptualization problem, or both? *Journal of Educational Measurement*, *15*, 77–89. http://dx.doi.org/10.1111/j.1745-3984.1978.tb00059.x.

Helmke, A. (2009). *Unterrichtsqualität und Lehrerprofessionalität: Diagnose, evaluation und Verbesserung des Unterrichts [Instructional quality and teacher professionalism. Diagnosis, evaluation and improvement of instruction].* Seelze, Germany: Klett-Kallmeyer.

Helmke, A., Helmke, T., Heyne, N., Hosenfeld, A., Hosenfeld, I., Schrader, F.-W., et al. (2008). Zeitnutzung im Grundschulunterricht. Ergebnisse der Unterrichtsstudie „VERA – Gute Unterrichtspraxis“ [Use of instructional time in elementary classrooms. Results of the assessment of lessons “VERA – Good instructional practice”]. *Zeitschrift für Grundschulforschung*, *1*, 23–36.

Helmke, T., Helmke, A., Schrader, F.-W., Wagner, W., Nold, G., & Schröder, K. (2008). Die Videostudie des Englischunterrichts [The video study of English lessons]. In E. Klieme, et al. (Eds.), *Unterricht und Kompetenzerwerb in Deutsch und Englisch. Ergebnisse der DESI-Studie* (pp. 345–363). Weinheim: Beltz.

Hiebert, J., Gallimore, R., Garnier, H., Givvin, K. B., Hollingsworth, H., & Jacobs, J., et al. (2003). *Teaching mathematics in seven countries. Results from the TIMSS 1999 video study*. Washington, DC: U.S. Department of Education, National Center for Education Studies.

Hiebert, J., & Grouws, D. A. (2007). The effects of classroom mathematics teaching on students’ learning. In F. K. Lester (Ed.), *Second handbook of research on mathematics teaching and learning* (pp. 374–404). Charlotte, NC: Information Age.

Hill, H. C., Charalambous, C. Y., & Kraft, M. A. (2012). When rater reliability is not enough: teacher observation systems and a case for the generalizability study. *Educational Researcher*, *41*, 56–64. http://dx.doi.org/10.3102/0013189X12437203.

Janík, T., Miková, M., Najvar P., & Najvarová, V. (2006). Unterrichtsformen und -phasen im tschechischen Physikunterricht: design und Ergebnisse der CPV Videostudie Physik [Teaching forms and teaching phases in Czech physics lessons: Design and results of the CPV video study in physics]. *Zeitschrift für Didaktik der Naturwissenschaften*, *12*, 219–238.

Janík, T., Seidel, T., & Najvar, P. (Eds.). (2009). *The power of video studies in investigating teaching and learning in the classroom*. Münster, Germany: Waxmann.

Kane, T. J., McCaffrey, D. F., Miller, T., & Staiger, D. O. (2013). *Have we identified effective teachers? Validating measures of effective teaching using random assignment* (Research paper prepared for the Bill and Melinda Gates Foundation). Retrieved from MET project website: http://metproject.org/downloads/MET_Validating_Using_Random_Assignment_Research_Paper.pdf

Kane, T. J., & Staiger, D. O. (2012). *Gathering feedback for teachers: Combining high-quality observations with student surveys and achievement gains* (Research paper prepared

for the Bill and Melinda Gates Foundation). Retrieved from MET project website: http://metproject.org/downloads/MET_Gathering_Feedback_Research_Paper.pdf

Kennedy, M. M. (2010). Attribution error and the quest for teacher quality. *Educational Researcher*, *39*, 591–598. http://dx.doi.org/10.3102/0013189X10390804.

Klieme, E. (2006). Empirische Unterrichtsforschung: Aktuelle Entwicklungen, theoretische Grundlagen und fachspezifische Befunde [Empirical instructional research: current developments, theoretical foundation and subject-specific results]. *Zeitschrift für Pädagogik, 52,* 765–773.

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

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. Forschungsbefunde, Reforminitiativen, Praxisberichte und Video-Dokumente* (pp. 43–57). Munich, Germany: Medienhaus Biering.

Kobarg, M., Dalehefte, I. M., & Menk, M. (2012). Der Einsatz systematischer Videoanalysen zur Untersuchung der Wirksamkeit des Unterrichtsentwicklungsprogramms SINUS an Grundschulen [The application of systematic video analyses to investigate the effectiveness of the instructional improvement program SINUS at primary schools]. In M. Kobarg, C. Fischer, I. M. Dalehefte, F. Trepke, & M. Menk (Eds.), *Maßnahmen zur Lehrerprofessionalisierung wissenschaftlich begleiten – verschiedene Strategien nutzen.* Münster, Germany: Waxmann.

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

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,* 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*, 494–509. http:dx.doi.org/10.1016/j.learninstruc.2007.09.002.

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. http://dx.doi.org/10.1037/a0032583.

Lakes, K. D., & Hoyt, W. T. (2008). What sources contribute to variance in observer ratings? Using generalizability theory to assess construct validity of psychological measures. *Infant and Child Development, 17,* 269–284. http://dx.doi.org/10.1002/icd.551.

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. http://dx.doi.org/10.1016/j.learninstruc.2008.11.001.

Lotz, M., Lipowsky, F., & Faust, G. (Eds.). (2013). [Documentation of the instruments of the project "Personality and learning development of primary school children" (PERLE) – part 3. Technical report regarding the PERLE video analyses]. *Materialien zur Bildungsforschung, Band 23/3*. *Dokumentation der Erhebungsinstrumente des Projekts „Persönlichkeits- und Lernentwicklung von Grundschulkindern" (PERLE) – Teil 3. Technischer Bericht zu den PERLE-Videoanalysen.* Frankfurt am Main, Germany: GFPF.

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,* 120–131. http://dx.doi.org/10.1016/j.cedpsych.2008.12.001.

Malmberg, L.-E., Hagger, H., Burn, K., Mutton, T., & Colls, H. (2010). Observed classroom quality during teacher education and two years of professional practice. *Journal of Educational Psychology, 102*, 916–932. http://dx.doi.org/10.1037/a0020920.

Mayer, D. P. (1999). Measuring instructional practice: can policymakers trust survey data? *Educational Evaluation and Policy Analysis*, *21*, 29–45. http://dx.doi.org/10.3102/01623737021001029.

Mayer, R. E. (2004). Should there be a three-strikes rule against pure discovery learning? The case for guided methods of instruction. *American Psychologist, 59*, 14–19. http://dx.doi.org/10.1037/0003-066X.59.1.14.

Medley, D. M., & Mitzel, H. (1963). Measuring classroom behavior by systematic observation. In N. L. Gage (Ed.), *Handbook of research on teaching* (pp. 247–328). Chicago: Rand-McNally.

Meyer, L., Seidel, T., & Prenzel, M. (2006). Wenn Lernsituationen zu Leistungssituationen werden: Untersuchung zur Fehlerkultur in einer Videostudie [When learning situations turn

into performance situations: investigating error culture in a video study]. *Schweizerische Zeitschrift für Bildungswissenschaften*, *28*, 21–41.

Neumann, K., Fischer, H. E., Labudde, P., & Viiri, J. (2009). Postersymposium Physikunterricht im Vergleich: Unterrichtsqualität in Deutschland, Finnland und Schweiz [Poster symposium physics instruction by comparison: Instructional quality in Germany, Finland and Switzerland]. In D. Höttecke (Ed.), *Chemie- und Physikdidaktik für die Lehramtsausbildung* (pp. 357–359). Berlin, Germany: LIT.

Newton, X. A. (2010). Developing indicators of classroom practice to evaluate the impact of district mathematics reform initiative: a generalizability analysis. *Studies in Educational Evaluation*, *36*, 1–13. http://dx.doi.org/10.1016/j.stueduc.2010.10.002.

Pauli, C., & Reusser, K. (2011). Expertise in Swiss mathematics instruction. In Y. Li, & G. Kaiser (Eds.), *Expertise in mathematics instruction: An international perspective* (pp. 85–107). New York, NY: Springer.

Pauli., C., Reusser, K., & Grob, U. (2007). Teaching for understanding and/or self-regulated learning? A video-based analysis of reform-oriented mathematics instruction in Switzerland. *International Journal of Educational Research, 46*, 294–305. http://dx.doi.org/10.1016/j.ijer.2007.10.004.

Petko, D., Waldis, M., Pauli, C., & Reusser, K. (2003). Methodologische Überlegungen zur videogestützten Forschung in der Mathematikdidaktik: Ansätze der TIMSS 1999 Video Studie und ihrer schweizerischen Erweiterung [Methodological considerations concerning video-based research in mathematic didactics]. *Zentralblatt für Didaktik der Mathematik*, *35*, 265–280.

Pianta, R. C., & Hamre, B. K. (2009). Conceptualization, measurement, and improvement of classroom processes: standardized observation can leverage capacity. *Educational Researcher*, *38*, 109–119. http://dx.doi.org/10.3102/0013189X09332374.

Pietsch, M. (2010). Evaluation von Unterrichtsstandards [Evaluation of classroom teaching standards]. *Zeitschrift für Erziehungswissenschaft*, *13*, 121–148. http://dx.doi.org/10.1007/s11618-010-0113-z.

Praetorius, A.-K., Lenske, G., & Helmke, A. (2012). Observer ratings of instructional quality: do they fulfill what they promise? *Learning and Instruction, 22*, 387–400. http://dx.doi.org/10.1016/j.learninstruc.2012.03.002.

Rakoczy, K. (2008). *Motivationsunterstützung im Mathematikunterricht: Unterricht aus der Perspektive von Lernenden und Beobachtern* [Motivational support in mathematics lessons: Instruction from the perspectives of learners and observers]. Münster, Germany:

Waxmann.

Rakoczy, K., & Pauli, C. (2006). Hoch inferentes Rating: Beurteilung der Qualität unterrichtlicher Prozesse [High-inference rating. Assessment of instructional quality]. In I. Hugener, C. Pauli, & K. Reusser (Eds.), *Videoanalysen. Dokumentation der Erhebungs- und Auswertungsinstrumente zur schweizerisch-deutschen Videostudie „Unterrichtsqualität, Lernverhalten und mathematisches Verständnis". Materialien zur Bildungsforschung* (Vol. 15); (pp. 206–233). Frankfurt am Main, Germany: GFPF.

Renkl, A., & Helmke, A. (1993). Prinzip, Nutzen und Grenzen der Generalisierungstheorie. [Principle, utility, and limits of generalizability theory]. *Empirische Pädagogik, 7*, 63–85.

Reusser, K. (2006). Konstruktivismus: Vom epistemologischen Leitbegriff zur Erneuerung der didaktischen Kultur. [Constructivism: from a key epistemological concept to a new didactic culture]. In M. Baer, M. Fuchs, P. Füglister, K. Reusser, & H. Wyss (Eds.), *Didaktik auf psychologischer Grundlage. Von Hans Aeblis kognitionspsychologischer Didaktik zur modernen Lehr- und Lernforschung* (pp. 151–168). Bern: h.e.p. verlag.

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. http://dx.doi.org/10.1037/a0027268.

Rosenshine, B., & Furst, N. (1973). Research on teacher performance criteria. In B. O. Smith. (Ed.), *Research in teacher education: A symposium.* Englewood Cliffs, NJ: Prentice Hall.

Roth, K. J., Druker, S. L., Garnier, H., Lemmens, M., Chen, C., Kawanaka, T., et al. (2006). *Teaching Science in five countries: Results from the TIMSS 1999 Video Study*. Washington, DC: National Center for Education Statistics.

Seidel, T., & Prenzel, M. (2006). Stability of teaching patterns in physics instruction: findings from a video study. *Learning and Instruction, 16*, 228–240. http://dx.doi.org/10.1016/j.learninstruc.2006.03.002.

Seidel, T., Prenzel, M., Duit, R., Euler, M., Geiser, H., Hoffmann, L., et al. (2002). „Jetzt bitte alle nach vorne schauen!" Lehr-Lernskripts im Physikunterricht und damit verbundene Bedingungen für individuelle Lernprozesse ["Can everybody look to the front of the classroom please?" Patterns of instruction in physics classrooms and its implication for students' learning]. *Unterrichtswissenschaft, 30*, 52–77.

Seidel, T., Prenzel, M., Rimmele, R., Dalehefte, I. M., Herweg, C., Kobarg, M., et al. (2006). Blicke auf den Physikunterricht: Ergebnisse der IPN Videostudie [Views on physics lessons: results from the IPN video study]. *Zeitschrift für Pädagogik*, *52*, 799–821.

Seidel, T., Prenzel, M., Schwindt, K., Rimmele, R., Kobarg, M., & Dalehefte, I. M. (2009).

The link between teaching and learning – investigating effects of physics teaching on student learning in the context of the IPN video study. In T. Janík, T. Seidel, & P. Najvar (Eds.), *The power of video studies in investigating teaching and learning in the classroom* (pp. 161–180). Münster, Germany: Waxmann.

Seidel, T., & Shavelson, R. (2007). Teaching effectiveness research in the past decade: the role of theory and research design in disentangling meta-analysis results. *Review of Educational Research, 77,* 454–499. http://dx.doi.org/10.3102/003465430731031.

Shavelson, R., & Dempsey-Atwood, N. (1976). Generalizability of measures of teaching behavior. *Review of Educational Research*, *46*, 553–611. http://dx.doi.org/10.3102/00346543046004553.

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

Shavelson, R. J., Webb, N. M., & Burstein, L. (1986). Measurement of teaching. In M. C. Wittrock (Ed.), *Handbook of research on teaching* (pp. 50–91). New York, NY: Macmillan.

Shumate, S., Surles, J., Johnson, R., & Penny, J. (2007). The effect of the number of scale and non-normality on the generalizability coefficient: a Monte Carlo study. *Applied Measurement in Education, 20*, 1–20. http://dx.doi.org/10.1080/08957340701429645.

Smith, P. L. (1978). Sampling errors of variance components in small sample generalizability studies. *Journal of Educational Statistics, 3*, 319–346.

Staub, F. C. (2007). Mathematics classroom cultures: methodological and theoretical issues. *International Journal of Educational Research, 46*, 319–326. http://dx.doi.org/10.1016/j.ijer.2007.10.007.

Stigler, J. W., Gallimore, R., & Hiebert, J. (2000). Using video surveys to compare classrooms and teaching across cultures: examples and lessons from the TIMSS video studies. *Educational Psychologist, 35*, 87–100. http://dx.doi.org/10.1207/S15326985EP3502_3.

Stigler, J., Gonzales, P., Kawanaka, T., Knoll, S., & Serrano, A. (1999). *The TIMSS-videotape classroom study* (Technical Report). Los Angeles, CA.

Tschannen-Moran, M., & Woolfolk Hoy, A. (2001). Teacher efficacy: capturing an elusive construct. *Teaching and Teacher Education, 17*, 783–805. http://dx.doi.org/10.1016/S0742-051X(01)00036-1.

Vieluf, S., & Klieme, E. (2011). Cross-nationally comparative results on teachers' qualification, beliefs, and practices. In Y. Li & G. Kaiser (Eds.), *Expertise in mathematics*

instruction (pp. 295–325). New York, NY: Springer.

Waldis, M., Grob, U., Pauli, C., & Reusser, K. (2010). Der schweizerische Mathematikunterricht aus der Sicht von Schülerinnen und Schülern und in der Perspektive hochinferenter Beobachterurteile [Swiss mathematics instruction from the perspective of students and high-inference observer ratings]. In K. Reusser, C. Pauli, & M. Waldis (Eds.), *Unterrichtsgestaltung und Unterrichtsqualität. Ergebnisse einer internationalen und schweizerischen Videostudie zum Mathematikunterricht* (pp. 171–208). Münster, Germany: Waxmann.

Williams, E. J. (1959). The comparison of regression variables. *Journal of the Royal Statistical Society, Series B, 21*, 396–399.