Learning and Instruction 31 (2014) 2–12

Contents lists available at ScienceDirect

Learning and Instruction

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

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

Anna-Katharina Praetorius [a,*], Christine Pauli [b], Kurt Reusser [b], Katrin Rakoczy [c], Eckhard Klieme [c]

[a] Department of Psychology, University of Augsburg, Universitaetsstraße 10, 86159 Augsburg, Germany
[b] Institute of Education, University of Zurich, Freiestrasse 36, 8032 Zurich, Switzerland
[c] German Institute for International Educational Research, Schlossstrasse 29, 60486 Frankfurt am Main, Germany

A R T I C L E I N F O

Article history:
Received 29 January 2013
Received in revised form
12 December 2013
Accepted 15 December 2013

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

A B S T R A C T

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 these 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.

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

* Corresponding author. Tel.: +49 821 598 5519; fax: +49 821 598 5289.
E-mail address: anna.praetorius@phil.uni-augsburg.de (A.-K. Praetorius).

0959-4752/$ – see front matter
http://dx.doi.org/10.1016/j.learninstruc.2013.12.002

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.

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^{2}Ca	Bürgermeister et al. (2011)	Mathematics	2
CPV video study	Janík et al. (2006)	Physics	4–8
QuiP	Neumann, Fischer, Labudde, and Viiri (2009)	Physics	2
DESI	T. Helmke et al. (2008)	English	2
IPN video study	Seidel et al. (2009)	Physics	2
LPS	Clarke, Keitel, and Shimizu (2006)	Mathematics	10
PERLE	Lotz, Lipowsky, and Faust (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, Dalehefte, and Menk (2012)	Mathematics & science	3
TIMSS	Baumert et al. (1997) Stigler, Gallimore, and Hiebert (2000)	Mathematics	1(–3[a])
TIMSS 1999 video study	Hiebert et al. (2003)/ Roth et al. (2006)	Mathematics & science	1
VERA	A. Helmke et al. (2008)	German	0–2[b]
		Mathematics	0–2[b]
		Other subjects	0–1[b]

[a] In the part of the TIMS study that took place in Germany, some of the teachers were videoed three times (Kunter, 2005).
[b] In this study, neither the subject nor the number of lessons per teacher were set, so not all teachers were videoed teaching every subject.