Learning and Instruction 22 (2012) 387–400

Contents lists available at SciVerse ScienceDirect

Learning and Instruction

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

Observer ratings of instructional quality: Do they fulfill what they promise?☆

Anna-Katharina Praetorius [a,*], Gerlinde Lenske [a], Andreas Helmke [b]

[a] DFG-Graduate School "Teaching and Learning Processes", University of Koblenz-Landau, Thomas-Nast-Straße 44, 76829 Landau, Germany
[b] Department of Psychology, University of Koblenz-Landau, Germany

ARTICLE INFO

Article history:
Received 19 July 2011
Received in revised form
3 March 2012
Accepted 11 March 2012

Keywords:
Instructional quality
Observer ratings
Generalizability theory
Reliability
Validity

ABSTRACT

Despite considerable interest in the topic of instructional quality in research as well as practice, little is known about the quality of its assessment. Using generalizability analysis as well as content analysis, the present study investigates how reliably and validly instructional quality is measured by observer ratings. Twelve trained raters judged 57 videotaped lesson sequences with regard to aspects of domain-independent instructional quality. Additionally, 3 of these sequences were judged by 390 untrained raters (i.e., student teachers and teachers). Depending on scale level and dimension, 16–44% of the variance in ratings could be attributed to instructional quality, whereas rater bias accounted for 12–40% of the variance. Although the trained raters referred more often to aspects considered essential for instructional quality, this was not reflected in the reliability of their ratings. The results indicate that observer ratings should be treated in a more differentiated manner in the future.

1. Introduction

Instructional quality is considered one of the major modifiable factors influencing students' achievement (Hattie, 2009). In order to improve the quality of instruction it is important to identify the strengths and weaknesses of teachers concerning their teaching. To measure these strengths and weaknesses, ratings are indispensable. However, these ratings are error-prone (for an overview see Hoyt, 2000). If so, it is difficult to draw conclusions for research and practice on the basis of rating data. In instructional research, the amount and causes of observer rater bias when measuring instructional quality have rarely been investigated to date. Therefore, the aim of this study is to shed light on this topic using generalizability theory (Brennan, 2001; Shavelson & Webb, 1981, 1991) and content analysis of qualitative interviews (Mayring, 2004).

After a short introduction to theoretical issues regarding instructional quality and possibilities to measure it, ratings as a data collection method and problems associated with this method are addressed. Afterward, generalizability theory and qualitative measures are described as ways of gaining insight into the amount and causes of rater bias.

1.1. *What is instructional quality?*

One of the most influential research traditions regarding instructional quality is teacher effectiveness research. The central paradigm of teacher effectiveness research is the process-mediation-product paradigm (Brophy, 2000; Shuell, 2001). Within this paradigm it is assumed that teachers can only provide opportunities to learn (=process). The utilization of these opportunities (=mediation) has to be performed by the students and may lead to gains in learning (=product) in a next step.

One popular model—especially in German-speaking countries—which summarizes the most important aspects of instructional quality, was developed by Klieme et al. (see e.g., Klieme, Lipowsky, Rakoczy, & Ratzka, 2006). The model deals with domain-independent instructional quality, measured via high-inference measures (i.e., ratings which require a certain amount of interpretation). In the model, three basic dimensions of instructional quality are distinguished: (a) classroom management, (b) cognitive activation, and (c) personal learning support. According to Klieme, Schümer, and Knoll (2001), these dimensions are essential for good instruction in any school subject, school type and grade.

The first dimension, *classroom management*, focuses on providing time to learn for students. Its underlying assumption is that the more time students have for learning, the more opportunities they have to be involved in learning processes (Brophy, 2000; Walberg & Paik, 2000). To provide enough time for learning, teachers have to prevent or to deal effectively with disruptions and

☆ This research was supported in part by grants from the DFG (German Research Foundation). We thank Loredana Mihalca for helpful comments on an earlier version of the manuscript.

* Corresponding author. Tel.: +49 6341 280 33 261; fax: +49 6341 280 33 260.
E-mail address: praetorius@uni-landau.de (A.-K. Praetorius).

0959-4752/$ – see front matter © 2012 Elsevier Ltd. All rights reserved.
doi:10.1016/j.learninstruc.2012.03.002

disciplinary conflicts (Borich, 2007). The second dimension, *cognitive activation,* refers to the cognitively demanding processes of problem-solving and understanding (e.g., by providing challenging tasks; see Hugener et al., 2009; Reusser, 2006). The most important aspects for enhancing students' motivation to learn are subsumed in the third dimension, *personal learning support*, which includes individual learner support, a positive teacher–student relationship, constructive and positive teacher feedback as well as a positive approach toward students' errors.

1.2. Why should instructional quality be measured?

Waxman, Hilberg, and Tharp (2004) mentioned four reasons for investigating instructional quality: (a) description of instructional practices, (b) investigation of instructional inequities for different groups of students, (c) improving teacher education programs, and (d) improvement of teachers' classroom instruction. The first two reasons can be subsumed under the heading "scientific interest". Researchers and the general public are interested in seeing, for example, the differences in instruction between countries (e.g., Stigler & Hiebert, 1999), which students profit from which type of instruction (for an overview see Snow, 1989), whether one instructional method is better than another (e.g., Farkas, 2003) et cetera. Implementing the third point requires the findings of point one and two and deals with communicating them in teacher education.

A totally different focus—especially from the first two points—is pursued with improvement as a direct purpose for measuring instructional quality. In the last years a shift in educational policy can be observed toward evidence-based educational decisions (Oelkers & Reusser, 2008). In this context, teachers are more expected to diagnose their own teaching strengths and weaknesses, instead of relying on implicit assumptions about their own teaching. As extensive observations by trained raters are unaffordable, reciprocal observation of lessons by teachers to give feedback about instruction is demanded and supported by policy (Kultusministerkonferenz, 2004) and practice-orientated research (e.g., Helmke et al., 2011).

1.3. How can instructional quality be measured?

According to Clausen (2002), the most common method to measure instructional quality in the process-product paradigm is ratings by external observers. The main arguments for using external observer ratings are: First, they are "the most direct way to measure instructional quality" (Clare, Valdés, Pascal, & Steinberg, 2001, p. 2; see also Pianta & Hamre, 2009; Walberg & Haertel, 1980). Second, the complexity of instruction can be adequately depicted via observer ratings (Helmke, 2010; Petko, Waldis, Pauli, & Reusser, 2003). Third, actors see situations from a certain, biased angle (Jones & Nisbett, 1971; Storms, 1973).

However, observer ratings also have some drawbacks. Two of them are mainly mentioned in literature (Kunter, 2005; Lüdtke, Robitzsch, Trautwein, & Kunter, 2009): First, observers can only observe a very short time period. In video-based classroom research, usually one or a few lessons are videotaped. This may threaten the validity of such ratings. Second, using observers in studies is very expensive.

When comparing the advantages and disadvantages of observer ratings mentioned in literature, one could conclude that measuring the quality *of a certain lesson* with observer ratings seems appropriate. However, the argumentation that observer ratings of instructional quality are advantageous is mainly based on plausibility assumptions. Literature on rater bias shows that the judgment processes of trained raters can lead to biased ratings.

1.4. Problems associated with ratings

As many objects cannot be measured directly, observer ratings are one of the major methods of collecting data in research (Hoyt, 2000). For decades, researchers have been concerned with the problems associated with such ratings (for an early example, see Guilford, 1954). Especially rater bias (e.g., leniency/severity bias) is often mentioned as a drawback of this method. Rater bias is generally defined as a disagreement among raters that can be traced back to different interpretations of rating scales or unique, idiosyncratic perceptions of the target in question (Hoyt, 2000). Therefore, in most cases rater bias is regarded as measurement error.

Investigations concerning rater bias are conducted on a regular basis. A meta-analysis of Hoyt and Kerns (1999) summarizes findings in different research areas (e.g., psychotherapy or job evaluation) and concludes that about 37 percent of the variance in ratings is due to rater bias. The authors also investigated moderators of rater bias and concluded that the highest risk ratings are—among other characteristics (e.g., acquaintanceship)—inferential ratings by observers with little (i.e., less than 5 h) or no training. Thus, one could conclude that if raters are trained sufficiently, ratings of instructional quality can be carried out without big problems. However, rating instructional quality is highly complex and rater trainings concerned with complex objects are not automatically effective in dealing with rater bias and accuracy, as some studies have pointed out (cf. Lumley & McNamara, 1995). Researchers conducting video-based classroom studies with high-inference ratings assume as a rule that training is effective when there is consensus about a joint theoretical understanding in the training group (Rakoczy & Pauli, 2006; Seidel, 2005).[1] Thus, it remains unclear whether rater trainings really work as intended.

1.5. Investigating the amount of rater bias in measuring instructional quality

Regardless of whether the efficacy of rater trainings is directly investigated or not, every scientific study has to report whether the ratings they use for their conclusions about certain topics are sufficiently reliable. Indeed, different coefficients have been developed to quantify reliability (e.g., Cohens κ, Intraclass coefficient, Kendalls τ). One disadvantage of these is that they are not flexible. Thus, it is only possible to prove one kind of reliability at a certain time point. Another disadvantage is that no further information about the amount and causes for biases is given in addition to the reliability coefficient. To circumvent these disadvantages, Cronbach, Gleser, Nanda, and Rajaratnam (1972) developed generalizability theory.

1.5.1. What is generalizability theory and how does it work?

Generalizability theory (hereafter referred to as G theory) is a powerful framework to deeply scrutinize ratings. G theory enables the separation of multiple sources of error (called facets) via variance component analysis (Brennan, 2001; Shavelson & Webb, 1991) and thus serves as a framework for examining the dependability of behavioral measurements.

Depending on the objective, different coefficients are used to estimate the dependability of the measurement in question. The generalizability coefficient (ρ^2) is used if the aim is to undertake relative decisions, i.e., comparisons between or within persons and

[1] This assumption is made as high-inference measures of instructional quality are based on whole lessons. Carrying out a sufficient number of ratings to screen the efficacy of trainings would lead to very long training durations.