Fauth, Benjamin; Göllner, Richard; Lenske, Gerlinde; Praetorius, Anna-Katharina; Wagner, Wolfgang

Who sees what? Conceptual considerations on the measurement of teaching quality from different perspectives

Praetorius, Anna-Katharina [Hrsg.]; Grünkorn, Juliane [Hrsg.]; Klieme, Eckhard [Hrsg.]: Empirische Forschung zu Unterrichtsqualität. Theoretische Grundfragen und quantitative Modellierungen. 1. Auflage. Weinheim; Basel : Beltz Juventa 2020, S. 138-155. - (Zeitschrift für Pädagogik, Beiheft; 66)

66. Beiheft April 2020

ZEITSCHRIFT FÜR PÄDAGOGIK

Empirische Forschung zu Unterrichtsqualität. Theoretische Grundfragen und quantitative Modellierungen

BELTZ JUVENTA

Zeitschrift für Pädagogik · 66. Beiheft

Zeitschrift für Pädagogik · 66. Beiheft

Empirische Forschung zu Unterrichtsqualität

Theoretische Grundfragen und quantitative Modellierungen

Herausgegeben von
Anna-Katharina Praetorius, Juliane Grünkorn
und Eckhard Klieme

BELTZ JUVENTA

ISSN: 0514-2717
ISBN 978-3-7799-3534-6 Print
ISBN 978-3-7799-3535-3 E-Book (PDF)
Bestellnummer: 443534

1. Auflage 2020

in der Verlagsgruppe Beltz · Weinheim Basel
Werderstraße 10, 69469 Weinheim

Herstellung: Hannelore Molitor
Satz: text plus form, Dresden
Druck und Bindung: Beltz Grafische Betriebe, Bad Langensalza
Printed in Germany

Weitere Informationen zu unseren Autoren und Titeln finden Sie unter: www.beltz.de

Inhaltsverzeichnis

Themenblock V: Modellierung der Wirkungen von Unterrichtsqualität (englischsprachig)

Kommentare

Themenblock IV: Zur Bedeutung unterschiedlicher Perspektiven bei der Erfassung von Unterrichtsqualität

Benjamin Fauth/Richard Göllner/Gerlinde Lenske/Anna-Katharina Praetorius/ Wolfgang Wagner

Who Sees What?

Conceptual Considerations on the Measurement of Teaching Quality from Different Perspectives

Abstract: One puzzling finding in education research is that teachers, students, and external observers agree only marginally on their ratings of teaching quality. In this theoretical contribution, we summarize and reappraise previous findings on agreement between different raters of teaching quality. We explain these findings by thoroughly examining the instruments that have been used to measure teaching quality. Building on this, we propose a reference perspective matrix, which should be useful in explaining perspective-specific rating mechanisms behind responses to certain survey or observation items. The reference perspective matrix could thus afford a theoretical foundation for future studies on the assessment of teaching quality.

Keywords: Teaching Quality, Measurement, Validity, Classroom Management, Agreement

1. Introduction

Teaching quality is one of the most prominent and powerful predictors of student learning in school (Hattie, 2009). However, properly measuring teaching quality is still a huge challenge. Researchers and practitioners often assess teaching quality by using ratings from students, teachers, or trained observers. Although researchers assume that these different approaches assess the same target constructs, empirical studies have repeatedly found low or even zero correlations between these data sources when they are applied to the same sample of classes (e. g., Clausen, 2002; Fauth, Decristan, Rieser, Klieme & Büttner, 2014b; Kunter & Baumert, 2006; Wagner, Göllner, Werth, Voss, Schmitz & Trautwein, 2016).

In the present contribution, we seek explanations for these results. Drawing on a variety of theoretical traditions, including personality and social psychology approaches, we present theoretical considerations that should help to explain previous findings and that may serve as a framework for future studies. Our approach is to examine the items

found in commonly used survey and observation instruments in order to investigate how the specific wording of items might shape responses by students, teachers, and observers. Focusing on the construct of classroom management, we show that items refer either to teacher behavior, to student behavior, or to a mixture of both, and that these different item references have consequences for assessments of classroom management. On the basis of this observation, we present a matrix of item references and rater perspectives that can help us understand how a certain item's wording may shape answers to this item from a certain perspective.

2. Teaching Quality: The Problem of Alignment Between Perspectives

The increased use of direct measures of teaching quality has been accompanied by a growing interest among researchers in the psychometric quality of these measures. An important indicator of psychometric quality is the degree to which different data sources produce the same results in evaluating the same instructor. In a seminal study, Clausen (2002) drew on high-inference video ratings of the German TIMSS 1995 video data and compared these ratings to survey responses collected from students, and teacher self-reports. The study found that only 13 out of 36 correlations between corresponding scales (as measured from the three perspectives) yielded values differing significantly from zero. The average correlation between the perspectives was .16, with a range between −.28 and .45 (Clausen, 2002, p. 129). By September 2019, this study had been cited 450 times, according to Google Scholar: this indicates that other researchers have indeed paid attention to this finding.

In the last 15 years, numerous studies have applied a similar approach and found relative agreements between rater perspectives roughly in the range reported by Clausen (2002; Camburn & Barnes, 2004; Chaplin, Gill, Thompkins & Miller, 2014; De Jong & Westerhof, 2001; Desimone, Smith & Frisvold, 2010; Fauth et al., 2014b; Gitomer et al., 2014; Kunter & Baumert, 2006; Wagner et al., 2016; Wettstein, Ramseier, Scherzinger & Gasser, 2016; Mayer, 1999; Kaufman, Stein & Junker, 2016). These studies have two consistent findings: First, overall, the correlations between measures obtained from student ratings, teacher self-evaluations, and observation protocols are low. Second, the highest correlations are among indicators of classroom management, rather than indicators of other constructs, such as cognitive activation or student support.

2.1 Perspective-Specific Validities?

The relatively low correlations between perspectives have led researchers to think about the relations between different data sources in terms of validity rather than reliability (Kunter & Baumert, 2006). Additionally, it has been put into question whether it makes sense from a methodological standpoint to think of teaching quality as a perspective-independent construct (Clausen, 2002). Accordingly, *perspective-specific validities* have

been hypothesized for the different data sources: "It is conceivable that students' and teachers' perceptions tap different aspects of the classroom environment, rather than the same underlying construct" (Kunter & Baumert, 2006, p. 234). Indeed, from an epistemological perspective, we have to acknowledge that humans' perceptions of their environment are perspective specific in nature (Graumann, 1960). Both philosophers and psychologists have argued convincingly that our knowledge of the world is and will always be an idiosyncratic construction that is fundamentally affected by our individual preconceptions and schemes of perception. This idiosyncratic way of perceiving our environment is rooted in previous experiences during the life course. As these experiences naturally differ between persons, their perceptions of the environment will also differ.

The literature nowadays commonly refers to perspective-specific validities to explain and/or justify low correlations between perspectives (e. g., Fauth, Decristan, Rieser, Klieme & Büttner, 2014a; Wettstein et al., 2016). In the present paper we argue, however, that this approach has at least two pitfalls.

First, the term teaching quality is currently not used in a perspective-specific way, either by teachers or by those conducting substantive research. In the contexts where these measures are usually applied, most people are interested in *teaching quality,* not in *teaching quality as perceived from a certain perspective.* When we think about classroom interactions that foster student learning, we usually do not think about 'teachers' perceived classroom interactions' or 'students' perceived classroom interactions.' Thus, from a scientific perspective, knowing that human perceptions are perspective specific in nature should not limit the search for the best instruments to measure teaching quality.

Second, the plausibility of the concept of perspective-specific validities may vary for different constructs. The student's perspective on the feeling of being emotionally supported by the teacher might have a special relevance. Some degree of nonagreement will be the standard for these support dimensions, even within one rater group's perspective (e. g., students in a class; Schweig, 2016). In contrast, classroom disruptions or teacher strategies to ensure smooth transitions could be perceived differently from different perspectives. But we would not expect different disruptions or different transitions to be in evidence, depending on who is doing the rating. The events rated are relatively distinct ones in the classroom that – in principle – everyone should be able to rate accurately. Accordingly, deviations between perspectives should be understood in light of reliability rather than validity.

Consequently, at least for classroom management, we assume that there is a 'true score' and that while deviations between perspectives are possible, they require explanation. Having acknowledged that agreement between perspectives can be expected, nonagreement has to be explained. The concept of perspective-specific validities is potentially attractive for researchers, as it offers a plausible explanation for nonagreement. But the risk that this concept entails is that deviations between perspectives may be unquestioningly accepted, instead of being properly investigated.

2.2 Approaches to Explaining Low Correlations Between Perspectives

A number of reasons for perspective-specific deviations have been advanced in the literature. For instance, researchers have considered that students might find it difficult to judge the didactic value of specific math assignments, or that teachers might find it difficult to judge the correct learning speed for students (Kunter & Baumert, 2006; Mayer, 1999). Some researchers have expressed doubts on whether *students* sufficiently understand the pedagogic principles underlying teaching (e. g., Fauth et al., 2014b). Thus, their agreement with teachers and observers would be lower for constructs requiring an understanding of pedagogy (Clausen, 2002). One example would be the 'Socratic-dialogue practice' (e. g., "In math class, our teacher lets us keep making the wrong assumption until we notice it ourselves.").

As *external observers* usually only get a short look at what is going on in the classroom during the school year (usually one to five lessons; Praetorius, Pauli, Reusser, Rakoczy & Klieme, 2014), a "sampling effect" (Clausen, 2002, p. 90) is assumed to limit the accuracy of observer ratings. Consequently, for constructs of low observability (e. g., scales that refer to seldom-occurring events such as the 'social orientation' of the teacher being demonstrated, as in the following, sample item: "Our math teacher cares about students' problems") or high variability across lessons (Kane et al., 2012; Praetorius et al., 2014), observer ratings would not correlate highly with student and teacher ratings. In the case of *teacher* ratings, self-serving biases can be a problem. Hence, one would expect lower correlations to student and observer ratings for scales with a strong evaluative component. Clausen (2002, p. 91) cites "discipline" as an example (sample item: "The students in this class mess around a lot" – see Section 3 for a discussion of this scale).

In an attempt to account for these challenges, Clausen (2002) named one characteristic of each rating perspective (didactic understanding, small sample of lessons, self-serving bias) that limits the accuracy of ratings from that perspective. Each major characteristic would lead to deviations from the other two perspectives. According to Clausen, such deviations should differ in extent, depending on the characteristics of the target constructs (demands regarding didactic understanding, observability, evaluativeness). Clausen (2002) rated these characteristics of the target constructs for each scale used in the TIMSS video data, to test the assumptions of this model empirically. However, the results showed little support for these assumptions.

On second glance, the categorization of the target constructs and their measurement shows that the observability or evaluativeness of an item is difficult to determine. We discussed these questions extensively and uncovered one possible explanation: That the inconsistent findings may be explained by the specific wording of survey and observation items. More precisely, we suspected that item characteristics such as observability or evaluativeness strongly depend on whom or what an item refers to (item reference): that is, whether the item refers to the teacher's or to students' classroom behavior. In order to explain this, we need to briefly discuss the theoretical foundations of teaching quality.

2.3 Two Sides of a Coin – the Significance of Teacher and Student Behavior for Teaching Quality

From a theoretical standpoint, teaching is interactive in nature, and teaching quality thus can only be understood as an interplay between teachers and students. Current research agrees with this line of reasoning, and accordingly has conceptualized teaching quality as a complex social process that takes place in the interactions between students and teachers (Doyle, 2013). This also implies that teaching quality is not completely determined by the teacher's behavior (Göllner, Fauth, Lenske, Praetorius & Wagner, in this issue; Fauth et al., in press). Instead, teaching must be understood as a "social practice that is co-constructed by students and teachers" (Praetorius, Klieme, Herbert & Pinger, 2018, p. 6). This theoretical view of teaching quality as a co-construction between teachers and students is also reflected in theoretical conceptualizations of classroom management. The ecological approach to classroom management established by Doyle (2013), and the early work of Kounin (1970) describe the characteristics of interactions between teachers and students rather than specific teacher behaviors (Klieme, 2006).

The quality of a certain teaching strategy will always be hard to evaluate without knowing the students' reactions to it (or the students' behavior preceding the teachers' action). For instance, the same classroom management strategy will have a completely different impact when it is applied in a class of well-behaved students compared to poorly behaved students. Accordingly, Praetorius et al. stated that "Classroom management is both a condition for students getting attentive (e. g., through teacher monitoring) and an indication of students being attentive (e. g., lack of interruptions)" (2018, p. 6).

This conceptualization has consequences. Given that most teaching quality assessments are designed to evaluate the teacher, the notion that teaching quality also depends on the students is not trivial.

The current policy press is to develop measures that allow for inferences about *teacher* effectiveness. Using particular measures, the goal is to be able to make some type of claim about the qualities of a teacher. Yet, to varying degrees, the measures we examine do not tell us only about the teacher. A broad range of contextual factors also contributes to the evidence of the *teaching* quality, which is more directly observable. (Gitomer & Bell, 2013, p. 416).

This understanding of teaching quality as a co-construction between teachers and students is seemingly shared by many researchers in the field of education research and educational psychology. For example, in a study in Germany (Clausen, Schnabel & Schröder, 2002), 22 researchers from these two fields were asked to rate student survey scales in regard to similarities and differences ('free pile sorting'). Using multidimensional scaling, the authors showed that participants used the degree to which the scales referred to student or teacher behavior, to sort items (Clausen et al., 2002). This result shows that – at least in researchers' understandings of these scales – item reference plays a central role: Items refer in varying degrees to student or to teacher behavior.

3. The Perspective Reference Matrix: A Taxonomy for Understanding Perspective-Specific Rating Processes

The notion that teaching quality is not limited to teacher behavior, but also depends on student behavior, is reflected in the items that researchers have used to measure teaching quality. We believe that this insight can play a particularly important role in understanding previous results on correlations between perspectives. In the present paper, we argue that how an item rates in terms of observability or evaluativeness will always depend on the kinds of questions asked, and on the person who has to answer these questions. It is therefore necessary to analyze perspective-specific judgments at the level of specific items.

Consider the above-mentioned sample item for *discipline* ("The students in this class mess around a lot"): This item refers not to teacher behavior but to student behavior. Thus, it seems at least questionable whether it is really highly evaluative *when answered from the teacher's perspective* (see Section 2.2). The claim that students in a class tend to mess around a lot might even be a good, self-serving explanation for teachers in chaotic classrooms. This would make the item nonevaluative for teachers. In contrast, imagine a student who has to respond to the item "In this class, I mess around a lot". As the student's (mis-)behavior is now being evaluated, this item becomes highly evaluative – but only *when it is answered from the student's individual perspective.* Finally, external observers do not need excuses for chaotic classrooms; nor will they feel evaluated when students mess around. This simple example shows that evaluativeness and observability (see Section 2.2) are not attributes of constructs or items, but rather of item-rater combinations: How evaluative an item is will always depend on who is answering it and to whom the item refers. These examples also suggest that these differences may relate to differences between self- and other-ratings: The same item can require a self-rating from one perspective but an other-rating from the other perspective. This thought can be formalized in a matrix (Fig. 1) of rater perspectives (who is rating?) and item references (what is rated?).

To understand the specific mechanisms that underlie responses to a certain item, it is crucial to be aware both of the perspective from which an item is answered (teacher, student, or external observer) and of what the item refers to (teacher actions, student actions, or a mixture of both). Additionally, the response to a certain item from a certain perspective will be driven by the quality and the quantity of information that is available to a rater, as well as the degree of ego-involvement that a specific item wording implies for a certain rater. We describe these psychological mechanisms in detail in Section 3.3.

In the following sections, we first of all evaluate whether the parameters of this taxonomy are the most relevant ones. To do so, we first review currently used instruments, to show that the item reference dimension is relevant to them. Afterwards, we examine the psychological processes that are assumed to be responsible for the differences in self- and other-ratings that occur when different item references are rated from different perspectives.

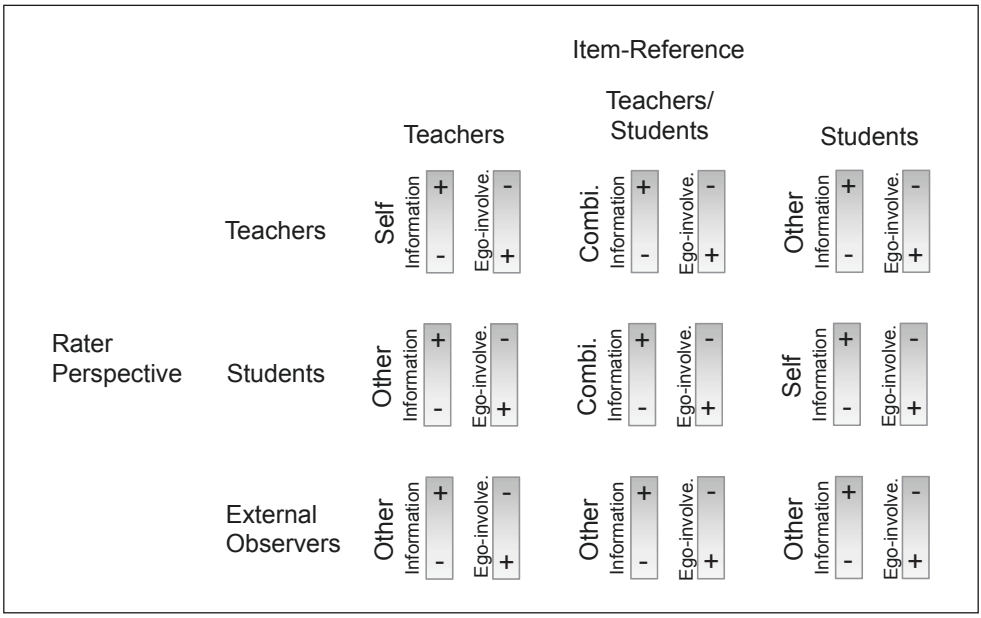

Fig. 1: Reference perspective matrix

3.1 Different Item References in Assessments of Classroom Management

In the following section, we review whether and how the different item references (columns in the taxonomy; see Fig. 1) relate to the measures that are commonly used to assess teaching quality from different perspectives (rows in the taxonomy). Here, we concentrate on the field of classroom management. The instruments discussed below were selected according to two criteria: First, we selected instruments and items for measuring classroom management from each of the three perspectives of external observations, student ratings, and teacher self-ratings. Second, within each of the perspectives we concentrated on those instruments that are either most popular in the educational system (e. g., the Tripod student survey in the US) or that are most frequently used in empirical education research. In our review, we did not make any a priori assumption that one of the perspectives (e. g., external observations) would be superior to the others, in the sense that one group of instruments would – in general – provide more accurate ratings than the others. Additionally, the following sections should not be read as an evaluation of specific instruments. Rather, the instruments reviewed below serve as examples of the way teaching quality is assessed from the different perspectives. The primary focus of this review is: To what extent do these instruments refer to teacher and/or student actions in the classroom?

External observations. The CLASS framework (Classroom Assessment Scoring System; Pianta & Hamre, 2009), which is one of the most frequently used classroom obser-

vation systems, explicitly considers the role of student behavior in successful classroom management. "In contrast to traditional observation protocols that focus on teacher actions, CLASS-S is representative of more recent evaluation protocols that focus on the actions and interactions of both teachers and students" (Gitomer et al., 2014, p. 9). In the CLASS manuals (e. g., Pianta, La Paro & Hamre, 2008), we find items both on teachers' classroom management behavior (e. g., "The teacher is consistently proactive and monitors the classroom effectively"; Pianta et al., 2008, p. 45) and on students' behavior (e. g., "There are few, if any, instances of student misbehavior in the classroom"; Pianta et al., 2008, p. 45). The authors point out in a footnote to the *behavior management* scale that in certain classrooms the teacher strategies described in the protocol might not be observable – "because behavior is so well managed. If there is no evidence of student misbehavior, it is assumed that effective behavioral strategies are in place and a classroom may score in the high range" (Pianta et al., 2008, p. 45). The fact that an assumption about effective behavioral strategies replaces observed teacher behavior in such cases shows that there is indeed a dependency of ratings on student behavior.

Other rating instruments also explicitly address students' behavior when evaluating classroom management. In the Framework for Teaching (FFT; Danielson, 2007), the absence of student misbehavior serves as an indicator of *managing student behavior.* In the video rating instruments that capture the three basic dimensions of teaching quality (Klieme, Pauli & Reusser, 2009), student disruptions are regularly assessed as an indicator of a teacher's classroom management (see Lipowsky et al., 2009; Praetorius et al., 2018).

Student ratings. The Tripod Classroom Environment Survey (Ferguson, 2010) is one of the most frequently applied student surveys in the United States, and has also been used in the Measures of Effective Teaching (MET) study (Kane et al., 2012). Interestingly, in the field of classroom management, the questionnaire asks only about student behavior, not about teacher behavior. These are the items of the *control* scale, which represents one of the Tripod's 7 Cs (sample item: "Student behavior in this class is a problem"; Wallace, Kelcey & Ruzek, 2016, p. 1857; the "Cs" refer to care, captivate, challenge, confer, clarify, consolidate, and control).

Göllner, Wagner, Rose, Fauth, and Nagengast (2018) reviewed student survey items from five large-scale studies conducted in Germany in recent years, and pooled the data collected in these studies into one integrated data set. This final data set included data from a total of 95,328 students. A scale was only included in this data set when it was applied in at least two different studies to 5 % of the whole sample (5,766 students). This approach justifies the authors' assumption that the scales they included constitute a representative sample of what is usually used to capture student ratings of teaching quality in German large-scale assessments. In the field of classroom management, each of the five large-scale studies included items on students' discipline or disruptions in the classroom (student reference). In three out of five studies, items on teachers' monitoring behavior were used (teacher reference). Additionally, two studies asked about the "inefficient use of time" in class (e. g., "In math, a lot of time is wasted"), where the item's wording leaves it open as to whom it refers (combined item reference).

Teacher self-ratings. In the teacher version of the Classroom Assessment Scoring System (CLASS-T; Hamre, 2008), teachers were asked to judge their "areas of strength and growth" (ranging from 1 = area of much growth to 5 = area of great strength). The item "Using time productively" is described as follows: "Productive classrooms are like 'well-oiled machines' – students in these classrooms know what they should be doing and always have something to do" (Gitomer et al., 2014, p. 30). Thus, this item refers to both teachers and students.

Studies that use teacher self-evaluations often use items similar to those used with students (where the item refers to teacher behavior, the third-person perspective is changed to a first-person formulation; e. g., Clausen, 2002; Kunter & Baumert, 2006; Wagner et al., 2016; Wettstein et al., 2016). Accordingly, we find a similar variety of item references (teachers, students, and a mixture of both) in teacher survey items and student survey items.

Summary. Summing up, we can conclude that all of the items used in the aforementioned studies can be categorized into three groups: (a) items that refer to student actions (e. g., student behavior, control, discipline, and disruption), (b) Items referring to teacher actions (e. g., monitoring, teacher awareness of student conduct), and (c) items that are open-ended as to whose actions exactly they are referring to (e. g., inefficient use of time; using time productively). We can draw a relatively well-founded distinction between items that clearly refer to teacher actions and items that clearly refer to student actions. In addition to these easily classifiable cases, there are items that do not spell out a specific referent but that allow the responder to easily infer to what or whom the items refer (e. g., "In math, the lesson is often disrupted," Göllner, Wagner, Rose et al., 2018, where responders will probably think of student misbehavior); there are also items without a clear referent that could be interpreted as referring to teachers, students, or interactions between both (e. g., "In math, a lot of time in class is wasted").

We found studies that use items referring only to student actions to operationalize classroom management (Kunter et al., 2013; Fauth et al., 2014b; Wagner et al., 2016; Wallace et al., 2016) and studies that use items referring only to teacher actions (de Jong & Westerhof, 2001; Wagner, Göllner, Helmke, Trautwein & Lüdtke, 2013; Mayr, Eder, Fartacek, Lenske & Pflanzl, 2013). Finally, there are studies that use both kinds of items, either combined in a single scale (Fauth et al., 2014a; Hochweber, Hosenfeld & Klieme, 2014) or separated in different scales (Clausen, 2002; Wettstein et al., 2016).

Interestingly, the studies mentioned above do not make a clear distinction between scales referring to the teacher's behavior and scales referring to the students' behavior. All of these different items are subsumed under the term classroom management. Additionally, studies tend to attribute well-managed classrooms to the teacher (e. g., Hochweber et al., 2014, p. 289). For example, the Tripod's *control* dimension (items referring only to student behavior) is meant to evaluate whether teachers "are able to manage the class in a way that teaching and learning occur efficiently, without being derailed by misbehavior or distractions" (Ferguson & Danielson, 2015, p. 106). As with the Tripod student survey, many studies use instruments that focus primarily on student behavior to measure classroom management. Oftentimes, students' discipline is the only

indicator of a teacher's classroom management (Fauth et al., 2014; Kunter et al., 2013; Wagner et al., 2016).

In the taxonomy presented here, we assume that the different item references described above would play a role in the relative agreement between different perspectives on teaching quality. In the following section, we explain our hypothesis that differences between self- and other-ratings play an important role in these rating mechanisms, in more detail.

3.2 Self- and Other-Ratings

The reference perspective matrix makes the assumption that the various item references described in the previous section can have an impact on item responses. Depending on who is responding to a certain item (students, teachers, or external observers), the same item can be an invitation to judge one's own behavior (e. g., a teacher's judgment of his/her own monitoring behavior) or another person's behavior (a student's judgment of the teacher's monitoring). The item reference, in combination with the identity of the person answering the item, determines whether an item represents a self- or an other-rating. We assume that this is a crucial issue in the assessment of teaching quality. In the following section, we further outline how the distinction between self- and other-ratings may be highly relevant for understanding perspective-specific ratings of teaching quality.

Think about an item like "Our teacher immediately notices when students start doing something else" (Baumert et al., 2009, p. 211). A student who has to judge this item will have to rely on indirect behavioral indicators that a teacher has noticed something (e. g., the teacher steps nearer to a student who is seemingly not paying attention, or the teacher starts staring at the student while continuing to speak), which the student would have to interpret correctly (see the realistic accuracy model of personality judgement: Funder, 1995). When the same item is answered from the teacher's perspective, we instantly notice significant differences. First, the teacher has privileged access to his/her own thoughts and thus is directly privy to his/her own noticing of something (although the teacher may find it difficult to judge whether he/she *immediately* notices when the students start doing something else). Second, teachers will be much less prone to error than students or external observers, who have to interpret a certain teacher behavior as an indicator of noticing students' attention behaviors. The students, in contrast, are limited to drawing inferences from the teacher's behavior and to interpreting overt behavior, to determine whether this indicates noticing, on the part of the teacher. External observers are in a similar position to students. However, they will have very specific indicators for a teacher noticing something in their rating manual. This will make their ratings more reliable. However, it is very unlikely that these indicators would be the same ones that students use. Teachers' thoughts are hard to read, even for trained observers.

3.3 Information and Motivation as Central Dimensions of Differences Between Self- and Other-Ratings

Such differences between self- and other-ratings also form the foundation of a theoretical model that has been developed in the field of personality research: the SOKA model (self-other knowledge asymmetry) of Vazire (2010). This model has been very influential in personality and social psychology, and has proven to be a powerful framework when it comes to explaining differences between self- and other-ratings in personality research. In the following discussion, we apply this model to our research on different perspectives on teaching quality.

The SOKA model assumes two major asymmetries between self- and other-ratings: (1) the quality and the quantity of information that is available and salient for a rater ("informational difference in perspective"), and (2) the degree of ego involvement that goes along with a rating ("motivational significance"; Vazire, 2010, p. 283). Regarding ego involvement, Vazire (2010) states that "judges have a lot more at stake when they are also the target than when they are judging someone else" (p. 284).

By considering the two aspects of information and motivation, this approach takes into account that "human perceivers act as both intuitive scientists and intuitive politicians – their judgments are influenced by both 'cold' information-processing goals (i. e., understanding and predicting the actor's behavior) and by 'hot' motivational goals (i. e., protecting or enhancing their own self-worth)" (Vazire, 2010, p. 283).

In our example of a teacher noticing whether students are paying attention, we have discussed the informational asymmetries between the teacher's and students' perspectives. What about the second asymmetry, of difference in motivation? Teachers certainly have a professional ethos that highlights that noticing what students do is good and necessary for teachers. So responding to this item will somehow activate the "intuitive politician" (Vazire, 2010) who is motivated to protect his or her self-worth. Additionally, teachers will differ in the importance they place on noticing student actions, and thus they will differ in how strong their intuitive politician is. Students will probably not be as interested in protecting their teachers' self-worth – they have less at stake in this evaluation. However, as described above, if they want to answer this item honestly they face another severe problem: They lack relevant *information.* In fact, the double asymmetry of *motivation* and *information* makes deviations between the three perspectives more than plausible.

Let us have a look at an item already discussed above: "The students in this class mess around a lot" (item from PISA 2003 assessment; see Kunter & Baumert, 2006, p. 245). The asymmetries in information and motivation will probably be less important in responses to this item from different perspectives. The *information* available will not be very different for the different perspectives. Students messing around do not refer to someone's thoughts, but to openly displayed behavior. Concerning *ego involvement,* teachers have much less at stake when student behavior is under scrutiny (in fact, it might even be self-protective for a teacher to agree with this item – see our discussion of this item example above). Students – who are rating their own behavior in this case –

are probably not explicitly motivated to answer the item in a certain way either. That is because this item is not a pure self-rating but a combination of self- and other-ratings. The item leaves open the degree to which an individual student is responsible for the messing around. This phenomenon will be very common in items referring to students. Many surveys contain items that refer to individual students ("I-form") as well as to the whole class ("We-form," according to Sirotnik, 1980; see also Den Brok, Brekelmans & Wubbels, 2006; Wagner et al., 2013). To assess items that focus on student behavior and that are rated by students, one should therefore always take into account whether the item asks respondents to assess their individual student behavior or to assess class behavior as a whole.

Thus, we have identified differences in information and ego involvement as two central factors that operate in a perspective-specific way rating teaching quality. Regarding the reference perspective matrix, we now have a better idea of why students, teachers, and external observers might disagree in their ratings of teaching quality. Different processes are at work depending on whether an item represents a self- or an other-rating. These different processes can be explained by differences in the information available to raters and how the information is used to respond to an item. To make predictions about the accuracy of a rater's response to an item, it is necessary to begin by identifying the cell where the item is located from a certain perspective, and then to evaluate the extent of information available to answer this item, and the degree of ego involvement that could motivate the rater to answer in a certain way.

4. Applying the Reference Perspective Matrix to Previous Findings

In this section, we revisit previous studies and reinterpret their results in light of the reference perspective matrix. The main questions in this section are: Does the item reference really make a difference? And second: Can these differences between perspectives be explained by informational and motivational differences in self- and other ratings?

4.1 Factorial Structure in Student Ratings

In Fauth et al. (2014a) the factorial structure of primary school students' teaching quality ratings were examined. The expected three factors could be distinguished relatively well. However, the model fit might have been artificially inflated, due to the fact that one factor consisted only of items referring to student behavior (classroom management), whereas all the items of the other two factors referred to the teacher. The comparably low correlations between the classroom management factor and the two other factors are in line with this interpretation (Fauth et al., 2014b, p. 6).

In a recent study on the Tripod student survey by Wallace et al. (2016), the authors presented a bi-factor model of teaching quality ratings, with one general factor representing all items and one specific factor in classroom management. Their interpretation

suggested that there was a general teaching competence and an additional specific classroom management competence. However, taking a closer look at the items, it turns out that the items on classroom management (the control dimension of Tripod) are also the ones that refer to students' actions (e. g., "Student behavior in this class is a problem"), whereas almost all of the other items[1] refer to the teacher (e. g., "My teacher explains difficult things clearly"). Thus, at this point we do not know whether the distinctive aspect of the specific classroom management factor is the substantive focus on classroom management or the reference to students' behavior rather than teacher behavior. A similar factor structure emerged in the analyses of Schweig (2014), who considered schools, as level-2 units, rather than classes.

4.2 Correlations Between Perspectives

These studies indicate that item references make a difference within the perspective of student ratings. When we take into account the reference perspective matrix and the considerations of informational and motivational differences, we would additionally predict higher between-perspective correlations for those scales that refer to student behavior (see the extensive discussion above on the "students messing around" items).

In Kunter and Baumert (2006), there was one factor (classroom management) that emerged in both the students' and in the teachers' responses (while all other items revealed a different factor structure in both perspectives). Again, the items on classroom management were the ones that referred to the students, while all other items focused on teacher behavior. In accordance with our assumptions, it was also this factor that showed the highest correlations among the few significant relationships between perspectives in this study (latent correlation of 0.64, see Kunter & Baumert, 2006, p. 240). This pattern was confirmed in Wagner et al. (2016) and Fauth et al. (2014a), where the highest correlations between teacher and student ratings were again found for classroom management measured with items referring to student behavior.

However, in the four studies mentioned above, the relationship between item reference and substantive focus on classroom management was confounded. That is, classroom management was measured using items referring to the students, whereas all other factors were measured with items referring to the teacher. In Wettstein et al. (2016), the authors included three scales with items referring to students (e. g., "Some students don't really listen to the teacher") and one scale referring to the teacher (e. g., "The teacher notices when students are not on task"). Correlations between teacher and student ratings were only found in the three scales that referred to students. No correlation was found in the scale referring to the teacher. Again, these results are in line with the assumption of different informational and motivational processes in different cells of the reference perspective matrix. Wettstein et al. (2016) also examined the correlations between student ratings of two different teachers teaching the same class. The results revealed a simi-

1 Four out of the remaining 29 items had a student referent.

lar pattern: No correlation between ratings of the two teachers on the scale referring to teacher behavior, but high correlations in scales referring to student behavior.

4.3 Different Item References for Different Perspectives

Clausen (2002) is another example of a study that considered items with both referents – teacher behavior (in the form of teachers' *monitoring*) and student behavior (in the form of students' *discipline*). Consistently with the results of Wettstein et al. (2016), the author found no significant correlations between ratings of teachers' monitoring behavior. In the case of discipline, only student and observer ratings were correlated. A closer look at the items used for the teacher scale of discipline reveals, however, that these items actually do not refer to student behavior but rather to the teacher (e. g., "Right in the beginning of a new course I explain the rules that students have to stick to in round terms"; Clausen, 2002, p. 219). Thus, regarding the perspective reference matrix, we have a comparably high correlation between those scales with the same reference (same column in the matrix) and no significant correlations between scales with different item references (different columns in the matrix; see Fig. 1).

In summary, the results of the studies reviewed above provide strong evidence that item reference really matters. Additionally, we have found indications that differences between self- and other-ratings have important implications for responses to measures of teaching quality. Certainly, the results presented above indicate that we should pay more attention to these issues in future research on the assessment of teaching quality.

5. Limitations and Future Research

In the present contribution, we started with the question of how to explain the low agreement between different rater perspectives in teaching quality assessments. With the reference perspective matrix developed here, we now offer more general considerations of how teaching quality can be best assessed. This contribution might thus be helpful when it comes to developing high-quality survey and observation instruments in the future.

The presented matrix provides a taxonomy that can serve as a heuristic for examining survey items and that may also be helpful in explaining results from previous studies. However, the explanations given above are as yet post hoc hypotheses. To properly validate the assumptions made above, we would need strong, preferably experimental research designs. One possibility would be to systematically manipulate assessment items regarding their item reference (teacher/student behavior) but also in regard to the differences in information and motivation that a certain item from a certain perspective is likely to trigger. These factors would not be easy to manipulate in an isolated way, but we believe that the effort invested in this endeavor would be worth it.

Another possibility would be to make more use of recent approaches to video observation data. For instance, the 'advocatory' approach proposed by Oser, Curcio &

Düggeli (2007) explicitly takes into account multiple perspectives on teaching as well as a combination of self- and other-ratings. In this approach, the competence of a teacher is not being evaluated with very broad items trying to capture what is going in the classroom *in general.* Instead, teachers are invited to judge very specific situations in videos of another teacher's instruction. The quality of these judgements can then serve as indicators of the observing teacher's competence. Although this approach targets teacher competence rather more than actual teaching quality, we believe that such innovations could also be helpful in addressing some of the issues raised in this paper.

The present contribution has limited application of the reference perspective matrix to the field of classroom management. Whether the matrix will also be applicable to other constructs of teaching quality such as cognitive activation and individual support, is as yet unclear. Idiosyncratic perceptions play a more important role in such constructs (Göllner, Wagner, Eccles & Trautwein, 2018). We have reason to believe that the reference perspective matrix will also prove helpful to understanding perspective-specific ratings in these areas, where perspective-specific differences in information and ego-involvement will also likely be crucial factors. This assumption will have to be examined in future contributions.

References

Baumert, J., Blum, W., Brunner, M., Dubberke, T., Jordan, A., Klusmann, U., Krauss, S., Kunter, M., Löwen, K., Neubrand, M., & Tsai, Y.-M. (2009). *Professionswissen von Lehrkräften, kognitiv aktivierender Mathematikunterricht und die Entwicklung von mathematischer Kompetenz (COACTIV): Dokumentation der Erhebungsinstrumente* (Materialien aus der Bildungsforschnung Nr. 83). Berlin: Max-Planck-Institut für Bildungsforschung.

Camburn, E., & Barnes, C.A. (2004). Assessing the validity of a language arts instruction log through triangulation. *The Elementary School Journal, 105,* 49–73.

Chaplin, D., Gill, B., Thompkins, A., & Miller, H. (2014). *Professional practice, student surveys, and value-added: Multiple measures of teacher effectiveness in the Pittsburgh public schools.* Regional Educational Laboratory Mid-Atlantic.

Clausen, M. (2002). *Unterrichtsqualität: Eine Frage der Perspektive?* Münster: Waxmann.

Clausen, M., Schnabel, K., & Schröder, S. (2002). Konstrukte der Unterrichtsqualitat im Expertenurteil. *Unterrichtswissenschaft, 30,* 246–260.

Danielson, C. (2007). *Enhancing professional practice: A framework for teaching.* Alexandria, VA: ASCD.

De Jong, R., & Westerhof, K.J. (2001). The quality of student ratings of teacher behaviour. *Learning Environments Research, 4,* 51–85.

Den Brok, P., Brekelmans, M., & Wubbels, T. (2006). Multilevel issues in research using students' perceptions of learning environments. *Learning Environments Research, 9,* 199–213.

Desimone, L.M., Smith, T.M., & Frisvold, D.E. (2010). Survey measures of classroom instruction comparing student and teacher reports. *Educational Policy, 24,* 267–329.

Doyle, W. (2013). Ecological approaches to classroom management. In C.M. Evertson & C.S. Weinstein (Eds.), *Handbook of classroom management* (pp. 107–136). Routledge.

Fauth, B., Wagner, W., Bertram, C., Göllner, R., Roloff-Bruchmann, J., Lüdtke, O., Polikoff, M., Klusmann, U., & Trautwein, U. (in press). Don't blame the teacher? The need to account for classrooms characteristics in evaluations of teaching quality. *Journal of Educational Psychology.*

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

Fauth, B., Decristan, J., Rieser, S., Klieme, E., & Büttner, G. (2014b). Grundschulunterricht aus Schüler-, Lehrer- und Beobachterperspektive: Zusammenhänge und Vorhersage von Lernerfolg. *Zeitschrift für Pädagogische Psychologie, 28,* 127–137.

Ferguson, R. (2010). *Student perceptions of teaching effectiveness.* Boston, MA: Harvard University.

Ferguson, R.F., & Danielson, C. (2014). How framework for teaching and tripod 7Cs evidence distinguish key components of effective teaching. In T.J. Kane, K.A. Kerr, & R.C. Pianta (Eds.), *Designing teacher evaluation systems* (pp. 98–143). San Francisco, CA: John Wiley & Sons.

Funder, D.C. (1995). On the accuracy of personality judgment: A realistic approach. *Psychological Review, 102,* 652–670.

Gitomer, D.H., & Bell, C.A. (2013). Evaluating teaching and teachers. In K.F. Geisinger (Ed.), *APA Handbook of Testing and Assessment in Psychology* (pp. 415–444). Washington, DC: American Psychological Association.

Gitomer, D., Bell, C., Qi, Y., McCaffrey, D., Hamre, B.K., & Pianta, R.C. (2014). The instructional challenge in improving teaching quality: Lessons from a classroom observation protocol. *Teachers College Record, 116*(6), 1–32.

Göllner, R., Wagner, W., Eccles, J.S., & Trautwein, U. (2018). Students' idiosyncratic perceptions of teaching quality in mathematics: A result of rater tendency alone or an expression of dyadic effects between students and teachers? *Journal of Educational Psychology, 110*(5), 709–725.

Göllner, R., Wagner, W., Rose, N., Fauth, B., & Nagengast, B. (2018). Students' perceptions of teaching quality in mathematics: An integrated data analysis of five large-scale assessments. Manuscript submitted for publication.

Graumann, C.F. (1960). *Grundlagen einer Phänomenologie und Psychologie der Perspektivität.* Berlin: de Gruyter.

Hamre, B.K. (2008). *My areas of strength and growth.* Unpublished manuscript, University of Virginia, Charlottesville, VA.

Hattie, J. (2009). *Visible learning.* New York: Routledge.

Hochweber, J., Hosenfeld, I., & Klieme, E. (2014). Classroom composition, classroom management, and the relationship between student attributes and grades. *Journal of Educational Psychology, 106,* 289–300.

Kane, T.J., Staiger, D.O., McCaffrey, D., Cantrell, S., Archer, J., Buhayar, S., Kerr, K., Kawakita, T., & Parker, D. (2012). *Gathering feedback for teaching: Combining high-quality observations with student surveys and achievement gains.* Seattle, WA: Bill & Melinda Gates Foundation.

Kaufman, J.H., Stein, M.K., & Junker, B. (2016). Factors associated with alignment between teacher survey reports and classroom observation ratings of mathematics instruction. *The Elementary School Journal, 116,* 339–364.

Klieme, E. (2006). Empirische Unterrichtsforschung: aktuelle Entwicklungen, theoretische Grundlagen und fachspezifische Befunde. *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: Waxmann.

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

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.

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

Lipowsky, F., Rakoczy, K., Pauli, C., Drollinger-Vetter, B., Klieme, E., & Reusser, K. (2009). Quality of geometry instruction and its short-term impact on students' understanding of the Pythagorean theorem. *Learning and Instruction, 19,* 527–537.

Mayer, D. P. (1999). Measuring instructional practice. *Educational Evaluation and Policy Analysis, 21,* 29–45.

Mayr, J., Eder, F., Fartacek, W., Lenske, G., & Pflanzl, B. (2013). *Linzer Diagnosebogen zur Klassenführung. Version LDK-P-WP.* Alpen-Adria-Universität Klagenfurt.

Oser, F., Curcio, G. P., & Düggeli, A. (2007). Kompetenzmessung in der Lehrerbildung als Notwendigkeit – Fragen und Zugänge. *Beiträge zur Lehrerinnen- und Lehrerbildung, 25,* 14–26.

Pianta, R. C., La Paro, K., & Hamre, B. K. (2008). *Classroom Assessment Scoring System (CLASS).* Baltimore: Paul H. Brookes.

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

Praetorius, A. K., Klieme, E., Herbert, B., & Pinger, P. (2018). Generic dimensions of teaching quality: The German framework of Three Basic Dimensions. *ZDM, 50,* 407–426.

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

Schweig, J. (2014). Cross-level measurement invariance in school and classroom environment surveys: Implications for policy and practice. *Educational Evaluation and Policy Analysis, 36,* 259–280.

Schweig, J. D. (2016). Moving beyond means: Revealing features of the learning environment by investigating the consensus among student ratings. *Learning Environments Research, 19,* 441–462.

Sirotnik, K. A. (1980). Psychometric implications of the unit-of-analysis problem. *Journal of Educational Measurement, 17*(4), 245–282.

Vazire, S. (2010). Who knows what about a person? The self–other knowledge asymmetry (SOKA) model. *Journal of Personality and Social Psychology, 98,* 281–300.

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

Wagner, W., Göllner, R., Werth, S., Voss, T., Schmitz, B., & Trautwein, U. (2016). Student and teacher ratings of instructional quality: Consistency of ratings over time, agreement, and predictive power. *Journal of Educational Psychology, 108,* 705–721.

Wallace, T. L., Kelcey, B., & Ruzek, E. (2016). What can student perception surveys tell us about teaching? Empirically testing the underlying structure of the tripod student perception survey. *American Educational Research Journal, 53,* 1834–1868.

Wettstein, A., Ramseier, E., Scherzinger, M., & Gasser, L. (2016). Unterrichtsstörungen aus Lehrer- und Schülersicht. *Zeitschrift für Entwicklungspsychologie und Pädagogische Psychologie, 48,* 171–183.

Zusammenfassung: Die Urteile von Schüler*innen, Lehrkräften und externen Beobachter*innen zur Unterrichtsqualität stimmen häufig nur in geringem Maße überein. In dem vorliegenden konzeptuellen Beitrag geben wir einen Überblick über bisherige empirische Befunde zu Perspektivenvergleichen und versuchen diese Befunde durch eine Analyse der eingesetzten Erhebungsinstrumente zu erklären. Darauf aufbauend skizzieren wir eine Referenten-Perspektiven-Matrix zur Klassifikation existierender Fragebogen- und Beobachtungsitems. Diese kann zur Erklärung der perspektivenspezifischen Mechanismen beitragen, welche der Beantwortung von Items zugrunde liegen. Die vorgestellte Matrix bietet damit auch eine Grundlage für künftige Arbeiten zur Erfassung von Unterrichtsqualität.

Schlagworte: Unterrichtsqualität, Messung, Validität, Klassenführung, Perspektivenvergleich

Contact

Prof. Dr. Benjamin Fauth, Institut für Bildungsanalysen Baden-Württemberg (IBBW),
Heilbronner Str. 172, 70191 Stuttgart, Germany
E-Mail: benjamin.fauth@ibbw.kv.bwl.de

Prof. Dr. Richard Göllner, University of Tübingen,
Hector Research Institute of Education Sciences and Psychology,
Europastraße 6, 72072 Tübingen, Germany
E-Mail: richard.goellner@uni-tuebingen.de

Prof. Dr. Gerlinde Lenske, University of Koblenz-Landau,
Institute of Primary Education,
Fortstraße 7, 76829 Landau, Germany
E-Mail: lenske@uni-landau.de

Prof. Dr. Anna-Katharina Praetorius, University of Zurich,
Institute of Education,
Freiestrasse 36, 8032 Zurich, Switzerland
E-Mail: anna.praetorius@ife.uzh.ch

Dr. Wolfgang Wagner, University of Tübingen,
Hector Research Institute of Education Sciences and Psychology,
Europastraße 6, 72072 Tübingen, Germany
E-Mail: wolfgang.wagner@uni-tuebingen.de