Chapter 7
Student Ratings of Teaching Quality Dimensions: Empirical Findings and Future Directions

Richard Göllner, Benjamin Fauth, and Wolfgang Wagner

Abstract This chapter discusses current issues in research on the validity of student ratings of teaching quality. We first discuss the advantages and limitations of student ratings of teaching quality based on theoretical considerations regarding the teaching quality concept. Research reveals that the validity of student ratings differs depending on the aspect of teaching quality being rated (i.e., classroom management, cognitive activation, or student support). Extending this research, we propose that future studies on the validity of student ratings should take into account students' cognitive processing while responding to survey items. We discuss three areas that seem promising for future research: the complexity and comprehensibility of survey items, the referent and addressee of items, and finally, the idiosyncratic nature of student ratings.

Keywords Student ratings · Teaching quality · Dimensions · Validity · Theoretical considerations

1 Introduction

Assuring reliable and valid measures is a key issue in assessing teaching quality in schools or classrooms for evaluative purposes. In general, student ratings represent a promising way to evaluate teaching because they provide firsthand impressions and are more efficient in assessing teaching quality than alternatives such as classroom observations. On the other hand, however, scholars have expressed concerns about

R. Göllner (✉) · W. Wagner
Hector Research Institute of Education Sciences and Psychology, University of Tübingen, Tübingen, Germany
e-mail: richard.goellner@uni-tuebingen.de

W. Wagner
e-mail: wolfgang.wagner@uni-tuebingen.de

B. Fauth
Institute for Educational Analysis (IBBW), Stuttgart, Germany
e-mail: Benjamin.Fauth@ibbw.kv.bwl.de

W. Rollett et al. (eds.), *Student Feedback on Teaching in Schools*,
https://doi.org/10.1007/978-3-030-75150-0_7

students' ability to provide reliable and valid information about teaching quality. In the following chapter, we first describe a common framework of teaching quality and then present recent findings on the differential validity of student ratings for conceptually different aspects of teaching quality. Finally, we show that the way in which students are asked about teaching quality in surveys raises awareness of the potential and limitations of student ratings and can help us identify existing gaps in the field of teaching quality research.

2 The Concept of Teaching Quality

Teaching quality is widely understood as rooted in a teacher's actual behavior, but it is also influenced by student–teacher interactions (Doyle, 2013; Fauth et al., 2020b; Göllner et al., 2020; Hamre & Pianta, 2010; Kunter et al., 2013). Thus, conceptually, teaching quality refers to teacher behavior in the classroom as well as students' reactions to this behavior and vice versa. One implication of this is that the context and conditions in which teaching takes place always need to be considered. Teaching quality has been described and assessed in a number of different frameworks, many of which show a great deal of overlap (e.g., Creemers & Kyriakides, 2008; Danielson, 2007; Pianta et al., 2008). A very common conception of teaching quality subdivides it into three superordinate quality domains, namely classroom management, teachers' learning support, and cognitive activation (see Hamre & Pianta, 2010; Praetorius et al., 2018). Classroom management has traditionally been seen as a central element of good teaching and has an important place in many conceptualizations of teaching quality. Relevant characteristics include a lack of student misbehavior and effective management of time and classroom routines (Evertson & Weinstein, 2006). Student support is based on a positive student–teacher relationship and a learning environment in which, for example, students are given constructive feedback on how to improve their performance or see the subject matter as more relevant (Brophy, 2000). Finally, cognitive activation encompasses, for example, providing challenging tasks that clarify the connection between different concepts or link new learning content to prior knowledge (e.g., Kunter et al., 2013). These aspects of quality have received substantial empirical attention in recent years. Most importantly for the present chapter, they serve as the foundation for survey instruments and observation protocols that can then be used to examine the empirical relevance of teaching quality for students' achievement and learning-related outcomes (e.g., students' interest, motivation, self-efficacy; e.g., Kunter et al., 2013).

3 Why Should Student Ratings Be Used to Assess Teaching Quality?

Teaching quality—in terms of teachers' classroom management, the support teachers provide to students, or the extent to which learning is cognitively demanding—can be assessed in different ways, each of which entails a number of advantages and disadvantages (Derry et al., 2010; Desimone et al., 2010; Fraser & Walberg, 1991; Wubbels et al., 1992). For instance, classroom observations are viewed as the gold standard in teaching quality research. They are considered the most objective method of measuring teaching practices and represent a central element in teacher training (Pianta et al., 2008). On the other hand, it is widely recognized that classroom observation is not without problems. Observers need to be specially trained, their observations provide only snapshots, and it is unclear whether the presence of observers systematically changes the behavior of teachers and students (e.g., Derry et al., 2010).

In contrast to classroom observations, student ratings of teaching quality are much easier to obtain. They are considered to be more cost effective, and they are directly tied to students' day-to-day classroom experiences. Moreover, they are not merely the result of a single or quite limited number of observations, and they ensure a reliable assessment of teaching quality (Lüdtke et al., 2009). Research has shown that the psychometric properties of a class' average teaching quality perceptions are not systematically inferior to those from observational measures (e.g., Clausen, 2002; de Jong & Westerhof, 2001; Maulana & Helms-Lorenz, 2016). In addition, there is empirical evidence that students are able to provide valid ratings of teaching quality, although differences between quality dimensions need to be taken into account (Fauth et al., 2014; Kuhfeld, 2017; Nelson et al., 2014; Schweig, 2014; Wagner et al., 2013; Wallace et al., 2016; see also Chap. 5 by van der Lans in this volume). Specifically, previous research has shown that student ratings of classroom management typically emerge as a clearly identifiable teaching quality aspect, which exhibits significant associations with observational as well as teacher self-report data and predicts students' learning in terms of their achievement, interest, and motivation (e.g., Kunter et al., 2007; Lipowsky et al., 2009). Furthermore, student ratings of classroom management are comparable across different learning contexts (e.g., different school subjects; Wagner et al., 2013) and even reveal time-specificity. That is, student ratings have proven to be sensitive enough to capture differences in teachers' classroom management over the course of several weeks or months (Wagner et al., 2016). In contrast, the psychometric properties of student ratings of learning support and cognitive activation are less clear. In the case of cognitive activation, this is because measures cannot be generally applied to all subjects but need to reflect the specificity and requirements of each individual subject (e.g., mathematics, languages, the arts, etc.). Consequently, the majority of existing student surveys of teaching quality do not include cognitive activation measures, making it much harder to evaluate the validity of student ratings with respect to this dimension. Nevertheless, the few studies that do exist show that even ratings by primary school students reveal substantial differences in cognitive activation between classrooms. In addition, cognitive activation

ratings have been shown to be separable from classroom management ratings and to a lesser extent from learning support ratings, and to be statistically significant associations with student learning outcomes (e.g., subject-related interest; Fauth et al., 2014). The situation for student ratings of learning support is even more complex. Previous research has shown that student ratings of learning support exhibit relatively low agreement with classroom observations and even low agreement across students in the same classroom. One potential explanation for this is that students' perceptions of teachers' learning support do not exclusively function as a quality characteristic that differs across classrooms but are also affected by students' individual experiences within classrooms (Aldrup et al., 2018; Atlay et al., 2019; den Brok et al., 2006a; Göllner et al., 2018). For a long time, these within-classroom differences were considered the result of factors external to teaching quality, such as students' rating tendencies (e.g., harshness or leniency) or perceptual mindsets (e.g., halo error; e.g., Lance et al., 1994). However, recent research has shown that these differences can also reflect effects stemming from the dyadic relationships between each individual student and his or her teacher. Specifically, a recent study by Göllner and colleagues (2018) used national longitudinal data from the Program for International Student Assessment (PISA) database and showed that rating differences in student perceptions of learning support partially result from teacher-independent rater tendencies, but also reflect the dyadic relationship between an individual student and one specific teacher. Therefore, students' ratings of teaching quality provide important information about their individual experiences in their classroom learning environments.

4 Future Directions for the Use of Students' Ratings

Although student ratings of teaching quality have become a prominent way to obtain student feedback on teaching quality in schools and classrooms, scholars and practitioners have also criticized their use in both summative and formative assessments (Abrami et al., 2007; Benton & Cashin, 2012). They emphasize the specific nature of student ratings, as students are not trained to provide valid assessments of teaching quality in the same way as adult observers. Thus, it is important to acknowledge potential limitations of student ratings, which raises the question of how student ratings for evaluative purposes can be improved. We believe that a more detailed examination of existing survey instruments can be a fruitful approach to finding out how student ratings work and what we can do to achieve reliable and valid ratings. From a very general perspective, a student survey can be seen as ordinary text material (i.e., textual information presented in the form of separate items), requiring students to read and interpret a question to understand what is meant, retrieve the requested information from memory, and form a judgment based on their knowledge and expertise (Tourangeau et al., 2000). Building upon this foundation, this chapter presents three areas of recent research that might help provide a deeper understanding of students' teaching quality rating and exploit future research directions.

4.1 Complexity and Comprehensibility

At first glance, existing student surveys fundamentally differ in their linguistic complexity, which shapes student responses (e.g., Krosnick & Presser, 2010; Tourangeau et al., 2000). It is surprising to see that even frequently used surveys are linguistically challenging, particularly for younger respondents (e.g., Fauth et al., 2014; Wagner et al., 2013). Consequently, it can be argued that many reporting problems (i.e., low interrater agreement) arise because students encounter difficulties in comprehending the survey. Survey items include many linguistic features, including surface aspects (e.g., the length of words and sentences) and characteristics that require more linguistic analysis (e.g., the number of complex noun phrases). For example, the following items might be used to assess teachers' sensitivity to and awareness of students' level of academic functioning: "In math, the individual students often do different tasks" and "In math lessons, the teacher asks different questions, depending on how able the student is." However, the items differ in their linguistic characteristics: number of words (9 vs. 15), structure of sentences (1 vs. 2 clauses), average word length (5.33 characters vs. 5.00 characters), and number of complex noun phrases per clause (2 vs. 0.5). In addition, students may be less familiar with certain words used in the items (e.g., "individual," "depending") or have to make many interpretations because single words do not refer to specific, denotable, and relatively objective behavior (i.e., high-inference ratings; e.g., Roch et al., 2009; Rosenshine, 1970). Despite the large body of literature on traditional best practices in the construction of survey questions (see Krosnick & Presser, 2010), only a few studies have examined the impact of these and other linguistic characteristics on student surveys' ability to reliably and validly assess teaching quality. One of these studies showed that the use of measures with a lower specificity and higher level of abstraction (high-inference ratings) leads to higher interrater reliability in student ratings, but lower agreement with expert assessments. Contrary to common expectations, rater agreement increased as the behavioral observability of the measures decreased (Roch et al., 2009). The authors argue that raters might compensate for uncertainty in high-inference ratings by more strongly adjusting their ratings to match their general impression, which might in turn be unrelated or only partially related to the teaching quality dimension in question. Such findings impressively demonstrate that the association between linguistic features and psychometric properties of student ratings is anything but trivial, and a more rigorous consideration of linguistic forms in existing surveys is needed.

4.2 Framing

Student surveys also differ in characteristics apart from linguistic complexity. Specifically, the referent and addressee of survey items are two salient characteristics that might affect the information obtained from student ratings of teaching quality but

received less attention in research on student perceptions of teaching quality (den Brok et al., 2004, 2006b; McRobbie et al., 1998). The referent can be defined as the subject to which an item refers. At first glance, student rating items that refer more to the classroom (e.g., "In math class, the lesson is often disrupted") than to the teacher (e.g., "Our math teacher always knows exactly what is happening in class") tend to exhibit more favorable psychometric properties in terms of interrater agreement or distinctiveness from other theoretically relevant aspects of teaching quality (see Fauth et al., 2020a; Göllner et al., 2020). However, the use of surveys that refer more to the classroom than to the teacher might result in serious constraints. First, items referring more to the classroom than to the teacher are frequently used to assess classroom management, but much rarer for items assessing learning support or cognitive activation. This raises the question of whether the well-established distinctiveness of classroom management compared to other quality aspects is also due to systematic differences in the referent used. Second, previous findings have shown that when classroom management items refer to the classroom, measures are more prone to classroom composition effects (e.g., proportion of male students or performance composition). Even though existing analytical procedures can be used to account for such differences in classroom composition, it is unclear whether such analytical adjustments result in fair comparisons or relatively favor or penalize certain individual teachers. Irrespective of this, classroom management measures referring more to students than to the teacher need to be seen from an interactionist perspective that includes both teachers and students they teach (Fauth et al., 2020a). In addition, the target of the teacher's behavior that is addressed in a survey is important. In the simplest case, this can be either the responding student him/herself (e.g., "The teacher motivates me") or all students in the classroom (e.g., "The teacher motivates us"). An examination of existing surveys shows that the "me-addressee" is predominantly used when assessing the support teachers provide to students, whereas the "we-addressee" is more frequently used for classroom management and cognitive activation (e.g., BIJU, Baumert et al., 1996; Tripod survey; e.g., Prenzel et al., 2013; Wallace et al., 2016). At the same time, previous studies have shown that student support dimensions usually fail to predict student learning outcomes on the classroom level but are more consistent predictors at the individual student level (e.g., Aldrup et al., 2018). These results raise the question of whether support can be better conceptualized as a dyadic phenomenon between a teacher and an individual student or whether they merely reflect how teacher support is assessed. Experimentally varying the addressee for items assessing multiple teaching quality dimensions will enable us to examine whether the addressee affects the information obtained from student ratings at the student and classroom level. The findings might also be interesting for analytical modeling procedures used in teaching quality research. First, findings from multilevel models applied to separate students' shared (student level) and non-shared (classroom level) perceptions of teaching quality might be directly affected by the used addressee. Whereas the "me-addressee" assumed to provide valid information about students individual learning experiences at the student level, the "we-addressee" might be more adequate to give insights in students' learning at the classroom level; or in other words, one cannot simply assume that different

item wordings can interchangeably be used at different levels of analysis (den Brok, 2001). Second, there is increased interest in more recent analytical procedures that model classroom heterogeneity in student ratings as an additional indicator of good teaching (e.g., Schenke et al., 2018). Applying these modeling procedures to surveys with a "me-addressee" might be a better way to assess student–teacher fit in classrooms and teacher adaptivity than surveys with a "we-addressee." If surveys with a "we-addressee" are considered, different levels of heterogeneity between classes might be more a reflection of class-specific measurement precision (i.e., more or less agreement in classes). In other words, the choice of addressee in surveys can be assumed to have very serious consequences for teaching quality assessment and the view we take on students' learning in classrooms.

4.3 The Idiosyncratic Nature of Student Ratings

Finally, it is important to ask what we can fundamentally expect from student ratings of teaching quality and to what extent student ratings of teaching quality reveal idiosyncrasies, i.e., are systematically different from alternative methods. Even when we take special care to use comprehensible and age-appropriate surveys and make more intentional decisions about the referent and addressee in survey items, the specific nature of student ratings needs to be considered. One main objective of previous research has been to determine the degree of idiosyncrasy in student ratings by comparing them to alternative assessment methods (e.g., Clausen, 2002; Kunter & Baumert, 2006). This research has shown that student ratings, particularly those assessing learning support and cognitive activation, exhibit substantial differences to classroom observations or teacher self-report data, which might lead to the conclusion that students are less able to provide valid information on teaching quality and its theoretically proposed dimensions (e.g., Abrami et al., 2007). However, this focus on limitations and biases of student ratings bears the risk of neglecting the expertise students naturally acquire through their everyday experiences in classrooms. Thus, future research needs to better appreciate the unique information we obtain from student ratings (e.g., Leighton, 2019). In order to do so, however, we need to learn much more about the mental models that underlie students' ratings and the extent to which these models differ from those of adult observers evaluating teaching quality. A recent study by Jaekel et al., (2021) found that student ratings of teaching quality in one school subject (mathematics or German language) did not only result from students' daily experiences in the subject at hand, but were also affected by their experiences in the respective other subjects. Students seem to make use of comparative information when objective criteria for good teaching is not available. In addition, there is a need to understand how a developmental perspective can help us understand idiosyncrasies in student ratings of teaching quality. That is, it is reasonable to assume that student ratings of teaching quality are affected by the age-related developmental stages in which ratings take place. For instance, students' need to define their own identity and stronger need for autonomy during adolescence (e.g., Eccles et al., 1993)

might function as a guiding perspective when students have to rate teaching quality. A recent study by Wallace and colleagues (2016) based on the Tripod survey identified two dimensions of students' ratings of teaching quality: one specific classroom management factor and one broad general factor. Interestingly, the quality indicators with the highest loadings on the general factor were indicators that clearly capture students' perceptions of teachers' learning support and student–teacher relationship (Schweig, 2014; Wallace et al., 2016). The same is true for student ratings of cognitive activation. It is interesting to note that even though cognitive activation is considered a central aspect explaining students' achievement, cognitive activation measures are much less common in existing surveys than classroom management or learning support measures. One major reason for this is that assessing teachers' ability to use stimulating learning materials, the quality of questions teachers ask during lessons, or the quality of classroom discussion from students perspective is seen as a particularly challenging task because it requires special knowledge and skills which is beyond students' firsthand experiences of participation in the classroom. Whether and to what extent students are really able to provide information on these and other aspects of cognitive activation in line with an adult view remains an open question that needs to be addressed in future research. As part of this process, we have to think about further refining existing measures that capture central aspects of cognitive activation in a wide variety of learning situations and by making more explicit use of other principles getting learners to learn long, complex, and difficult things. Alternative ways of conceptualizing and measuring effective learning contexts from related disciplines (e.g., discourse analysis in linguistic research; Turner & Meyer, 2000) or entirely different research fields (e.g., game-based learning; Gee, 2007) can provide a good foundation for improving existing cognitive activation measures.

5 Closing Remarks

As the work we reviewed in this chapter makes clear, student ratings have become a vibrant part of teaching quality research. We are particularly excited about two aspects of this research. The first is the usefulness of student ratings in research and practice. Even though differences across teaching quality dimensions need to be considered, students can provide a valid perspective on teaching quality and are thus in no way generally inferior to alternative assessments such as classroom observations or teacher self-reports. Second, students provide a plethora of information on teaching quality at both the classroom and the student level, with the latter referring to students' individual learning experiences within a classroom in a way that is beyond the scope of alternative assessments. As research on student ratings progresses, it will be critical to take a deeper and more consequential look at the characteristics of existing surveys to determine what we can learn about teaching quality from the students' perspective. We look forward to participating in work on these topics in the future.

References

Abrami, P. C., D'Apollonia, S., & Rosenfield, S. (2007). The dimensionality of student ratings of instruction: What we know and what we do not. In R. P. Perry & J. C. Smart (Eds.), *The scholarship of teaching and learning in higher education: An evidence-based perspective* (pp. 385–456). Springer.

Aldrup, K., Klusmann, U., Lüdtke, O., Göllner, R., & Trautwein, U. (2018). Social support and classroom management are related to secondary students' general school adjustment: A multilevel structural equation model using student and teacher ratings. *Journal of Educational Psychology, 110,* 1066–1083. https://doi.org/10.1037/edu0000256.

Atlay, C., Tieben, N., Fauth, B., & Hillmert, S. (2019). The role of socioeconomic background and prior achievement for students' perception of teacher support. *British Journal of Sociology of Education, 40,* 970–991. https://doi.org/10.1080/01425692.2019.1642737.

Baumert, J., Roeder, P. M., Gruehn, S., Heyn, S., Köller, O., Rimmele, R., et al. (1996). Bildungsverläufe und psychosoziale Entwicklung im Jugendalter (BIJU) [Educational pathways and psychosocial development in adolescence]. In K.-P. Treumann, G. Neubauer, R. Moeller, & J. Abel (Eds.), *Methoden und Anwendungen empirischer pädagogischer Forschung* [Methods and applications of empirical educational research] (pp. 170–180). Waxmann.

Benton, S. L., & Cashin, W. E. (2012). *Student ratings of teaching: A summary of research and literature* (IDEA Paper No. 50). Manhattan: Kansas State University, Center for Faculty Evaluation and Development.

Brophy, J. (2000). *Teaching. Educational practices series, 1.* Brüssel: International Academy of Education (IAE).

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

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

Danielson, C. (2007). *Enhancing professional practice: A framework for teaching* (2nd ed.). ASCD.

de Jong, R., & Westerhof, K. J. (2001). The quality of student ratings of teacher behaviour. *Learning Environments Research, 4,* 51–85. https://doi.org/10.1023/A:1011402608575.

den Brok, P. (2001). *Teaching and student outcomes.* W. C. C.

den Brok, P., Brekelmans, M., & Wubbels, T. (2004). Interpersonal teacher behaviour and student outcomes. *School Effectiveness and School Improvement, 15,* 407–442. https://doi.org/10.1080/09243450512331383262.

den Brok, P., Brekelmans, M., & Wubbels, T. (2006a). Multilevel issues in research using students' perceptions of learning environments: The case of the Questionnaire on Teacher Interaction. *Learning Environments Research, 9,* 199–213. https://doi.org/10.1007/s10984-006-9013-9.

den Brok, P., Fisher, D., Rickards, T., & Bull, E. (2006b). Californian science students' perceptions of their classroom learning environments. *Educational Research and Evaluation, 12,* 3–25. https://doi.org/10.1080/13803610500392053.

Derry, S. J., Pea, R. D., Barron, B., Engle, R. A., Erickson, F., Goldman, R., Hall, R., Koschmann, T., Lemke, J. L., Sherin, M. G., & Sherin, B. L. (2010). Conducting video research in the learning sciences: Guidance on selection, analysis, technology, and ethics. *Journal of the Learning Sciences, 19*(1), 3–53. https://doi.org/10.1080/10508400903452884.

Desimone, L. M., Smith, T. M., & Frisvold, D. E. (2010). Survey measures of classroom instruction: Comparing student and teacher reports. *Educational Policy, 24*(2), 267–329. https://doi.org/10.1177/0895904808330173.

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

Eccles, J. S., Midgley, C., Wigfield, A., Buchanan, C. M., Reuman, D., Flanagan, C., & Mac Iver, D. (1993). Development during adolescence: The impact of stage–environment fit on adolescents' experiences in schools and families. *American Psychologist, 48,* 90–101.

Evertson, C. M., & Weinstein, C. S. (2006). *Handbook of classroom management: Research, practice, and contemporary issues*. Lawrence Erlbaum Associates Publishers.

Fauth, B., Decristan, J., Rieser, S., Klieme, E., & Büttner, G. (2014). Student ratings of teaching quality in primary school: Dimensions and prediction of student outcomes. *Learning and Instruction, 29,* 1–9. https://doi.org/10.1016/j.learninstruc.2013.07.001.

Fauth, B., Göllner, R., Lenske, G., Praetorius, A.-K., & Wagner, W. (2020a). Who sees what? Conceptual considerations on the measurement of teaching quality from different perspectives. *Zeitschrift Für Pädagogik, 66,* 138–155.

Fauth, B., Wagner, W., Bertram, C., Göllner, R., Roloff-Bruchmann, J., Lüdtke, O., Polikoff, M. S., Klusmann, U., & Trautwein, U. (2020b). Don't blame the teacher? The need to account for classroom characteristics in evaluations of teaching quality. *Journal of Educational Psychology, 112,* 1284–1302. https://doi.org/10.1037/edu0000416.

Fraser, B. J., & Walberg, H. J. (1991). *Educational environments: Evaluation, antecedents and consequences*. Pergamon Press.

Gee, J. P. (2007). *What video games have to teach us about learning and literacy* (2nd ed.). Palgrave Macmillan.

Göllner, R., Fauth, B., Lenske, G., Praetorius, A.-K., & Wagner, W. (2020). Do student ratings of classroom management tell us more about teachers or classrooms composition? *Zeitschrift Für Pädagogik, 66,* 156–172.

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,* 709–725. https://doi.org/10.1037/edu0000236.

Hamre, B. K., & Pianta, R. C. (2010). Classroom environments and developmental processes: Conceptualization and measurement. In J. Meece & J. Eccles (Eds.), *Handbook of research on schools, schooling, and human development* (pp. 25–41). Routledge.

Jaekel, A-K., Göllner, R., & Trautwein, U. (2021). How students' perceptions of teaching quality in one subject are impacted by the grades they receive in another subject—Dimensional comparisons in student evaluations of teaching quality. *Journal of Educational Psychology*. https://doi.org/10.1037/edu0000488.

Krosnick, J. A., & Presser, S. (2010). Questionnaire design. In J. D. Wright & P. V. Marsden (Eds.), *Handbook of survey research* (2nd ed.). Emerald Group.

Kuhfeld, M. (2017). When students grade their teachers: A validity analysis of the Tripod Student Survey. *Educational Assessment, 22*(4), 253–274. https://doi.org/10.1080/10627197.2017.1381555.

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. https://doi.org/10.1007/s10984-006-9015-7.

Kunter, M., Baumert, J., Blum, W., Klusmann, U., Krauss, S., & Neubrand, M. (Eds.). (2013). *Cognitive activation in the mathematics classroom and professional competence of teachers: Results from the COACTIV project*. Springer.

Kunter, M., Baumert, J., & Köller, O. (2007). Effective classroom management and the development of subject-related interest. *Learning and Instruction, 17,* 494–509. https://doi.org/10.1016/j.learninstruc.2007.09.002.

Lance, C. E., LaPointe, J. A., & Fisicaro, S. A. (1994). Tests of three causal models of halo rater error. *Organizational Behavior and Human Decision Processes, 57,* 83–96. https://doi.org/10.1006/obhd.1994.1005.

Leighton, J. P. (2019). Students' interpretation of formative assessment feedback: Three claims for why we know so little about something so important. *Journal of Educational Measurement, 56,* 793–814. https://doi.org/10.1111/jedm.12237.

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. https://doi.org/10.1016/j.learninstruc.2008.11.001.

Lüdtke, O., Robitzsch, A., Trautwein, U., & Kunter, M. (2009). Assessing the impact of learning environments: How to use student ratings in multilevel modelling. *Contemporary Educational Psychology, 34,* 123–131. https://doi.org/10.1016/j.cedpsych.2008.12.001.

Maulana, R., & Helms-Lorenz, M. (2016). Observations and student perceptions of the quality of preservice teachers' teaching behaviour: Construct representation and predictive quality. *Learning Environments Research, 19,* 335–357. https://doi.org/10.1007/s10984-016-9215-8.

McRobbie, C. J., Fisher, D. L., & Wong, A. F. L. (1998). Personal and class forms of classroom environment instruments. In B. J. Fraser & K. G. Tobin (Eds.), *International handbook of science education* (pp. 581–594). Kluwer.

Nelson, P. M., Demers, J. A., & Christ, T. J. (2014). The responsive environmental assessment for classroom teaching (REACT): The dimensionality of student perceptions of the instructional environment. *School Psychology Quarterly, 29,* 182–197. https://doi.org/10.1037/spq0000049.

Pianta, R. C., La Paro, K. M., & Hamre, B. K. (2008). *Classroom assessment scoring system (CLASS: PreK-3).* Brookes.

Praetorius, A.-K., Klieme, E., Herbert, B., & Pinger, P. (2018). Generic dimensions of teaching quality: The German framework of Three Basic Dimensions. *ZDM Mathematics Education, 50,* 407–426. https://doi.org/10.1007/s11858-018-0918-4.

Prenzel, M., Baumert, J., Blum, W., Lehmann, R., Leutner, D., Neubrand, M.,Pekrun, R., Rolff, H.-G., Rost, J., & Schiefele, U. (2013). *PISA-I-Plus 2003, 2004.* IQB—Institute for Educational Quality Improvement.

Roch, S. G., Paqin, A. R., & Littlejohn, T. W. (2009). Do raters agree more on observable items? *Human Performance, 22,* 391–409. https://doi.org/10.1080/08959280903248344.

Rosenshine, B. (1970). Evaluation of classroom instruction. *Review of Educational Research, 40,* 279–300. https://doi.org/10.3102/00346543040002279.

Schenke, K., Ruzek, E., Lam, A. C., Karabenick, S. A., & Eccles, J. S. (2018). To the means and beyond: Understanding variation in students' perceptions of teacher emotional support. *Learning and Instruction, 55,* 13–21. https://doi.org/10.1016/j.learninstruc.2018.02.003.

Schweig, J. (2014). Multilevel factor analysis by model segregation: New applications for robust test statistics. *Journal of Educational and Behavioral Statistics, 39*(5), 394–422. https://doi.org/10.3102/1076998614544784.

Tourangeau, R., & Rips, L. J., & Rasinski, K. (2000). *The psychology of survey response.* Cambridge University Press.

Turner, J. C., & Meyer, D. K. (2000). Studying and understanding the instructional contexts of classrooms: Using our past to forge our future. *Educational Psychologist, 35,* 69–85. https://doi.org/10.1207/S15326985EP3502_2.

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 domain-generalizability of domain-independent assessments. *Learning and Instruction, 104,* 148–163. https://doi.org/10.1016/j.learninstruc.2013.03.003.

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. https://doi.org/10.1037/edu0000075.

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. https://doi.org/10.3102/0002831216671864.

Wubbels, T., Brekelmans, M., & Hooymayers, H. P. (1992). Do teacher ideals distort the self-reports of their interpersonal behavior? *Teaching and Teacher Education, 8*(1), 47–58.

Richard Göllner is a Professor of Educational Effectiveness and Trajectories at the Hector Research Institute of Education Sciences and Psychology at the University of Tübingen (Germany). His work focuses on teaching quality, specifically on the measurement of instructional

practice from an interdisciplinary perspective, and its impact on students' achievement. Furthermore, he is interested in students' personality development within schools and the use of simulated learning contexts in experimental research in education.

Benjamin Fauth is Head of the Department for Empirical Educational Research at the Institute for Educational Analysis (IBBW) in Stuttgart (Germany) and Associate Professor at the Hector Research Institute of Education Sciences and Psychology at the University of Tübingen (Germany). His research focuses on the quality of teaching, in particular questions of the theoretical conceptualization, assessment, and the impact of teaching quality. Furthermore, his research focuses on the professional competence of teachers and on questions of applied evaluation research.

Wolfgang Wagner studied psychology at the University of Koblenz-Landau (Germany) and now works as a research assistant at the Hector Research Institute of Education Sciences and Psychology at the University of Tübingen (Germany). His main research interests include the assessment of characteristics of learning environments and their effects on the development of targeted outcomes (in particular, academic achievement), as well as methodological issues in the field of (multilevel) latent variable models.