Tosca Panetta[1,2]

Richard Göllner[1], Evelin Ruth-Herbein[2], Julia Maier[2], Ann-Kathrin Jaekel[1], Ulrich Trautwein[1], Benjamin Fauth[1,2]

[1]Hector Research Institute of Education Sciences and Psychology, University of Tübingen
[2]Institute for Educational Analysis Baden-Württemberg, Stuttgart

PhD student

Achieving high-quality observation ratings of teaching quality – investigation of a newly developed classroom observation instrument

Abstract (300 words)

Teaching quality is one of the central determinants of students' performance-related and motivational outcomes (e.g., Burroughs et al., 2019; Fauth et al., 2019; Hattie, 2009). Thus, educational systems need to provide reliable and valid teaching assessments for quality development in education. One of the central perspectives to gather feedback on teaching are classroom observations by external raters (e.g., Kane & Staiger, 2012). In Germany however, no classroom observation instrument has yet been established for the use in school practice. To fill this gap, Fauth et al. (2021) developed an eleven-item classroom observation form based on the framework of the three basic dimensions of teaching quality (Klieme et al., 2006; Praetorius et al., 2018). The observation form aims to assess the central aspects of teaching quality that have shown to be predictive for students' outcomes in previous studies (e.g., Lipowsky et al., 2009). In the present investigation, we evaluated the newly developed observation form in terms of psychometric quality in two studies. In the first study, $N = 10$ experienced mathematics teachers participated in a classroom observation training and rated short classroom video sequences at five different time points. We investigated consistency and accuracy of the observation ratings in the course of the training. In the second study, the 10 previously trained mathematics teachers rated $N = 34$ classroom videos from the Pythagoras study (Klieme et al., 2009). We investigated construct validity of the ratings. Overall, the results provide evidence that the newly developed observation form offers a successful assessment of teaching quality based on the three basic dimensions. Limitations of the present investigation and implications for educational research and school practice are discussed.

Extended summary (1000 words, excluding reference list) Include introduction, theoretical background, methods, aims, preliminary findings/findings, results, theoretical and education significance, relevance to the QUINT ambition and the reference list.

Theoretical Background

The question of what defines high-quality teaching is crucial in both educational research and school practice. Thus, teaching quality aspects have been widely investigated in the last decades. Different conceptualizations of the construct of teaching quality have been developed (Senden et al., 2021). Upon these conceptualizations, the framework of the three basic dimensions (Klieme et al., 2006; Praetorius et al., 2018), comprising cognitive activation, student support and classroom management, offers a useful approach to describe student-teacher interactions in the classroom and provides a suitable theoretical foundation for the assessment of teaching quality.

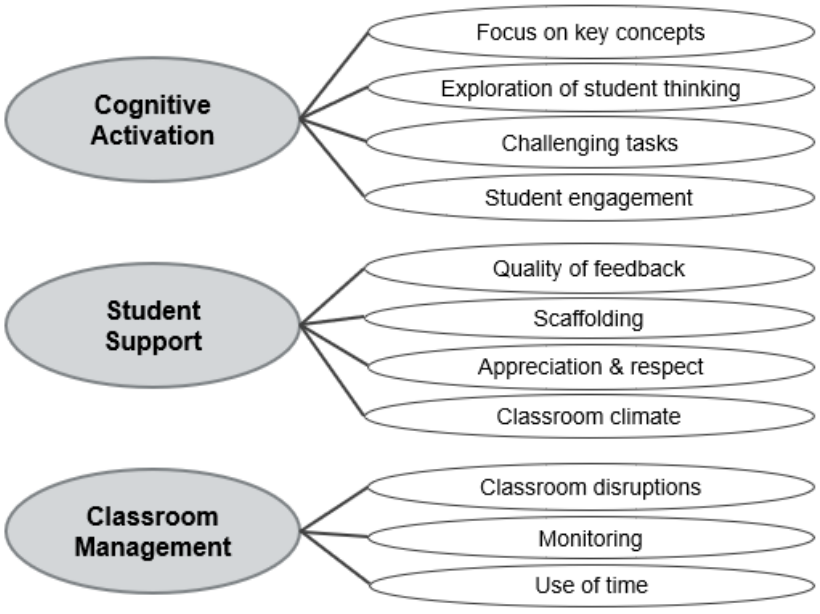
One widely used approach to assess teaching quality are classroom observations by external raters (e.g., Fauth et al., 2020). A variety of classroom observation systems already exists (Bell et al., 2019). However, approaches to assess teaching quality in German-speaking countries are either very

extensive or lack scientific foundation and have not been validated in empirical studies. So far, no scientific classroom observation instrument suitable for the use in school practice has yet been established. A newly developed observation form (Fauth et al., 2021) seeks to fill this gap.

The eleven-item classroom observation form is based on the framework of the three basic dimensions of teaching quality (Klieme et al., 2006; Praetorius et al., 2018). The instrument consists of the observation form with eleven single items and an accompanying observation manual. The observation manual comprises an introduction to the theoretical foundations underpinning the three basic dimensions and an introduction to how to use the instrument during classroom observations. Additionally, the manual contains a short description of the theoretical background to each item as well as a set of observable indicators that individuals rating this item will have to consider. With its eleven items, the observation form does not claim to cover the construct of teaching quality in a comprehensive way, but focuses on central teaching quality aspects that proved to be predictive for students' learning outcomes (e.g., Fauth et al., 2014; Lipowsky et al., 2009; Wagner et al., 2016). As teaching quality has proven to be co-constructed by teachers and students (e.g., Göllner et al., 2020), the observation form contains items focusing teachers' behavior (e.g., exploration of student thinking) and items focusing students' behavior (e.g., student engagement). Figure 1 gives an overview of the eleven teaching aspects assessed by the observation form and their assignment to the three basic dimensions.

Figure 1

Teaching quality aspects assessed by the observation form (adapted from Fauth et al., 2021)



Like every newly developed instrument, the observation form must undergo a testing of its psychometric quality prior to its practical use. The aim of this investigation is to examine the psychometric quality of the observation form based on two studies. In the first study, we investigated consistency and accuracy of the observation ratings by teachers in the course of a classroom observation training. We examined differences between items focusing teachers' behavior and items focusing students' behavior. In the second study, we investigated the construct validity of the ratings. We examined if validity shows with ratings of another observation instrument based on the three basic dimensions.

Methods

The first study was embedded in a classroom observation training for *N* = 10 experienced mathematics teachers from the German state of Baden-Württemberg. At five time points, the teachers rated short sequences of classroom videos of the TALIS video study (Grünkorn et al., 2020). Consistency and accuracy of the ratings were determined by interrater-agreement and interrater-reliability (LeBreton & Senter, 2008). Interrater-agreement was estimated using the "average absolute deviation index" (AD_M; Burke et al., 1999). Interrater-reliability was estimated using different types of intraclass-correlations (ICC1 and ICC2; LeBreton & Senter, 2008; Lüdtke et al., 2009).

In the second study, the previously trained teachers rated 45-minute classroom videos (*N* = 34) of the Pythagoras study (Klieme et al., 2009). For the investigation of construct validity, we investigated convergent and discriminant correlations with observation ratings of the classroom observation instrument from the Pythagoras study (Rakoczy & Pauli, 2006), using a multitrait-multimethod approach (Campbell & Fiske, 1959).

Results

Results of the first study indicated a satisfactory agreement between teachers in teaching quality ratings after the rater training (AD_M = 0.20 - 0.61). The group mean of the teachers was a reliable measure of the teaching aspects assessed (ICC2 = .72 - .95). Differences in interrater-agreement and interrater-reliability occurred between items focusing students' behavior and items focusing teachers' behavior, where student-focused items were rated more consistently.

In the second study, results of the multitrait-multimethod analysis provided evidence for convergent validity. We found high convergent correlations with the observation ratings of the three basic dimensions assessed by the observation instrument from the Pythagoras study (*r* = .55 - .67). Intercorrelations between different teaching quality aspects indicated limited discriminant validity.

Theoretical and Education Significance

The newly developed observation form provides a useful approach for the assessment of teaching quality by classroom observation in school practice. In the future, additional material (e.g., classroom videos) will enrich the generic observation form with subject-specific demonstrations of the teaching quality aspects assessed. The observation form can be used for formative feedback or self-reflection on teaching quality but also in teacher education.

Relevance to the QUINT Ambition

As a new instrument for the assessment of teaching quality, the development and evaluation of the observation form shows a great overlap with the topics addressed in the QUINT conference "Theorizing and Measuring Teaching Quality: Instruments, Evidence and Interpretations". As classroom videos will be developed to provide a subject-specific demonstration for the generic items that will be part of rater trainings, this project fits all four QUINT research themes.

References

Bell, C. A., Dobbelaer, M. J., Klette, K., & Visscher, A. (2019). Qualities of classroom observation systems. *School effectiveness and school improvement*, *30*(1), 3–29. https://doi.org/10.1080/09243453.2018.1539014

Burroughs, N., Gardner, J., Lee, Y., Guo, S., Touitou, I., Jansen, K., & Schmidt, W. (2019). A review of the literature on teacher effectiveness and student outcomes. In N. Burroughs, J. Gardner, Y. Lee, S. Guo, I. Touitou, K. Jansen, & W. Schmidt (Eds.), *Teaching for excellence and equity: Analyzing teacher characteristics, behaviors and student outcomes with TIMSS* (pp. 7–17). Springer. https://doi.org/10.1007/978-3-030-16151-4_2

Burke, M. J., Finkelstein, L. M., & Dusig, M. S. (1999). On Average Deviation Indices for Estimating Interrater Agreement. *Organizational Research Methods, 2*(1), 49–68. https://doi.org/10.1177/109442819921004

Campbell, D. T., & Fiske, D. W. (1959). Convergent and discriminant validation by the multitrait-multimethod matrix. *Psychological Bulletin, 56*(2), 81–105. https://doi.org/10.1037/h0046016

Fauth, B., Decristan, J., Decker, A. T., Büttner, G., Hardy, I., Klieme, E., & Kunter, M. (2019). The effects of teacher competence on student outcomes in elementary science education: The mediating role of teaching quality. *Teaching and Teacher Education, 86*, 102882. https://doi.org/10.1016/j.tate.2019.102882

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. (2020). Who sees what? Conceptual considerations on the measurement of teaching quality from different perspectives. *Zeitschrift für Pädagogik, 66*(1), 138–155. https://doi.org/10.3262/ZPB2001138

Fauth, B., Herbein, E., & Maier, J. L. (2021). Beobachtungsmanual zum Unterrichtsfeedbackbogen Tiefenstrukturen. Institut für Bildungsanalysen Baden-Württemberg. Retrieved from https://ibbw-bw.de/site/pbs-bw-km-root/get/documents_E-523136125/KULTUS.Dachmandant/KULTUS/Dienststellen/ibbw/Empirische%20Bildungsforschung/Programme-und-Projekte/Unterrichtsfeedbackbogen/IBBW_Unterrichtsfeedbackbogen_Manual_Juni%202021.pdf

Grünkorn, J., Klieme, E., Praetorius, A.-K., & Schreyer, P. (2020). *Mathematikunterricht im internationalen Vergleich. Ergebnisse aus der TALIS-Videostudie Deutschland.* Leibniz-Institut für Bildungsforschung und Bildungsinformation (DIPF). https://doi.org/10.25656/01:21156

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*(1), 156–172. https://doi.org/10.3262/ZPB2001156

Hattie, J. (2009). *Visible Learning. A synthesis of over 800 meta-analyses relating to achievement.* Routledge. https://doi.org/10.4324/9780203887332

Kane, T. J., & Staiger, D. O. (2012). *Gathering Feedback for Teaching: Combining High-Quality Observations with Student Surveys and Achievement Gains.* Bill and Melinda Gates Foundation. Retrieved from https://files.eric.ed.gov/fulltext/ED540960.pdf

Klieme, E., Lipowsky, F., Rakoczy, K., & Ratzka, N. (2006). Qualitätsdimensionen und Wirksamkeit von Mathematikunterricht. Theoretische Grundlagen und ausgewählte Ergebnisse des Projekts "Pythagoras". In M. Prenzel & L. Allolio-Näcke (Eds.), *Untersuchungen zur Bildungsqualität von Schule* (pp. 127–146). Waxmann.

Klieme, E., Pauli, C., & Reusser, K. (2009). The Pythagoras Study. Investigating effects of teaching and learning in Swiss and German mathematics classrooms. In T. Janik & T. Seidel (Eds.), *The power of video studies in investigating teaching and learning in the classroom* (pp. 137–160). Waxmann.

LeBreton, J. M., & Senter, J. L. (2008). Answers to 20 Questions About Interrater Reliability and Interrater Agreement. *Organizational Research Methods, 11*(4), 815–852. https://doi.org/10.1177/1094428106296642

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*(6), 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 of classroom or school characteristics in multilevel modeling. *Contemporary Educational Psychology, 34*(2), 120–131. https://doi.org/10.1016/j.cedpsych.2008.12.001

Praetorius, A.-K., Klieme, E., Herbert, B., & Pinger, P. (2018). Generic dimensions of teaching quality: The German framework of Three Basic Dimensions. *ZDM, 50*(3), 407–426. https://doi.org/10.1007/s11858-018-0918-4

Rakoczy, K., & Pauli, C. (2006). Hoch inferentes Rating: Beurteilung der Qualität unterrichtlicher Prozesse. In I. Hugener, E. Klieme, C. Pauli & K. Reusser (Eds.), Dokumentation der Erhebungs- und Auswertungsinstrumente zur schweizerisch-deutschen Videostudie "Unterrichtsqualität, Lernverhalten und mathematisches Verständnis". 3. Videoanalysen (pp. 206-233). Gesellschaft zur Förderung Pädagogischer Forschung (GFPF), Deutsches Institut für Internationale Pädagogische Forschung (DIPF). https://doi.org/10.25656/01:3130

Senden, B., Nilsen, T., & Blömeke, S. (2021). Instructional Quality: A Review of Conceptualizations, Measurement Approaches, and Research Findings. In K. Klette, M. Blikstad-Balas, & M. Tengberg (Eds.), *Ways of Analyzing Teaching Quality.Potentials and Pitfalls* (pp. 140–172). Universitetsforlaget. https://doi.org/10.18261/9788215045054-2021-05

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*(5), 705–721. https://doi.org/10.1037/edu0000075