Teaching quality in higher education: agreement between teacher self-reports and student evaluations

Martin Daumiller, Stefan Janke, Julia Hein, Raven Rinas, Oliver Dickhäuser, Markus Dresel

Angaben zur Veröffentlichung / Publication details:

Daumiller, Martin, Stefan Janke, Julia Hein, Raven Rinas, Oliver Dickhäuser, and Markus Dresel. 2023. “Teaching quality in higher education: agreement between teacher self-reports and student evaluations.” *European Journal of Psychological Assessment* 39 (3): 176–81. https://doi.org/10.1027/1015-5759/a000700.

Teaching Quality in Higher Education: Agreement Between Teacher Self-reports and Student Evaluations

Martin Daumiller
University of Augsburg

Stefan Janke
University of Mannheim

Julia Hein
University of Mannheim

Raven Rinas
University of Augsburg

Oliver Dickhäuser
University of Mannheim

Markus Dresel
University of Augsburg

Teaching quality is a crucial factor within higher education. Research on this topic often requires assessing teaching quality as a global construct through self-reports. However, such instruments are criticized due to the lack of alignment between teacher and student reports of instructional practices. We argue that while teachers might over- or under-estimate specific dimensions of teaching quality, the aggregation of these dimensions in the form of overarching teaching quality well reflects differences in teaching quality between teachers. Accordingly, we test a ten-item measure that allows faculty to self-report their teaching quality based on the aspects distinguished in the SEEQ (Marsh, 1982, 2007). Using 15,503 student assessments of teaching quality in 889 sessions taught by 97 faculty members, we conducted Doubly Latent Multi Level Modelling while considering bias and unfairness variables to model overarching teaching quality assessed by students, and simultaneously corrected for measurement error and potential distortions through the assessment situation. This global factor of teaching quality was strongly associated with teacher self-reported teaching quality (ρ = .74), which we interpret as evidence that global teacher reports of teaching quality can serve as sensible indicators of overarching teaching quality for nomothetic research in higher education.

Keywords: teaching quality, self-report, alignment, student ratings, education

Teaching quality is a crucial factor within higher education and is frequently considered as a key variable in empirical investigations (see Wagner et al., 2016). Besides ideographic research (underlying school or teacher evaluations) and analyses of specific instructional practices, there is a high interest in nomothetic research that strives to understand more general processes associated with high teaching quality (e.g., how motivations, experiences, and behaviors of faculty relate to teaching quality). To this end, researchers often seek to assess teaching quality as a global construct through teacher self-reports, as ratings by students are typically not feasible in large-scale surveys involving many institutions. Such instruments can, however, be criticized, as teacher reports of specific teaching practices often do not align well with student reports. We argue that while teachers might over- or under-estimate specific dimensions of teaching quality, the aggregation of these dimensions in the form of overarching teaching quality forms a sensible assessment that reflects differences in teaching quality between different teachers. We propose a measure that allows faculty to self-report their teaching quality based on the theoretically and empirically well-substantiated aspects distinguished in a widely used conceptualization of higher education teaching quality, the SEEQ (Marsh, 2007). To gauge how well this measure reflects differences in teaching quality, we consider student evaluations of

Correspondence concerning this article should be addressed to Martin Daumiller, Department of Psychology, University of Augsburg, Universitätsstr. 10, 86159 Augsburg, Germany; Martin.Daumiller@phil.uni-augsburg.de. ORCID: 0000-0003-0261-6143

The research reported in this article was supported by German Research Foundation Grant DI 929/5-1 to Oliver Dickhäuser and Grant DR 454/8-1 to Markus Dresel.

teaching quality as a benchmark while optimizing their validity and reliability through considering bias and unfairness variables as well as multiple student assessments.

Teaching Quality: Conceptualizations and Measurements

Teaching quality is frequently based on student ratings (Marsh, 2007). In higher education, the questionnaire "Student Evaluation of Educational Quality" (SEEQ; Marsh, 2007) is widely used in this regard. To describe teaching quality, this approach encompasses the dimensions of (1) learning/value and (2) overall evaluation, as well as (3) instructor enthusiasm, (4) organization/clarity, (5) group interaction, (6) individual rapport, (7) breadth of coverage, (8) examinations/grading, and (9) assignments/reading. For contexts where student-directed teaching methods are prevalent (such as Germany), another factor may be (10) the quality of student contributions that are facilitated and moderated by the instructor (Daumiller, Grassinger, et al., 2021). Research indicates that these dimensions can be aggregated into an overarching (second order) factor capturing teaching quality as a whole, which is of particular interest to researchers investigating antecedents and consequences of teaching quality at a broader level (Apodaca & Grad, 2005; Burdsal & Harrison, 2008; Rollett et al., 2021). While the SEEQ distinguishes different process and product dimensions of teaching quality – that vary regarding their observability and stability across different courses and are highly informative for better understanding instructional practices – such a second-order factor can serve as an overarching proxy for global teaching quality. Accordingly, in the present study we focus on this overarching factor of teaching quality instead of the specific dimensions of teaching quality. Prior research suggests that when examining teaching quality of many teachers from different institutions, student ratings may not be feasible, which is why self- reports are often utilized in such situations (e.g., Daumiller, Dickhäuser, et al., 2019; Porter, 2002).

Agreement Between Student and Teacher Reports

Regarding instruction, low correlations between teacher and student surveys are typically reported (e.g., Desimone et al., 2010; Kunter & Baumert, 2006; see Fauth et al., 2020). For example, Clausen (2002) found a relative agreement between teacher and student ratings ranging from r = –.28 to .42 for 12 dimensions of teaching quality, with especially low agreement for less observable behaviors such as autonomy support. Similarly, Lazarides and Schiefele (2021) reported some agreement between students and teachers for classroom management (r = .39), but little agreement for aspects such as emotional support (r = .22) or cognitive activation (r = –.02). It should be noted that most research on teacher and student agreement has been conducted in secondary education, while much less is known about higher education. Nevertheless, regarding the SEEQ, researchers have found that teachers and students distinguish between the same dimensions (Roche & Marsh, 2002), for which there is similar agreement as in secondary education (r = .26–.40) – but why do teacher and student reports of facets of teaching quality frequently not converge?

On the one hand, this may be a function of the theoretical construct being conceptualized by means of specific dimensions of teaching quality rather than teaching quality as a whole (e.g., Roche & Marsh, 2002). While the evaluation of specific dimensions of teaching quality (especially less directly inferable ones) may diverge between teachers and students, these deviations might be counterbalanced regarding the overall construct of teaching quality, resulting in greater agreement. Especially when being interested in a global measure of teaching quality, as in the present work, students' assessments may be used in the form of a higher order factor as a benchmark. It should be noted that for other research purposes, a specific consideration of the different dimensions and differences in agreement between teachers and students can be very informative to better understand instructional processes and the role of teachers and students therein (e.g., Lazarides & Schiefele, 2021).

On the other hand, the low agreement between teacher and student reports may also be due to limitations in how past research analyzed the association. To provide a sensible benchmark, the reliability and validity of the student assessments need to be optimized. Specifically, student ratings can be influenced by bias factors stemming from students or teachers, such as students' gender, prior interest, perceived difficulty, reason for participation, course format, as well as teachers' age and gender, which can lead to distorted teaching evaluations if not controlled for (Marsh, 2007). Further, due to variation of classroom instruction and different underlying reference periods of student ratings (e.g., lesson topic), agreement regarding student ratings at individual time points might be impaired (Wagner et al., 2016). As the global teaching quality, we are interested in this work is a rather stable pattern of teacher behavior in the classroom, the extraction of the consistent components from student ratings over several time points and different courses might provide more valid measures of teaching quality than single assessments prone to situational distortion and fluctuation (Wagner et al., 2016). Finally, the aggregation of all dimensions into an overarching latent second-order factor focuses on the consensus that underlies the aspects of teaching quality rather than on the assumption that single aspects of teaching can be accurately judged.

The Present Research

The present research is based on the notion that teacher self-reports of teaching quality can be criticized given their unclear validity. To gauge how well teacher self-reports (regarding a new measure of global teaching quality) reflect differences in teaching quality between different teachers, we use student assessments of teaching quality as a benchmark by optimizing their validity and reliability. Specifically, we consider bias and unfairness variables, extract the shared student ratings across time and courses, and model overarching teaching quality as a second-order factor. Confirming that student and teacher reports converge to a substantial extent would attest to the usefulness of teacher reports as a useful instrument for nomothetic research on overarching teaching quality.

Method

We used data from a larger research project in which student assessments of teaching quality were considered as consequences of faculty motivations (Daumiller, Siegel, et al., 2019). Students of the participating faculty members were asked to assess teaching quality across multiple sessions and courses, allowing for fine-grained insights. Besides these student reports that allow for the extraction of shared scores of teaching quality, faculty members were also asked to self- report their teaching quality. To this end, we used an instrument that considers each dimension of teaching quality with a single item by providing instructors with a clear and comprehensive definition of what each dimension entails. This scale can thus be considered an economic self-report scale that carefully matches the SEEQ dimensions of teaching quality. All materials, data, and code underlying this study are available in an open access repository (https:// osf.io/2pnkx/; Daumiller, Dresel, et al., 2021).

Sample

Our sample consisted of 15,503 student assessments of 7,126 students regarding 889 sessions of 194 courses taught by 97 faculty members. The students were

Table 1. Descriptive statistics of the Self-Reported Teaching Quality Scale

Item	Range	*M*	*SD*	Skew	Kurtosis
Learning/value (Extent to which students learn in your courses and develop a better understanding and interest for the subject)	2–8	6.07	1.22	−0.95	0.92
Overall rating (Extent to which your courses and you as a lecturer are rated/assessed by the students as a whole)	2–8	5.77	1.27	−0.53	0.10
Instructor enthusiasm (Extent to which you are committed and enthusiastic in teaching and how well you manage to make your courses interesting, active, dynamic and humorous)	2–8	5.96	1.46	−0.80	0.17
Organization/clarity (Extent to which you structure and explain the content of teaching as well as the used teaching materials; e.g., explaining the content of teaching in a comprehensive way, emphasizing important content, properly linking content)	2–8	6.13	1.32	−0.97	1.02
Group interaction (Extent to which you encourage your students to participate during your courses; e.g., contributing their own knowledge, asking questions, participating during discussions, etc.)	2–8	6.32	1.26	−0.79	0.91
Individual rapport (Extent to which you create a friendly atmosphere in dealing with students; including your interest in your students, taking their concerns seriously – inside and outside of classes)	2–8	6.66	1.34	−1.32	1.87
Breadth of coverage (Extent to which you teach taking different perspectives into consideration; e.g., inclusion of current scientific developments, consideration of different theoretical views and backgrounds – also if they differ from your own ways of thinking)	2–8	5.58	1.55	−0.23	−0.69
Examinations/grading (Extent to which your performance assessments of students are fair, benefit the students, and clearly meet the pre-discussed criteria, e.g., for contributions during the course)	1–8	5.81	1.65	−0.81	0.42
Assignments/reading (Extent to which your homework/exercises and given literature contribute to a deeper understanding of your contents of teaching)	1–8	5.45	1.49	−0.58	0.02
Contributions of students (Extent to which you select meaningful contributions for your students and moderate or complement them; e.g., contributions during discussions, presentations, group work)*	1–8	6.05	1.29	−0.70	1.60

Note. *N* = 97 faculty members. *is not an aspect of the original SEEQ but included in its German adaption (SEEQ-DE; Daumiller, Grassinger, et al., 2021) as an additional aspect of teaching quality that matters especially for contexts where student-directed teaching methods are prevalent (such as the higher education context in Germany).

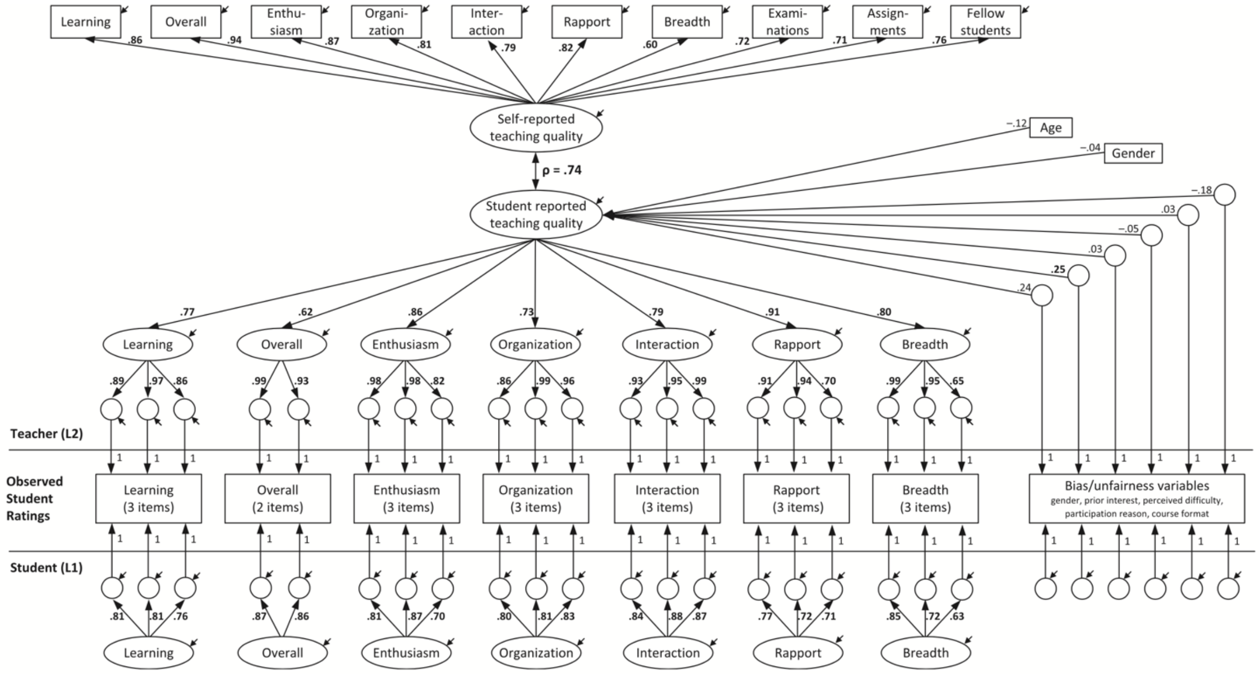

Figure 1. Results of the Doubly Latent Multilevel Model examining the alignment between self-reported teaching quality with shared student reports thereof ($\chi^2(938, N = 15{,}503) = 2{,}550, p < .001$; CFI = .966, TLI = .962, RMSEA = .011, SRMR = .070). Bold faced coefficients are statistically significant at $p < .05$.

33.2% female, 65.6% male, and 1.2% diverse; their average age was 23.0 years (SD = 2.5). The faculty members were on average 41.2 years old (SD = 10.6); 46 were male and 42 were female (9 did not state their gender). Their average teaching experience was 10.1 (SD = 7.9) years.

Measures

In the self-reported teaching quality scale (Daumiller, Dickhäuser, et al., 2019), faculty members referred their answers to the entirety of their current courses and assessed ten facets of their teaching (see Table 1) corresponding to the dimensions of the SEEQ on a scale from 1 (= very bad) to 8 (= very well). The internal consistency was high (McDonalds ω = .95).

For student assessments, we used the German adaption of the SEEQ (Marsh, 2007; Daumiller, Grassinger, et al., 2021) to measure learning/value (3 items; ω = .83), overall rating (2 items; r = .76), instructor enthusiasm (3 items; ω = .80), organization/clarity (3 items; ω = .81), group interaction (3 items; ω = .97), individual rapport (3 items; ω = .76), and breadth of coverage (3 items; ω = .76). Due to the study design (involving students rating multiple sessions) we reduced the number of items and did not include examinations/grading and assignments/readings, as these may not apply to all sessions. Intra-class correlations (ICC1 = .13–.41, ICC2 = .96–.99; see Table E1 in the Electronic Supplementary Material, ESM 1) showed moderate differences between the different teachers and that these dimensions were reliably assessed by the students.

As potential bias and unfairness variables, we considered students' gender, prior interest in the topic (single item from Marsh, 2007; scale: 1 = very little to 5 = very high), perceived difficulty (single item from Marsh, 2007; scale: 1 = very easy to 5 = very hard), and reason for participation (single item adapted from Marsh, 2007; recoded to two dummy coded variables representing whether participation was mandatory and/or out of interest), as well as the course format (seminar or lecture), and teacher age and gender.

Results

To answer our research questions, we conducted Doubly Latent Multilevel Analyses in which the shared ratings of teaching quality were modeled using a second-order factor on the between level (see Figure 1). We considered all dimensions of teaching quality conceptualized in the SEEQ, that is, both process dimensions as well as the product dimensions ("learning", "overall") to form the second-order factor and to adequately reflect all dimensions of the SEEQ. The shared component on the teacher level (L2) reflects the overarching teaching quality in which both measurement error as well as individual deviations through the assessment situations are corrected. Further, the bias and unfairness variables were regressed onto this factor to correct for potential distortions of student evaluations (Marsh, 2007). This factor thus describes differences between the participating teachers in what we equate most closely as overarching teaching quality assessed by students. Optimized in terms of validity and reliability, this forms the benchmark for which teacher self-reported teaching quality was compared to. To this end, our analysis indicated that the global factor of teaching quality as assessed by students was strongly and positively correlated with the factor of teacher self-reported teaching quality (ρ = .74, p = .008; 95% CI [.63, .82]).1 This means that factor scores of teacher self-assessments went along with respective differences in the overarching student-reported teaching quality. We interpret this as strong evidence for the measured scores of the self-report scale being validly interpretable.

Discussion

We considered a crucial aspect of higher education, teaching quality, and investigated how well a self-report scale can reflect differences in teaching quality. While teachers and students may not agree well on specific instructional behaviors, we contend that teacher reports of global teaching quality form a sensible indicator of differences in overarching teaching quality. As a benchmark, we used student assessments that were modeled in a way through which we know that their validity and reliability are optimized. Considering bias and unfairness variables, extracting the shared student ratings across time and different courses, and modeling overarching teaching quality as a second-order factor, we found a large positive correlation between both constructs. We consider this as evidence for the usefulness of teacher reports of global teaching quality as a short scale research instrument for future research on overarching teaching quality. However, two important aspects need to be considered when interpreting these findings.

First, while the data basis of the present investigation was fairly large, it should be considered that the faculty members voluntarily agreed to participate. Consequently, these may be teachers who are interested in their teaching and have thought about it already. Past research has indicated that exposure to student evaluations of teaching quality and feedback on one's teaching could lead to stronger associations between self-report and student-reported teaching quality (Roche & Marsh, 2002). Related to this, we only considered bias variables on the student side as we sought to optimize the students' evaluations of teaching quality in terms of their validity, but it should be considered that teachers may also be prone to biases (e.g., self-serving biases, lacking a frame of reference). Future research might consider investigating further variables at the teacher level to understand interindividual differences in the agreement between teachers and students.

Second, we focused on a self-report measure to be used in nomothetic research. However, beyond such research, this approach is not appropriate for high stakes testing as it can be easily manipulated (see Daumiller, Siegel, et al., 2019), or for when individual dimensions of teaching quality are of interest. However, for teaching quality as a whole, we consider teacher self-reports as a valid indicator and find that, besides their face and predictive validity, they also converge well with student reports. We hope that this helps to alleviate concerns about the use of teacher self-reports of their global teaching quality and facilitates future research on the processes associated with high teaching quality.

Electronic Supplementary Materials

The electronic supplementary material is available with the online version of the article at https://doi.org/10.1027/ 1015-5759/a000700

ESM 1. Basic sample and scale score statistics. Results of the Doubly Latent Multilevel Model.

References

Apodaca, P., & Grad, H. (2005). The dimensionality of student ratings of teaching. Studies in Higher Education, 30(6), 723–748. https://doi.org/10.1080/03075070500340101

Burdsal, C. A., & Harrison, P. D. (2008). Further evidence supporting the validity of both a multidimensional profile and an overall evaluation of teaching effectiveness. Assessment & Evaluation in Higher Education, 33(5), 567–576. https://doi.org/10.1080/ 02602930701699049

Clausen, M. (2002). Unterrichtsqualität [Reaching quality]. Waxmann. Daumiller, M., Dickhäuser, O., & Dresel, M. (2019). University instructors' achievement goals for teaching. Journal of Educational Psychology, 111(1), 131–148. https://doi.org/10.1037/ edu0000271

Daumiller, M., Dresel, M., Rinas, R., & Janke, S. (2021). Data and materials for "Teaching quality in higher education: Agreement between teacher self-reports and student evaluations". https:// osf.io/2pnkx/

Daumiller, M., Grassinger, R., Engelschalk, T., & Dresel, M. (2021). SEEQ-DE. Diagnostica, 67(4), 176–188. https://doi.org/ 10.1026/0012-1924/a000274

Daumiller, M., Siegel, S., & Dresel, M. (2019). Construction and validation of a Short Multidisciplinary Research Performance Questionnaire (SMRPQ). Research Evaluation, 28(3), 241–252. https://doi.org/10.1093/reseval/rvz009

Desimone, L. M., Smith, T. M., & Frisvold, D. E. (2010). Survey measures of classroom instruction. Educational Policy, 24(2), 267–329. https://doi.org/10/bdb76q

Fauth, B., Göllner, R., Lenske, G., Praetorius, A. K., & Wagner, W. (2020). Who sees what? Zeitschrift für Pädagogik, 66(1), 138– 155. https://doi.org/10.3262/ZPB2001138

Kunter, M., & Baumert, J. (2006). Who is the expert, Learning Environments Research, 9(3), 231–251. https://doi.org/ 10.1007/s10984-006-9015-7

Lazarides, R., & Schiefele, U. (2021). The relative strength of relations between different facets of teacher motivation and core dimensions of teaching quality in mathematics. Learning and Instruction, 76, Article 101489. https://doi.org/10.1016/j. learninstruc.2021.101489

Marsh, H. W. (1982). SEEQ: A reliable, valid, and useful instrument for collecting students' evaluations of university teaching. British Journal of Educational Psychology, 52(1), 77–95. https://doi.org/10.1111/j.2044-8279.1982.tb02505.x

Marsh, H. (2007). Students' evaluations of university teaching. In R. Perry & J. Smart (Eds.), The scholarship of teaching and learning in higher education (pp. 319–383). Springer.

Porter, A. C. (2002). Measuring the content of instruction. Educational Researcher, 31(7), 3–14. https://doi.org/10.3102/ 0013189X031007003

Roche, L. A., & Marsh, H. (2002). Teaching self-concept in higher education. In N. Hativa & J. Goodyear (Eds.), Teacher thinking, beliefs and knowledge in higher education (pp. 179–218). Springer.

Rollett, W., Bijlsma, H., & Röhl, S. (Eds.). (2021). Student feedback on teaching in schools. Springer. https://doi.org/10.1007/978- 3-030-75150-0

Wagner, W., Göllner, R., Werth, S., Voss, T., Schmitz, B., & Trautwein, U. (2016). Student and teacher ratings of instructional quality. Journal of Educational Psychology, 108(5), 705– 721. https://doi.org/10.1037/edu0000075