Learning Environments Research (2022) 25:775–802
https://doi.org/10.1007/s10984-021-09397-4

ORIGINAL PAPER

The perception problem: a comparison of teachers' self-perceptions and students' perceptions of instructional quality

Benedikt Wisniewski[1] · **Sebastian Röhl**[2] · **Benjamin Fauth**[3]

Received: 2 February 2021 / Accepted: 19 October 2021 / Published online: 7 November 2021

Abstract
Teachers' self-perceptions and their students' perceptions of the three basic dimensions of instructional quality were compared based on a sample of 171 classes and their teachers in German secondary education. Low to moderate correlations (r = .35 to .50) were found between the two perspectives. Differences in perceptions vary across teachers based on favorable and less favorable students' assessments. Results from latent profile analyses based on perception combinations of teachers and their classes hint at four differential profiles, reflecting to a large extent patterns of under- and overestimation of people's own competence identified in previous research. Significant differences in gender among individuals assigned to the four profiles could be found. Implications of identifying the divergence between teachers' and students' perceptions of instructional quality for reflective practice are discussed.

Keywords Dunning–Kruger-effect · Instructional quality · Latent class analysis · Teachers' self-perceptions

Theoretical framework

Based on the paradigm of reflective practice (Schön, 1983), teachers' self-reflection on teaching is considered an important prerequisite for professional development (Bengtsson, 2003; Ross & Bruce, 2007). Within the paradigm, reflection-on-practice focuses on how practitioners can change their methods or how they should move in new directions (Schön, 1983) and, with respect to teachers, this means becoming more-reasoned actors by questioning routines (Cruickshank, 1987) as well as by confronting personal assumptions and

✉ Benedikt Wisniewski
benedikt.wisniewski@phil.uni-augsburg.de

[1] Lehrstuhl für Schulpädagogik, University of Augsburg, Universitätsstr. 10, 86159 Augsburg, Germany

[2] Institute for Educational Science, University of Education, Freiburg, Germany

[3] Department of Empirical Education Research, Hector Institute for Empirical Educational Research, Tübingen, Germany

values as a consequence of experiencing perturbations in practice (Smyth, 1992). Consequently, reflective practice involves obtaining evidence about one's impact on how students benefit from teaching and how this impact could be improved (Benade, 2015; Hattie 2009). When teachers rely on their own perceptions of instructional quality, different mechanisms of bias can lead to miscalibrated perceptions and, consequently, to a learning environment that does not fit students' specific needs. Reflective practice therefore involves comparing self-perceptions with external data (e.g. students' perceptions). For characteristics of instructional quality, such comparisons are possible by obtaining both self and other perspectives with the help of standardized feedback questionnaires.

Previous research shows that teachers' self-perceptions and others' (students' or external observers') perceptions differ considerably from one another, with most studies reporting only small to moderate correlations between measurements of self and others' perspectives (Clausen, 2002; Fauth et al., 2014; Kunter & Baumert, 2006; Maulana et al., 2012; Wagner et al., 2016; Wisniewski et al., 2020). This can be accounted for by perspective-specific validities (Fauth et al., 2014; Wettstein et al., 2016), meaning that different points of view "tap different aspects of the classroom environment, rather than the same underlying construct" (Kunter & Baumert, 2006, p. 234). However, for some measures, the construct structure of instructional quality based on teachers' and students' perceptions is similar (Kunter et al., 2008; Wisniewski et al., 2020), making a comparison of the two perspectives possible.

In general, as discussed by Clausen (2002), the correlation between teachers' and students' perceptions of instructional quality is lower for high-inference characteristics that are partly influenced by students' preconditions (e.g. motivation, interest or prior knowledge). Additionally, they are lower when more-difficult characteristics are observable. The three basic dimensions of instructional quality are based on items with relatively low inference but, on the other hand, they are not easily observable but result from classroom interaction (Wisniewski et al., 2020). This means that teachers and students both judge the instruction provided, but divergence arises from subjective assessments of how effective this instruction is.

When reflective practice is considered as the questioning of one's own assumptions about teaching, then comparisons of perceptions can be used to adapt teaching to students' needs as well as to identify miscalibrations of one's own perceptions. The measurement of instructional quality characteristics from different perspectives allows such comparisons based on relevant criteria. The comparison of perception differences based on questionnaire data is considered a common way to identify blind spots in one's own perception and adapt teaching to students' needs (Helmke & Lenske, 2013). They help to question one's own assumptions and prevent reflection based merely on validating and perpetuating one's own view (Larrivee, 2006). Therefore, Hattie (2009) suggests that critical self-reflection as proposed by Schön (1983) needs to be enriched by evidence in the form of external data (in the form of feedback). Antoniou and Kyriakides (2011) state that critical reflective practice by teachers must utilize a combination of feedback from others, educational research, and their own perceptions. Instead of using criteria that are only subjectively perceived as relevant—different models of instructional quality provide characteristics that correlate with high achievement gains (Klieme et al., 2006 Slavin, 1994; Wisniewski et al., 2020).

Estimating the degree of disagreement between teachers and students is also important for understanding how teachers can benefit from feedback. Accordingly, studies of differences between student and teacher ratings of instructional quality can not only advance the current state of research, but also benefit educational practice. Existing research shows that inaccurate self-perceptions are relatively stable towards feedback (Hacker et al., 2000;

Helzer & Dunning, 2012; Simons, 2013) and low performers have difficulties in calibrating their self-perceptions based on feedback (Brett & Atwater, 2001; Kruger & Dunning, 1999). These individuals do improve the accuracy of their self-assessment through feedback when it is continuous, concrete, and specific (Miller & Geraci, 2011), which is consistent with more-general findings showing that feedback is ineffective when it focuses on very general observations (Braga et al., 2014; Kornell & Hausman, 2016), or when it is related to personality characteristics instead of actual behavior (Hattie & Timperley, 2007; Kluger & DeNisi, 1996). This means that people who over-estimate their performance need different feedback from those who assess their own performance realistically or tend to under-estimate themselves.

In this article—after a brief outline of the state of research on differences between students' and teachers' perceptions of instructional quality—we report how teachers' self-perceptions and students' perceptions are related, how differences in perception vary across teachers based on assessments of favorable and less-favorable students' assessments, and how these differences are moderated by person and context variables.

Assessing instructional quality

A common method for assessing characteristics of instructional quality are questionnaires including items that are associated with high learning outcomes (Kunter & Voss, 2013). These can usually be answered from different outside perspectives (observers, students) and from a teacher self-perspective (Helmke & Lenske, 2013). With students providing the least- expensive way to obtain formative evaluation feedback, it is disputed if and to what respect students' knowledge and experience enables them to provide reliable and valid information on these characteristics (Lamb, 2017). Primary students have been shown to distinguish teaching quality and popularity insufficiently (Fauth et al., 2014) and it has been shown that—even for college students—discriminant validity can be compromised by very trivial interventions such as giving students chocolate before they evaluate the teaching quality (Hessler et al., 2018; Youmans & Jee, 2007).

Nonwithstanding, many studies have shown that, when applying questionnaires that are based on sound theory, students' perceptions of instructional quality exhibit a high degree of discriminant validity (Balch, 2012; Ferguson, 2012; Gaertner, 2014; Kane & Staiger, 2012; van Petegem et al., 2008; Wagner et al., 2013; Wisniewski et al., 2020) and relevant outcomes of teaching (e.g., student achievement) are predicted more accurately by students' perceptions than by teachers' self-perceptions (De Jong & Westerhof, 2001; Fraser, 1991; Clausen, 2002; Kunter & Baumert, 2006; Pham et al., 2012; Seidel and Shavelson 2007). Additionally, as aggregated data, students' perceptions are more reliable than teachers' self-perceptions (Kyriakides et al. 2014). Taking these findings into account, comparing teachers' self-perceptions with students' perceptions of instructional quality is still a comparison of subjective data with other subjective data does not allow a decision to be made about which perspective is more accurate. However, students' perceptions can be considered an important source of information for identifying teachers' miscalibrated self-assessments that can adversely affect the effectiveness of teaching.

A comparison can be based on different quality criteria. There are several frameworks of instructional quality (see Wisniewski et al., 2020 for an overview). One of the most prominent models, the so-called three basic dimensions, includes the generic factors of classroom management, cognitive activation, and student support (Klieme et al., 2006 Praetorius et al., 2018). Classroom management includes rules and procedures, measures

of coping with disruption, and smooth transitions, which directly influence time on task (Seidel and Shavelson 2007). These three dimensions have been shown to predict students' cognitive, motivational and affective learning outcomes (Praetorius et al., 2018) and can be obtained independently of students' age, school subject or school type (Wisniewski et al., 2020). Cognitive activation integrates challenging tasks, connects newly-introduced concepts to prior knowledge, and encourages students in elaborate thinking and classroom discussion (Lipowsky et al. 2009), stimulating deep forms of thinking and conceptual understanding during learning (Klieme et al., 2006). Supportive climate refers to aspects of the teacher–student relationship, including a caring behavior, a productive way of dealing with errors, and constructive feedback (Klieme et al., 2006). With its very strong theoretical foundation and multiple verifications by confirmatory factor analyses (Fauth et al., 2014; Kunter and Voss 2013), the model offers a parsimonious structure for operationalizing the construct of instructional quality. Previous research has shown different results of comparability of the three basic dimensions across different perspectives: Kunter and Baumert (2006) found different factor structures for both perspectives and concluded that they are indeed different constructs. They also reported that only one of the three factors, classroom management, was comparable between teachers and students and that there was significant agreement between the two groups when this factor is assessed. Other research has demonstrated that the assessment of the three basic dimensions is indeed invariant across teachers and students (Kunter et al., 2008; Wisniewski et al., 2020). In order to make claims about the agreement or divergence of teachers and students when assessing instructional quality, measurement invariance of the two perspectives is a basic prerequisite.

In discussing the differences between generic aspects and subject-specific aspects of instructional quality, Lipowsky et al. (2018) show that the basic dimensions of instructional quality can be supplemented by subject-specific instructional quality, whereby the generic and subject-specific factors are largely independent of each other.

Also, in differently-designed questionnaires for surveying the learning environment, considerable divergences between student and teacher perceptions are found (Fraser, 2007). Findings from research on teacher–student interpersonal relationships and behavior revealed that a large proportion of the teachers tend to overestimate aspects of their behavior which are positively related with students' motivation and achievement compared with their students' perceptions. Additionally, teachers' tendency for underestimation of teaching aspects which are perceived as negative are widespread (Den Brok et al., 2006; Maulana et al., 2012).

Explanations for the divergence of perceptions

There are several explanations for self–other perception differences related to different tasks from personality and social psychology. While self-perceptions are characterized by privileged access to thoughts and are not dependent on interpretation of behavior, the perceptions of others rely on indirect behavioral indicators, drawing inferences, and interpreting behavior (Fauth et al., 2020). On the other hand, people's difficulty in detecting their own stable (positive and negative) behavioral tendencies stems from a lack of awareness similar to the phenomenon that fish are said to find it difficult to detect water (Kolar et al., 1996; Leising et al., 2006). Self- perceptions are less associated with actual behavior than the perceptions of others (Kolar et al., 1996). Summarizing differences between self and other ratings, the SOKA model (self–other knowledge asymmetry) by Vazire (2010) assumes that there are differences in the information that is available for a rater, as well as

differences in the processing of that information. Regarding the latter, the model accounts for the degree of ego involvement that differs between self and other ratings. In turn, this can lead to miscalibrated self-perceptions: "Judges have a lot more at stake when they are also the target than when they are judging someone else" (Vazire, 2010, p. 284).

Another reason for self-perceptions differing from others' perceptions is a specific cognitive bias phenomenon defined by Kruger and Dunning (1999). Less-skilled people usually overestimate their performance because they are less able to reflect accurately on what they do, whereas highly-skilled performers underestimate their skills because of their overestimation of other people's skills (and therefore underestimate their own relative competence) and out of modesty (Dunning et al., 2003). Although there is a low to moderate correlation ($r = 0.39$) between self-assessment and actual ability (e.g. with respect to examination taking), people who perform particularly poorly are unaware of their incompetence, while those who perform particularly well tend to underestimate themselves and are not fully aware of their good performance compared with peer group members (Kruger & Dunning, 1999). This effect—often called the Dunning–Kruger-effect—was originally attributed to metacognitions, assuming that the same skills that are necessary to solve a cognitive task are necessary to recognize whether the processing of that cognitive task is successful. Because people who are less competent in a task cannot produce a correct result, they cannot recognize a correct result. This leads them to overestimate their own abilities. When Kruger and Dunning (1999) split subjects into quartiles based on their ability in different tasks, those in the bottom quartile overestimated their performance the most strongly, while those in the top quartile slightly underestimated their performance. Until now, various alternatives have been presented to explain Kruger and Dunning's (1999) findings. Krueger and Mueller (2002) argue that the observable effect is a distortion that arises as a result of a regression to the middle of the self-assessment: both subjects with above-average competence and subjects with below-average competence tend to assess their abilities as average. This could also explain why the best performers in Kruger and Dunning's (1999) surveys underestimated their abilities. However, recent research with statistical control for this tendency of regression to the mean shows that the Dunning–Kruger effect can be reduced somewhat but cannot be fully explained by regression to the mean (McIntosh et al., 2019).

One moderator of the association between self-assessment and objective performance is the specificity of the items that are used to measure perceived ability (Ackerman et al., 2002; Zell & Krizan, 2014). Self-assessments are generally more precise for narrowly-defined areas of behavior that are based on clear criteria (e.g., "I intervened quickly and consistently when students ignored classroom rules") than for broader areas that are based on an overall impression (e.g., "I'm good at classroom management"). These findings are relevant for the present study because they suggest a specific and criterion-based operationalization of instructional quality.

Gender influences

Generally, misjudgments of people's own abilities or performances are moderated by different variables, with gender being one of the most influential ones (Lindeman et al., 1995; Lundeberg et al., 1994, 2000). Significant gender differences are found for many kinds of tasks, with men tending to assess their performances more positively, while women's self-evaluations are generally inaccurately low. Existing research has also shown that women especially underestimate their achievements in masculine-stereotyped tasks or domains

(Beyer & Bowden, 1997). Ehrlinger and Dunning (2003) found no actual differences between female and male college students' performance on a science test, but female students underestimated their performances because they thought less of their general scientific reasoning abilities. Similarly, female managerial students assess their own abilities that qualify them for a leadership position significantly lower than their male counterparts do (Bosak and Sczesny 2008). These findings indicate a pervasive gender bias in self-concepts related to performance. However, to the best of our knowledge, the question of how gender can affect the self-perceptions of teachers (especially in comparison to their students' perceptions) is still unresolved.

The present study

Previous research shows that teachers *generally* perceive instructional quality characteristics differently from students. However, it is unclear how teachers differ from each other in perceiving these characteristics compared with their students' perceptions. Previous findings do not address the question of how the differences vary between those teachers whose instructional characteristics are perceived favorably by students and those whose instructional characteristics are perceived less favorably. The purpose of the present study was to further explore the relationship between teachers' and students' perceptions of instructional quality by investigating patterns of teachers' over- and under-estimation of characteristics of generic instructional quality compared with students' perceptions. On the one hand, this study aimed to generate further findings about the different perception perspectives on teaching and explain these differences.

On the other hand, these findings can be used to provide teachers with important information for reflecting on their own teaching based on student feedback and self-perceptions. This knowledge about possible explanations for differences in perception, for example, could lead to a more self-critical attitude toward one's own teaching when teachers overestimate themselves to a particularly high degree. At the same time, it might encourage teachers who under-estimated their teaching (especially female teachers) to think somewhat more positively about their own teaching. Therefore, the research design encompassed the testing of hypotheses about differences in perception and the investigation of differential perception profiles.

After testing whether teachers' and students' perceptions obtained with the instrument used were comparable or, in other words, whether the measurement of instructional quality is invariant across the two (RQ1), we investigated how self-assessment and external assessment by students of instructional quality are correlated and whether and to what extent this correlation differs depending on the external assessment (RQ2). Following up on previous research dealing with self-perceptions that has shown different patterns for women and men when assessing their own behavior, we investigated whether gender effects on self-assessment known from other contexts can also be transferred to teachers' assessments of teaching (RQ3). To this end, we put forward four specific hypotheses:

- The perception of instructional quality can be obtained by the same measurement model for students and teachers (H1).
- Teachers' self-perceptions and students' perceptions are moderately correlated (H2.1).

- Teachers' self-perceptions differ in a systematic way from students' perceptions regarding the dimensions of classroom management (H2.2.1), cognitive activation (H2.2.2) and student support (H2.2.3):
  - (a) Overestimation of instructional quality characteristics is largest among those whose lessons are perceived unfavorably by their students.
  - (b) Correct estimation or underestimation can be found among those assessed favorably by their students.
- Male teachers show a tendency to significantly overestimate their own instructional quality (H3).

The three hypotheses are related to perception differences with respect to three separate characteristics of instructional quality. If patterns of over- and under-estimation of teachers' perceptions compared with students' perceptions can be shown for these three dimensions—given the multidimensionality of instructional quality—it must be clarified if the simultaneous consideration of all three dimensions allows the identification of teacher profiles that reflect their perception of instructional quality in general relative to their students' perceptions. Therefore, in a next step, we used a more-exploratory method to investigate typical inter-personal patterns in the deviation of self-assessment from an external assessment with regard to instructional quality in general. We analyzed the extent to which different profiles occur in the combination of teacher and student assessments of the three dimensions of instructional quality (RQ4). Finally, we explored if and to what extent personal and context variables (grade, school type, school subject, teacher gender) are associated with the perception profiles to which teachers belong (RQ5).

Because student assessments of teaching in primary and secondary schools are unsuitable or only suitable to a very limited extent for accountability purposes of teachers to supervisors (Röhl & Rollett, 2021), we limit interpretation of our findings to their relevance for teachers' self-reflection on their instruction.

Sample

The sample consisted of 171 teachers (51% female) teaching classes in grades 5–12 at eight German schools from three different school types. The corresponding student sample consisted of 4108 students. These three school types are university preparatory high schools (Gymnasium), intermediate secondary schools (Realschule) and vocational schools (Berufliche Schule). Within the German school context, these are three types of secondary schools, with the first two starting from grade 5 and both requiring certain entry grades from grade 4 (primary school), and vocational schools starting from grade 10 and requiring the completion of junior high school.

Data collection

Students' perceptions of instructional quality (with an average cluster size of 22.3) were surveyed during the period from September 2017 to October 2019 via an online portal. The data stem from the everyday school context, rather than being obtained for research purposes: teachers collected feedback for their professional development and

then provided the results for scientific analysis. This is why a period of this duration was chosen. No extra incentive was provided for teachers or students. For every survey, teachers assessed themselves on the same items as their students before they had received the students' assessments. Because of the technical nature of the online portal and privacy restrictions, no personalized student data were obtained. Personalized teacher data were anonymized before being transferred to us for analysis. The heterogenous database reflects the variety of different school types in the German school system, but it does not constitute a representative sample of the school system in a narrower sense. Teachers decided to use the online feedback portal voluntarily, which means that the sample was restricted to those teachers who were willing to reflect on their teaching based on student feedback or simply wanted to try out this instrument.

Measure

Students' and teachers' perceptions of instructional quality were surveyed using the teaCh questionnaire (Wisniewski & Zierer, 2020) consisting of 29 items, which refer to the seven categories of care, control, conferment, clarity, challenge, consolidation, and captivation (see Table 1). The items for teachers are identical to the student version but are formulated from the teacher's perspective. As most of the item formulations focus on the teacher (e.g., "The teacher had high expectations of me"; "I had high expectations of the students"), the questionnaire measures comparable self–other perceptions of teacher's behavior in the classes.

As latent second-order factors, these categories load on the known three basic dimensions of instructional quality, namely, classroom management, cognitive activation, and student support (Praetorius et al., 2018), which were used for analyses in this study. Both versions were rated on a four-point Likert-type scale, ranging from 1: I Don't Agree to 4: I Agree. The instrument has been shown to measure general instructional quality in a valid way and that the measurement is generalizable across school types, school subjects, and grade levels in a secondary school context (Wisniewski et al., 2020). It also allows a valid comparison of student and teacher perspectives.

Statistical analyses

Using the actual sample, we conducted a two-level confirmatory factor analysis of the assumed factor structure of seven first-order factors and three second-order factors referring to the basic dimensions of instructional quality. To compare the values of the student and teacher perspective of instructional quality in our analyses, at least a metric invariance between the two perspectives is necessary. For testing this, the models with the restrictions were compared between the groups with the less restrictive precursor. χ^2 statistics have proven to be an unreliable indicator of measurement invariance for large samples because significant p-values can be obtained almost irrespective of actual differences of model fit (Cheung & Rensvold, 2002; Kline, 2016). As an alternative, goodness-of-fit indices can be used as a more-reliable source of information to test for measurement invariance. As proposed by Meade et al. (2008), $\Delta CFI \leq 0.002$ was chosen as a comparative indicator. To ensure an even better comparability of the two perspectives, equal item loadings for students and teachers were specified for the subsequent analyses.

Table 1 Dimensions of teaCh questionnaire

Dimension	2nd order factor	Number of items	Example of item	ω_t
Captivation	Student support	6	The contents of the lesson were taught by the teacher in an interesting way	.94
Conferment	Student support	4	The teacher gave me helpful feedback on my performance	.94
Care	Student support	3	The teacher met me in a friendly and appreciative way	.87
Clarity	Cognitive activation	4	The teacher has tied in content that was already known to me	.92
Challenge	Cognitive activation	2	The teacher had high expectations of me	.90
Consolidation	Cognitive activation	4	During the lesson there were plenty of opportunities to practice the new content	.97
Control	Classroom management	6	When students violated the rules, the teacher intervened quickly and consistently	.94

ω_t, McDonald's Omega total. Examples of items translated from German

To show systematic miscalibrations of self-perceptions, most research on the Dunning–Kruger effect uses the percentile ranks of actual performance and self-assessments (Kruger & Dunning, 1999). Therefore, factor scores were scaled in the same way to test the relevant hypotheses.

For the in-depth regression analyses, we used the factor scores for the three basic dimensions of instructional quality provided by the program MPlus 8.2 (Muthen & Muthen 2012–2019). A latent profile analysis (LPA) based on teachers' self-perception compared with the students' perceptions of their classes was conducted, also using the three basic dimensions of instructional quality. Because teachers were nested in schools, we accounted for possible dependencies in our data by correcting standard errors and a chi-square test of model fit (TYPE = COMPLEX), with schools used as clusters. To identify the best-fitting profile solution, we estimated fit indices (BIC, aBIC, AIC), likelihood ratio tests (Vuong likelihood ratio test and Lo-Mendell-Rubin likelihood ratio test, entropy) and the number of subjects per assumed class for different solutions.

After identifying perception profiles, we tested if person and context variables were associated with the assignment of teachers to these profiles. We used the grade that was taught, the school type, the school subject, and the teacher's gender for this analysis.

Results

Descriptive results

All observed item means were slightly above the theoretical mean (ranging from 2.99 to 3.54 for teachers and from 2.98 to 3.35 for students), with standard deviations ranging from 0.60 to 0.92 for students and from 0.62 to 0.88 for teachers. Responses were approximately normally distributed with skewness ranging from − 1.44 to − 0.33 for teachers and from − 1.43 to − 0.49 for students. Kurtosis values ranged from − 0.61 to 1.79 for teachers and from − 0.61 to 1.77 for students. More specific descriptive data can be found in the Table 2 in the "Appendix".

Measurement model

The test for measurement invariance across teachers and their classes pointed to an acceptable fit for the sample. Using $\Delta CFI \leq 0.002$ as comparative indicator (Meade et al., 2008), results pointed to strong measurement invariance between the groups (see Table 3). The measurement model with equal item loadings for teachers and students used in the following pointed to an acceptable fit for the sample (CFI = 0.93 TLI = 0.92, RMSEA = 0.02 $SRMR_{within} = 0.03$, $SRMR_{between} = 0.10$). Intraclass correlations for the 29 items on the student level were substantial with ICC_1 ranging from 0.08 to 0.24 (median 0.17) and ICC_2 ranging from 0.68 to 0.88 (median 0.82).

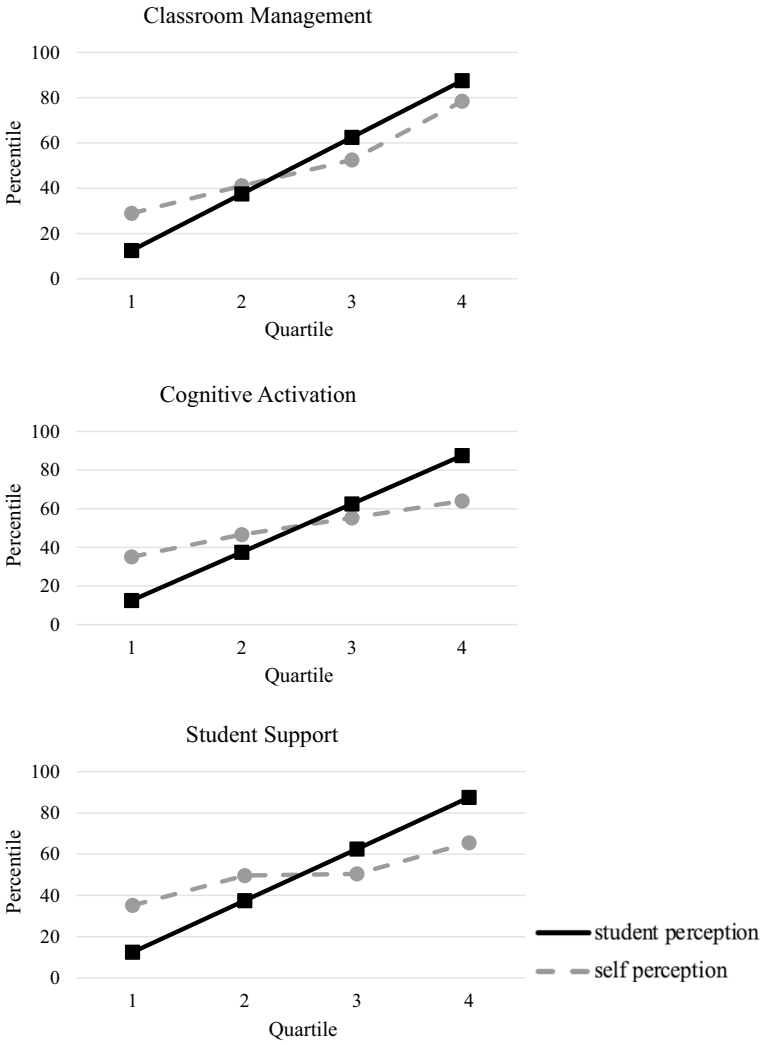

Fig. 1 Self-perception and aggregated student perceptions of instructional quality

Deviations between teachers' self-perceptions and students' perceptions

The correlations between self-perception and aggregated students' perception were $r = 0.52$ ($p < 0.001$) for classroom management, $r = 0.35$ ($p < 0.001$) for cognitive activation, and $r = 0.40$ ($p < 0.001$) for student support. For all three basic dimensions, teachers in the bottom quartile overestimated their performance based on student perceptions, whereas teachers in the three other quartiles either agreed with their students' perceptions or underestimated their performance (Fig. 1).

Influence of gender

Female teachers received better ratings than male teachers, with small but significant differences for all three basic dimensions (t_{CM} [127] = 2.52, $p < 0.05$; t_{CA}[129] = 3.08, $p < 0.01$; t_{SS}[132] = 3.34, $p < 0.01$). Self-perceptions were significantly different between female and male teachers (with more favorable self-perceptions for men) for classroom management and cognitive activation, but not for student support (t_{CM} [166] = − 2.57, $p < 0.05$; t_{CA} [169] = − 2.04, $p < 0.05$; t_{SS}[169] = − 1.77, $p > 0.05$), See Table 4.

In predicting teachers' self-perceptions using student perceptions and teachers' gender, both variables showed significant effects. Together, student perceptions and teachers' gender explained between 18 and 36% of variance in the self-perception for the three dimensions of classroom management, cognitive activation, and student support (Table 5.)

Identification of different perception profiles

The fit indices, likelihood ratio tests and number of subjects per assumed class for solutions for classes 1–12 are shown in Table 6. Both, the Vuong and the Lo-Mendell-Rubin likelihood ratio test supported the four-class-model. In addition, the difference of aBIC to the next number of classes decreased most significantly from 3 to 4 (ΔaBIC = 148) and from 4 to 5 (ΔaBIC = 98). Therefore, the four-class-model was selected for further analysis.

Class 1 (10.53%) was characterized by the lowest student assessments of instructional quality and significant differences between students' and self-perceptions for classroom management and cognitive activation ($p < 0.001$), whereas no perception difference was found for student support ($p = 0.10$) for this class. Class 2 (21.05%) was characterized by the second-lowest student assessments and significant perception differences for classroom management ($p < 0.05$), cognitive activation ($p < 0.001$) and student support ($p < 0.01$). Class 3 (40.35%) was characterized by teachers' underestimation of all three dimensions of classroom management ($p < 0.01$), cognitive activation ($p < 0.01$) and student support ($p < 0.05$) compared with their students' perceptions. Class 4 (28.07%) was characterized by the most-positive student assessments, agreement of self with students' perceptions for classroom management ($p = 0.40$) and student support ($p = 0.80$), and a significant underestimation of cognitive activation compared with students' perceptions ($p < 0.001$). Figure 2 shows the average factor scores for students' and self-perceptions of the latent variables.

Perception profiles and their association with grade, school type, school subject, and teacher gender

No significant differences were found for grade ($\chi^2 = 4.81$, $df = 6$, $p = 0.57$), school type ($\chi^2 = 1.59$, $df = 6$, $p = 0.95$) or taught school subject ($\chi^2 = 22$, $df = 18$, $p = 0.23$) among individuals assigned to the four classes could be found, However, a chi-squared test revealed significant differences in teacher gender ($\chi^2 = 28$, $df = 3$, $p < 0.001$, see Table 7), with a higher proportion of men in the overestimating classes 1 and 2.

Table 8 shows the assignment to the four classes differentiated by the taught subject, whereas

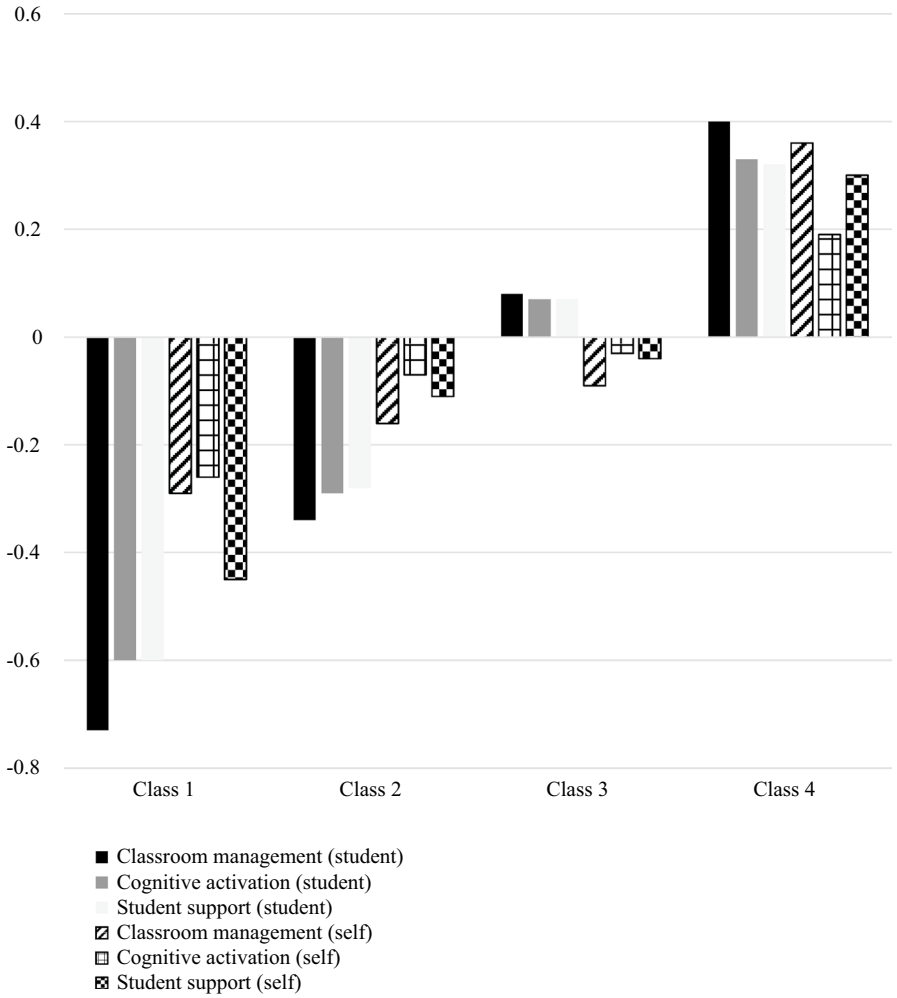

Fig. 2 Average factor scores for students' and self-perceptions of the latent variables

Table 9 shows the assignment to the four classes differentiated by teacher gender and taught subject ($\chi^2 = 51$, $df = 18$, $p < 0.001$).

Discussion

The aim of the present study was to investigate the divergence between teachers' self-perceptions and students' perceptions of instructional quality. We argue that this divergence is the basis for meaningful reflective practice because, when teachers misjudge instructional quality characteristics in comparison to their students, they cannot adapt the learning environment to their students' needs. A regression analysis and

latent profile analysis were applied to explore disagreement between teachers and students on characteristics of instructional quality.

Comparing perceptions across different groups requires that the measured constructs have the same meaning across these groups, and that comparisons of sample estimates are not distorted by group-specific attributes. As a prerequisite for our investigation, we therefore confirmed the assumed factor structure of the instrument used to obtain self and students' perceptions for our sample, and that the three superordinate dimensions of instructional quality were invariant across teachers and students, allowing a comparison of the two perspectives (H1).

In line with previous research (Clausen, 2002; Fauth et al., 2014; Kunter & Baumert, 2006; Maulana et al., 2012; Wagner et al., 2016), the overall correlation between teachers' self-perceptions and students' perception was only low to moderate (H2.1).

However, correlations based on the whole sample offer no informational value about how teachers differ in the accuracy of their self-perceptions. A classification of teachers into quartiles according to their students' perceptions produced a pattern similar to the pattern found by Kruger and Dunning (1999) for different tasks. Teachers with more-unfavorable students' perceptions overestimated their performance. The results from the latent profile analysis can hardly be explained by a regression to the middle of self-assessments as proposed by Krueger and Mueller (2002) because there is substantial overestimation of teachers with unfavorable students' assessments on the one hand, but little underestimation of teachers with favorable students' assessments. Teachers' self-perceptions are more accurate when the external perception is more favorable. Consequently, our data support Kruger's and Dunning's (1999) hypothesis, understanding over-estimations as a consequence of the inability to assess characteristics of instructional quality that someone puts into practice inadequately.

Teachers who overestimated their performance (classes 1 and 2) accounted for about one third of the sample. The 11% of teachers who received the lowest ratings for all three basic dimensions of instructional quality overestimated their performance regarding classroom management and cognitive activation the most. The two classes with the most-favorable student assessments accounted for more than two-thirds of the sample and were characterized by more-precise self-perceptions and underestimations.

While the assignment to the four identified classes was independent of grade level, school type, and taught subjects, significant gender differences were found, with female teachers being under-represented and male teachers being over-represented in the profile defined by the highest overestimation of one's own performance. These associations were not dependent on the school subjects taught. Effects of women especially underestimating their achievements in masculine-stereotyped tasks or domains (Beyer & Bowden, 1997) were not replicated in the way in which female teachers assessed their instructional quality more negatively in subjects like mathematics, physics or IT that are traditionally stereotyped as male domains (Makarova et al., 2019).

There are certain limitations of the study that need to be borne in mind when interpreting the results. Firstly, our sample included teachers from two German federal states and three school types and is therefore not representative for the whole German school system. Secondly, we could not test for effects of some personal variables. We could not consider students' gender because no individual-related student data were collected for data-protection reasons. It has been shown in previous research that there are interaction effects of teacher and student gender (Boring, 2017; Mitchel and

Martin 2018) for evaluations of teaching in higher education, but we could not expand on this research. We also were unable to use information on teachers' age or professional experience and could therefore not check if theses variables influence patterns of self-perception. Finally, teachers' perceptions of instructional quality were only compared with (also subjective) students' perceptions. Future research should expand on this by using actual teacher performance data (e.g. in the form of value added) so that claims of miscalibrations can be made based on objective data.

Despite these limitations, practical implications can still be drawn: although a favorable self-assessment can have beneficial effects (Bandura, 1977, 1997; Fox et al., 2009; Mosing et al., 2009), under- and over-estimations can also cause detrimental motivational consequences (Dunlosky & Rawson, 2012). Most importantly, over-estimations of one's own performance hinders people from developing their professional skills because they are unaware of reasons for improvement. This is especially disadvantageous when those who require professional development the most are least aware of its necessity.

Our findings point to the importance of adaptive feedback for teachers, considering different patterns of disagreement and counteracting under- and over-estimation. Because of the relative stability of inaccurate self-perceptions towards feedback (Hacker et al., 2000; Helzer & Dunning 2012; Simons, 2013) and the difficulties of low-performers in calibrating their self-perceptions based on feedback (Brett & Atwater, 2001; Kruger & Dunning, 1999), it is noteworthy that teachers who over-estimated the instructional quality of their lessons compared with how students perceived it did not benefit from simply getting the information that there is a divergence, and therefore need support to improve (Röhl & Gärtner, 2021). Feedback which is based on specific observations of actual behavior is more likely to provide opportunity for improvement (Braga et al., 2014; Kluger & DeNisi, 1996; Kornell & Hausman, 2016; Miller & Geraci, 2011).

In his seminal study, Goodlad (1984, quoted from Lamb, 2017) claimed that, by not using students' feedback, an essential part of reflective practice has been neglected. While research on the effectiveness of feedback to teachers from peers or traditional supervisors has provided conflicting results (Scheeler et al., 2004), student feedback seems to lead to improved instructional quality (Röhl & Gärtner, 2021). Consequently, questionnaire data from student feedback, as used for our analysis, can be one important feedback source used by teachers to reflect on teaching and improve on instructional practices. Our results point toward the necessity for this feedback to be based on clearly-defined dimensions of instructional quality and for those who overestimate their performance to be supported with measures of counselling and coaching.

Appendix

See Tables 2, 3, 4, 5, 6, 7, 8.

Table 2 Formulations of the items used to assess instructional quality and descriptive statistics

Item	1st-order factor	2nd-order factor	Item formulation	Students				Teachers			
				M	SD	γ_1	γ_2	*M*	SD	γ_1	γ_2
car1	Care	Student Support	The teacher met me in a friendly and appreciative way [Die Lehrperson begegnete mir freundlich und wertschätzend]	3.54	0.67	−1.43	1.79	3.5	0.73	−1.19	0.38
car2	Care	Student Support	The teacher created an atmosphere free of fear [Die Lehrperson sorgte für eine angstfreie Atmosphäre]	3.46	0.75	1.30	1.09	3.48	0.69	−0.94	−0.39
car3	Care	Student Support	The teacher was interested in whether I really learned something [Die Lehrperson interessierte sich dafür. ob ich wirklich etwas gelernt habe]	3.15	0.84	−0.69	−0.25	3.17	0.73	−0.53	−0.19
conf1	Conferement	Student Support	The teacher assessed my performance fairly [Die Lehrperson beurteilte meine Leistungen fair]	3.37	0.81	−1.24	0.91	3.46	0.74	−1.21	0.79
conf2	Conferement	Student Support	The teacher gave me helpful feedback on my performance [Die Lehrperson gab mir zu meinen Leistungen ein hilfreiches Feedback]	3.06	0.9	−0.63	−0.45	3.04	0.74	−0.32	−0.41
conf3	Conferement	Student Support	The teacher was fair and unbiased towards me and my classmates [Die Lehrperson hat sich mir gegenüber fair und unvoreingenommen gezeigt]	3.35	0.84	−1.21	0.74	3.52	0.74	−1.53	1.86
conf4	Conferement	Student Support	The teacher gave me meaningful feedback on my contributions [Die Lehrperson hat mir sinnvolle Rückmeldungen zu meinen Beiträgen in der Stunde gegeben]	3.24	0.83	−0.88	0.06	3.30	0.70	−0.78	0.48
capt1	Captivation	Student Support	The content of the lesson was taught by the teacher in an interesting way [Die Inhalte der Stunde wurden durch die Lehrperson auf interessante Art vermittelt]	3.09	0.87	−0.65	−0.34	3.20	0.70	−0.43	−0.36
capt3	Captivation	Student Support	I was able to see personal learning progress through the lessons [Ich konnte während der Stunde einen persönlichen Lernfortschritt feststellen]	3.03	0.89	−0.55	−0.56	3.11	0.73	−0.30	−0.68
capt4	Captivation	Student Support	The requirement level in the lesson was appropriate for me [Das Anforderungsniveau der Stunde war für mich angemessen]	3.19	0.83	−0.81	0.02	3.17	0.70	0.54	0.19

Table 2 (continued)

Item	1st-order factor	2nd-order factor	Item formulation	Students				Teachers			
				M	SD	γ_1	γ_2	*M*	SD	γ_1	γ_2
capt5	Captivation	Student Support	The learning pace in the class was appropriate for me [Das Lerntempo in der Stunde war für mich angemessen]	3.26	0.77	−0.87	0.36	3.29	0.63	−0.51	0.32
capt6	Captivation	Student Support	During the lesson I was able to apply strategies that are also useful for other problems/topics/areas [Im Unterricht konnte ich Strategien anwenden. die auch für andere Probleme/Themen/Gebiete nützlich sind]	3.21	0.86	−0.90	0.09	3.34	0.66	−0.49	−0.76
chal1	Challenge	Cognitive Activation	The tasks in the lesson were challenging for me [Die Aufgabenstellungen in der Stunde waren für mich herausfordernd]	2.98	0.87	−0.52	−0.47	3.04	0.81	−0.57	−0.17
chal2	Challenge	Cognitive Activation	The teacher had high expectations of me [Die Lehrperson hat hohe Erwartungen an mich gestellt]	3.22	0.87	−0.90	0	3.27	0.77	−0.95	0.6
clar1	Clarification	Cognitive Activation	The teacher has tied in content that was already known to me [Die Lehrperson hat an Inhalte angeknüpft. die mir schon bekannt waren]	3.23	0.83	−0.83	−0.04	3.12	0.81	−0.51	−0.58
clar2	Clarification	Cognitive Activation	The lesson had a clearly recognizable thread [Die Stunde hatte einen klar erkennbaren roten Faden]	3.05	0.90	−0.62	−0.50	3.05	0.88	−0.4	−0.95
clar3	Clarification	Cognitive Activation	The teacher showed me what the new content is related to [Die Lehrperson hat mir gezeigt. womit die neuen Inhalte zusammenhängen]	3.26	0.77	−0.81	0.16	3.37	0.65	−0.73	0.28
clar4	Clarification	Cognitive Activation	The teacher showed me what I could use the new content for [Die Lehrperson hat mir gezeigt. wofür ich die neuen Inhalte brauchen kann]	3.11	0.81	−0.67	−0.03	2.96	0.77	−0.60	0.28
capt2	Consolidation	Cognitive Activation	The course of the lesson was varied [Der Ablauf der Stunde war abwechslungsreich]	3.02	0.82	−0.54	−0.25	2.91	0.80	−0.54	0.01
cons1	Consolidation	Cognitive Activation	During the lesson. learning and practice phases alternated [In der Stunde wechselten sich Lern– und Übungsphasen ab]	3.19	0.86	−0.82	−0.12	3.21	0.83	−0.66	−0.5

Table 2 (continued)

Item	1st-order factor	2nd-order factor	Item formulation	Students				Teachers			
				M	SD	γ_1	γ_2	*M*	SD	γ_1	γ_2
cons2	Consolidation	Cognitive Activation	During the lesson. the teacher showed me exactly how I could solve certain tasks [Die Lehrperson hat mir genau gezeigt. wie ich eine bestimmte Aufgaben-stellung lösen kann]	3.10	0.84	−0.67	−0.20	3.08	0.78	−0.35	−0.71
cons3	Consolidation	Cognitive Activation	I had enough time to concentrate on the content of the lesson [Ich hatte genügend Zeit. mich intensiv mit den Inhalten der Stunde zu beschäftigen.]	3.00	0.90	−0.52	−0.59	3.15	0.77	−0.47	−0.53
cons4	Consolidation	Cognitive Activation	During the lesson there were plenty of opportunities to practice the new content [In der Stunde gab es ausreichend Gelegenheiten. die neuen Inhalte zu üben]	3.01	0.88	−0.49	−0.61	3.13	0.80	−0.53	−0.51
cont1	Classroom Management	Classroom Management	During the lesson. clear rules were discernible. which the teacher set and enforced [In der Stunde waren klare Regeln erkennbar. die die Lehrperson vorgab und durchsetzte]	3.18	0.83	−0.75	−0.17	3.05	0.71	−0.22	−0.55
cons3	Consolidation	Cognitive Activation	I had enough time to concentrate on the content of the lesson [Ich hatte genügend Zeit. mich intensiv mit den Inhalten der Stunde zu beschäftigen.]	3.00	0.90	−0.52	−0.59	3.15	0.77	−0.47	−0.53
cons4	Consolidation	Cognitive Activation	During the lesson there were plenty of opportunities to practice the new content [In der Stunde gab es ausreichend Gelegenheiten. die neuen Inhalte zu üben]	3.01	0.88	−0.49	−0.61	3.13	0.80	−0.53	−0.51
cont1	Classroom Management	Classroom Management	During the lesson. clear rules were discernible. which the teacher set and enforced [In der Stunde waren klare Regeln erkennbar. die die Lehrperson vorgab und durchsetzte]	3.18	0.83	−0.75	−0.17	3.05	0.71	−0.22	−0.55

Table 2 (continued)

Item	1st-order factor	2nd-order factor	Item formulation	Students				Teachers			
				M	SD	γ_1	γ_2	*M*	SD	γ_1	γ_2
cont2	Classroom Management	Classroom Management	The teacher did not waste time due to delays or idling [Die Lehrperson verschwendete keine Zeit durch Verzögerungen oder Leerlauf]	3.12	0.92	−0.74	−0.40	3.07	0.85	−0.46	−0.71
cont3	Classroom Management	Classroom Management	The teacher provided a trouble–free working atmosphere [Die Lehrperson hat für eine störungsfreie Arbeitsatmosphäre gesorgt]	3.17	0.85	−0.74	−0.26	3.31	0.75	−0.80	0.01
cont4	Classroom Management	Classroom Management	The teacher had a good overview of what was happening in the class [Die Lehrperson hatte einen guten Überblick über das Geschehen in der Klasse]	3.15	0.86	−0.74	−0.28	3.27	0.75	−0.73	−0.10
cont5	Classroom Management	Classroom Management	When students violated the rules. the teacher intervened quickly and consistently [Bei Regelübertretungen durch Schüler griff die Lehrperson schnell und konsequent ein]	3.10	0.86	−0.63	−0.37	3.09	0.80	−0.57	−0.23
cont6	Classroom Management	Classroom Management	The course of instruction was smooth [Die Übergänge zwischen den Phasen waren reibugslos]	3.15	0.83	−0.69	−0.19	3.22	0.79	−0.72	−0.15

Presented are translations of the original German items that are not yet validated in the English language. *SD* represents the standard deviation, γ1 the skewness and γ2 the kurtosis

Table 3 Measurement invariance for student perception and teacher self-perception

Type of invariance	χ^2	df	CFI	TLI	RMSEA	SRMR	ΔCFI	ΔRMSEA	TRd	Δdf	p
Configural invariance	3,923	712	.929	.919	.046	.036					
Metric invariance	3,963	734	.929	.921	.046	.038	0	0	35.54	22	< .05
Scalar invariance	4,039	763	.928	.923	.045	.038	.001	–.001	57.54	29	< .001
Strict invariance	4,126	792	.926	.924	.045	.040	.002	0	80.92	29	< .001

TRd, Satorra-Bentler scaled chi-square difference test

Table 4 Descriptive statistics for gender

Gender	*n*	Self-perception			Student perception		
		Classroom management	Cognitive activation	Student support	Classroom management	Cognitive activation	Student support
		M(SD)	M(SD)	M(SD)	M(SD)	M(SD)	M(SD)
Female	87	−0.09 (0.49)	−0.04 (0.29)	−0.06 (0.46)	0.07 (0.26)	0.07 (0.21)	0.08 (0.22)
Male	84	0.09 (0.41)	0.05 (0.28)	0.06 (0.43)	−0.08 (0.48)	−0.08 (0.39)	−0.08 (0.38)

M refers to factor score means

Table 5 Multiple regression for gender and student perception on teachers' self-perceptions

Scale	Predictor	Standardized estimate	*SE*	R^2
Classroom management	Gender	.30***	.06	.36
	Student perception	.58***	.07	
Student support	Gender	.25***	.03	.18
	Student perception	.41***	.04	
Cognitive activation	Gender	.25***	.04	.22
	Student perception	.46***	.06	

Dummy-coding for gender. 0: female, 1: male

*$p < .05$, **$p < .01$, ***$p < .001$

Table 6 Comparison of fit indices and likelihood ratio tests for different number of classes

Model	Fit indices			Likelihood ratio tests			Number of subjects in class											
	BIC	aBIC	AIC	VLM*p*R	*p* LMR	Entropy	1	2	3	4	5	6	7	8	9	10	11	12
1-class	892	853.98	854.30															
2-class	484	424.15	424.60	0.01	0.01	0.94	54	117										
3-class	331	248.54	249.20	0.43	0.44	0.89	44	75	52									
4-class	**245**	**140.90**	**141.70**	**0.04**	**0.04**	**0.92**	**18**	**36**	**69**	**48**								
5-class	169	42.19	43.20	0.14	0.14	0.93	30	18	50	53	20							
6-class	149	−0.04	1.10	0.40	0.40	0.94	21	52	10	45	23	20						
7-class	128	−42.58	−41.20	0.35	0.36	0.94	10	22	9	45	21	44	20					
8-class	103	−90.54	−89.00	0.37	0.37	0.93	18	15	9	24	26	20	19	40				
9-class	77	−138.27	−136.60	0.48	0.49	0.94	10	17	25	14	15	21	9	41	19			
10-class	73	−164.36	−162.50	0.56	0.57	0.94	15	18	10	2	37	25	18	21	14	11		
11-class	63	−196.49	−194.50	0.43	0.43	0.94	27	15	17	10	4	17	15	16	2	14	34	
12-class	65	−217.16	−215.00	0.75	0.75	0.95	10	14	4	24	1	17	1	15	35	20	11	19

BIC: Bayes's information criterion; aBIC: sample size-adjusted Bayesian Information Criterion; *AIC*: Akaike's information criterion; *p VLMR*: Vuongh likelihood ratio test for n versus n–1 classes; *p LMR*: Lo-Mendell-Rubin likelihood ratio test for n versus n–1 classes

Table 7 Association of perception profile and gender

Gender	*n*	Class 1 (%)	Class 2 (%)	Class 3 (%)	Class 4 (%)
Male	84	20.24	25.00	23.81	30.95
Female	87	1.15	17.24	56.32	25.29

Table 8 Association of perception profile and subject

Subject	*N*	Class 1 (%)	Class 2 (%)	Class 3 (%)	Class 4 (%)
Maths, physics, and IT	26	11.54	30.77	38.46	19.23
Biology and chemistry	16	12.50	12.50	37.50	37.50
German	18	27.78	11.11	33.33	27.78
English, French, and Spanish	28	3.57	28.57	28.57	39.29
Geography, economics, and history	16	6.25	25.00	56.25	12.50

Table 9 Association of perception profile and gender and subject

Subject	Gender	*N*	Class 1 (%)	Class 2 (%)	Class 3 (%)	Class 4 (%)
Maths, physics, and IT	Male	7	28.57	42.86	0.00	28.57
	Female	19	5.26	26.32	52.63	15.79
Biology and chemistry	Male	6	33.33	16.67	16.67	33.33
	Female	10	0.00	10.00	50.00	40.00
German	Male	10	50.00	20.00	20.00	10.00
	Female	8	0.00	0.00	50.00	50.00
English, French, and Spanish	Male	16	6.25	31.25	18.75	43.75
	Female	12	0.00	25.00	41.67	33.33
Geography, economics, and history	Male	7	14.29	28.57	42.86	14.29
	Female	9	0.00	22.22	66.67	11.11

Funding Open Access funding enabled and organized by Projekt DEAL.

References

Ackerman, L., Beier, E., & Bowen, R. (2002). What we really know about our abilities and our knowledge. *Personality and Individual Differences, 33*, 587–605. https://doi.org/10.1016/S0191-8869(01)00174-X

Antoniou, P., & Kyriakides, L. (2011). The impact of a dynamic approach to professional development on teacher instruction and student learning: Results from an experimental study. *School Effectiveness and School Improvement, 22*(3), 291–311.

Balch, R. T. (2012). *The validation of a student survey on teacher practice*. Vanderbilt University.

Bandura, A. (1977). Self-efficacy. Toward a unifying theory of behavioral change. *Psychological Review, 84*, 191–215.
Bandura, A. (1997). *Self-efficacy: The exercise of control*. Freeman.
Benade, L. (2015). Teachers' critical reflective practice in the context of twenty-first century learning. *Open Review of Educational Research, 2*(1), 42–54.
Bengtsson, J. (2003). Possibilities and limits of self-reflection in the teaching profession. *Studies in Philosophy and Education, 22*, 295–316. https://doi.org/10.1023/A:1022813119743
Beyer, S., & Bowden, E. (1997). Gender differences in self-perceptions. *Personality and Social Psychology Bulletin, 23*, 157–172. https://doi.org/10.1177/0146167297232005
Boring, A. (2017). Gender biases in student evaluations of teachers. *Journal of Public Economics, 145*, 27–41. https://doi.org/10.1016/j.jpubeco.2016.11.006
Braga, M., Paccagnella, M., & Pellizzari, M. (2014). Evaluating students' evaluations of professors. *Economics of Education Review, 41*, 71–88. https://doi.org/10.1016/j.econedurev.2014.04.002
Brett, J., & Atwater, L. (2001). 360° feedback: Accuracy, reactions, and perceptions of usefulness. *Journal of Applied Psychology, 86*, 930–942. https://doi.org/10.1037/0021-9010.86.5.930
Cheung, G. W., & Rensvold, R. B. (2002). Evaluating goodness-of-fit indexes for testing measurement invariance. *Structural Equation Modeling: A Multidisciplinary Journal, 9*(2), 233–255.
Clausen, M. (2002). *Unterrichtsqualität: Eine Frage der Perspektive? [Instructional quality: A question of perspective?]*. Waxmann.
Cruickshank, D. R. (1987). *Reflective teaching: The preparation of students of teaching*. Association of Teacher Educators.
De Jong, R., & Westerhof, K. J. (2001). The quality of student ratings of teacher behaviour. *Learning Environments Research, 4*(1), 51–85.
Den Brok, P., Bergen, T., & Brekelmans, M. (2006). Convergence and divergence between teachers' and students' perceptions of instructional behavior in Dutch secondary education. In D. L. Fisher & M. S. Khine (Eds.), *Contemporary approaches to research on learning environments: Worldviews* (pp. 125–160). World Scientific.
Dunlosky, J., & Rawson, K. (2012). Overconfidence produces underachievement: Inaccurate self-evaluations undermine students' learning and retention. *Learning and Instruction, 22*, 271–280. https://doi.org/10.1016/j.learninstruc.2011.08.003
Dunning, D., Johnson, K., Ehrlinger, J., & Kruger, J. (2003). Why people fail to recognize their own incompetence. *Current Directions in Psychological Science, 12*, 83–87. https://doi.org/10.1111/1467-8721.01235
Ehrlinger, J., & Dunning, D. (2003). How chronic self-views influence (and potentially mislead) estimates of performance. *Journal of Personality and Social Psychology, 84*, 5–18. https://doi.org/10.1037/0022-3514.84.1.5
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 [journal for Pedagogy], 66*(Beiheft 1/20), 138–155.
Fauth, B., Decristan, J., Rieser, S., Klieme, E., & Büttner, G. (2014). Student ratings of instructional 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
Ferguson, R. F. (2012). Can student surveys measure teaching quality? *Phi Delta Kappan, 94*(3), 24–28.
Fox, E., Ridgewell, A., & Ashwin, C. (2009). Looking on the bright side. *Biological Sciences, 276*, 1747–1751. https://doi.org/10.1098/rspb.2008.1788
Fraser, B. (1991). Two decades of classroom environment research. In B. J. Fraser & H. J. Walberg (Eds.), *Educational environments: Evaluation, antecedents, and consequences* (pp. 3–27). Pergamon.
Fraser, B. J. (2007). Classroom learning environments. In S. K. Abell & N. G. Lederman (Eds.), *Handbook of research on science education* (pp. 103–124). Routledge.
Gaertner, H. (2014). Effects of student feedback as a method of self-evaluating the quality of teaching. *Studies in Educational Evaluation, 42*, 91–99.
Goodlad, J. I. (1984). *A place called school: Prospects for the future*. McGraw-Hill.
Hacker, D., Bol, L., Horgan, D., & Rakow, E. (2000). Test prediction and performance in a classroom context. *Journal of Educational Psychology, 92*, 160–170. https://doi.org/10.1037//0022-0663.92.1.160
Hattie, J. (2009). *Visible learning: A synthesis of over 800 meta-analyses relating to achievement*. London: Routledge.
Hattie, J., & Timperley, H. (2007). The power of feedback. *Review of Educational Research, 77*, 81–112. https://doi.org/10.3102/003465430298487

Helmke, A., & Lenske, G. (2013). Unterrichtsdiagnostik als Grundlage für Unterrichtsentwicklung [Instructional diagnostics as a basis for instructional development]. *Beiträge Zur Lehrerbildung [contributions to Teacher Education], 31*(2), 214–233.

Helzer, E. G., & Dunning, D. (2012). Why and when peer prediction is superior to self-prediction: The weight given to future aspiration versus past achievement. *Journal of Personality and Social Psychology, 103*(1), 38.

Hessler, M., Pöpping, D. M., Hollstein, H., Ohlenburg, H., Arnemann, P. H., Massoth, C., Seidel, L. M., Zarbock, A., & Wenk, M. (2018). Availability of cookies during an academic course session affects evaluation of teaching. *Medical Education, 52*(10), 1064–1072.

Kane, T. J., & Staiger, D. O. (2012). *Gathering feedback for teaching: Combining high-quality observations with student surveys and achievement gains*. Bill & Melinda Gates Foundation: MET Project.

Klieme, E., Lipowsky, F., Rakoczy, K., & Ratzka, N. (2006). Qualitätsdimensionen und Wirksamkeit von Mathematikunterricht [Quality dimensions and effectiveness of mathematics teaching]. In M. Prenzel & L. Allolio-Näcke (Eds.), *Untersuchungen zur Bildungsqualität von Schule* (pp. 127–146). Waxmann.

Kline, R. B. (2016). *Principles and practice of structural equation modeling* (4th ed.) (Methodology in the Social Sciences). The Guilford Press.

Kluger, A., & DeNisi, A. (1996). The effects of feedback interventions on performance. *Psychological Bulletin, 119*, 254–284. https://doi.org/10.1037/0033-2909.119.2.254

Kolar, D., Funder, D., Colvin, C. (1996). Comparing the accuracy of personality judgments by the self and knowledgeable others. *Journal of Personality, 64*, 311–337. https://doi.org/10.1111/j.1467-6494.1996.tb00513.x

Kornell, N., & Hausman, H. (2016). Do the best teachers get the best ratings? *Frontiers in Psychology, 7*, 570. https://doi.org/10.3389/fpsyg.2016.00570

Krueger, J., & Mueller, R. (2002). Unskilled, unaware, or both? The better-than-average heuristic and statistical regression predict errors in estimates of own performance. *Journal of Personality and Social Psychology, 82*, 180–188. https://doi.org/10.1037/0022-3514.82.2.180

Kruger, J., & Dunning, D. (1999). Unskilled and unaware of it. *Journal of Personality and Social Psychology, 77*, 1121–1134.

Kunter, M., & Voss, T. (2013). The model of instructional quality in COACTIV. In M. Kunter, J. Baumert, W. Blum, U. Klusmann, S. Krauss, & M. Neubrand (Eds.), *Cognitive activation in the mathematics classroom and professional competence of teachers* (pp. 97–124). Springer. https://doi.org/10.1007/978-1-4614-5149-5_6

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., Tsai, Y.-M., Klusmann, U., Brunner, M., Krauss, S., & Baumert, J. (2008). Students' and mathematics teachers'perceptions of teacher enthusiasm and instruction. *Learning and Instruction, 18*, 468–482. https://doi.org/10.1016/j.learninstruc.2008.06.008

Kyriakides, L., Creemers, B., Panayiotou, A., Vanlaar, G., Pfeifer, M., Cankar, G., & McMahon, L. (2014). Using student ratings to measure quality of teaching in six European countries. *European Journal of Teacher Education, 37*, 125–143. https://doi.org/10.1080/02619768.2014.882311

Lamb, J. (2017). How do teachers reflect on their practice? A study into how feedback influences teachers' reflective practice. *The STeP Journal (student Teacher Perspectives), 4*(4), 94–104.

Larrivee, B. (2006). The convergence of reflective practice and effective classroom management. In C. M. Evertson & C. S. Weinstein (Eds.), *Handbook of classroom management: Research, practice, and contemporary issues* (pp. 983–1001). Lawrence Erlbaum.

Lindeman, M., Sundvik, L., & Rouhiainen, P. (1995). Under- or over-estimation of self? *Journal of Social Behavior and Personality, 10*, 123–134.

Lipowsky, F. Drollinger-Vetter, B., Klieme, E., Pauli, C. & Reusser, K. (2018). Generische und fachdidaktische Dimensionen von Unterrichtsqualität – Zwei Seiten einer Medaille? [Generic and subject didactic dimensions of teaching quality – Two sides of the same coin?]. In M. Martens, K. Rabenstein, K. Bräu, M. Fetzer, H. Gresch, I. Hardy & C. Schelle (Eds.), *Konstruktionen von fachlichkeit* [Constructions of subject matter](pp. 183–202). Klinkhardt.

Lundeberg, M. A., Fox, P. W., Brown, A. C., & Elbedour, S. (2000). Cultural influences on confidence. *Journal of Educational Psychology, 92*, 152–159. https://doi.org/10.1037/0022-0663.92.1.152

Lundeberg, M., Fox, P., & Punćcohaŕ, J. (1994). Highly confident but wrong: Gender differences and similarities in confidence judgments. *Journal of Educational Psychology, 86*, 114–121. https://doi.org/10.1037/0022-0663.86.1.114

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.

Leising, D., Rehbein, D., & Sporberg, D. (2006). Does a fish see the water in which it swims? A study of the ability to correctly judge one's own interpersonal behavior. *Journal of Social and Clinical Psychology, 25*(9), 963–974.

Makarova, E., Aeschlimann, B., & Herzog, W. (2019). The gender gap in STEM fields: The impact of the gender stereotype of math and science on secondary students' career aspirations. *Frontiers in Education*. https://doi.org/10.3389/feduc.2019.00060

Maulana, R., Opdenakker, M., Brok, P., & Bosker, R. J. (2012). Teacher–student interpersonal relationships in Indonesian lower secondary education: Teacher and student perceptions. *Learning Environments Research, 15*, 251–271. https://doi.org/10.1007/s10984-012-9113-7

McIntosh, R., Fowler, E., Lyu, T., & Della Sala, S. (2019). Wise up: Clarifying the role of metacognition in the Dunning–Kruger effect. *Journal of Experimental Psychology: General, 148*, 1882–1897. https://doi.org/10.1037/xge0000579

Meade, A., Johnson, E., & Braddy, P. (2008). Power and sensitivity of alternative fit indices in tests of measurement invariance. *Journal of Applied Psychology, 93*, 568–592. https://doi.org/10.1037/0021-9010.93.3.568

Miller, T., & Geraci, L. (2011). Unskilled but aware: Reinterpreting overconfidence in low-performing students. *Journal of Experimental Psychology, 37*, 502–506. https://doi.org/10.1007/s11409-011-9083-7

Mitchell, K. M., & Martin, J. (2018). Gender bias in student evaluations. *Political Science & Politics, 51*, 648–652. https://doi.org/10.1017/S104909651800001X

Mosing, M., Zietsch, B., Shekar, S., Wright, M., & Martin, N. (2009). Genetic and environmental influences on optimism and its relationship to mental and self-rated health. *Behavior Genetics, 39*, 597–604.

Muthen, L. K., & Muthen, B. (2012–2019). *MPlus Version 8.4*. Los Angeles, CA: Muthén & Muthén.

Pham, G., Koch, T., Helmke, A., Schrader, F., Helmke, T., & Eid, M. (2012). Do teachers know how their teaching is perceived by their pupils? *Procedia-Social and Behavioral Sciences, 46*, 3368–3374. https://doi.org/10.1016/j.sbspro.2012.06.068

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

Röhl, S., & Rollett, W. (2021). Jenseits von Unterrichtsentwicklung: Intendierte und nicht-intendierte Nutzungsformen von Schülerfeedback durch Lehrpersonen [Beyond instructional development: Intended and unintended uses of student feedback by teachers]. In K. Göbel, C. Wyss, K. Neuber, & M. Raaflaub (Eds.), *Quo vadis Forschung zu Schülerrückmeldungen zum Unterricht* [Quo vadis research on student feedback on instruction]. Springer. https://doi.org/10.1007/978-3-658-32694-4.

Röhl, S., & Gärtner, H. (2021). Relevant conditions for teachers' use of student feedback. In W. Rollett, H. J. E. Bijlsma, & S. Röhl (Eds.), *Student feedback on teaching in schools: Using student perceptions for the development of teaching and teachers.* Cham: Springer.

Ross, J., & Bruce, C. (2007). Teacher self-assessment: A mechanism for facilitating professional growth. *Teaching and Teacher Education, 23*, 146–159. https://doi.org/10.1016/j.tate.2006.04.035

Scheeler, M. C., Ruhl, K. L., & McAfee, J. K. (2004). Providing performance feedback to teachers: A review. *Teacher Education and Special Education, 27*(4), 396–407.

Schön, D. (1983). *The reflective practitioner*. Temple Smith.

Simons, D. (2013). Unskilled and optimistic: Overconfident predictions despite calibrated knowledge of relative skill. *Psychonomic Bulletin & Review, 20*, 601–607. https://doi.org/10.3758/s13423-013-0379-2

Seidel, T., & Shavelson, R. J. (2007). Teaching effectiveness research in the past decade: The role of theory and research design in disentangling meta-analysis results. *Review of Educational Research, 77*(4), 454–499.

Slavin, R. E. (1994). Quality, appropriateness, incentive, and time: A model of instructional effectiveness. *International Journal of Educational Research, 21*(2), 141–157.

Smyth, J. (1992). Teachers' work and the politics of reflection. *American Educational Research Journal, 29*, 267–300.

van Petegem, P., Deneire, A., & de Maeyer, S. (2008). Evaluation and participation in secondary education: Designing and validating a self-evaluation instrument for teachers to solicit feedback from pupils. *Studies in Educational Evaluation, 34*, 136–144.

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. https://doi.org/10.1037/a0017908

Wagner, W., Göllner, R., Helmke, A., Trautwein, U., & Lüdtke, O. (2013). Construct validity of student perceptions of instructional quality is high, but not perfect. *Learning and Instruction, 28*, 1–11. 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. *Journal of Educational Psychology, 108*, 705–721. https://doi.org/10.1037/edu0000075

Wettstein, A., Ramseier, E., Scherzinger, M., & Gasser, L. (2016). Unterrichtsstörungen aus Lehrer-und Schülersicht [Teaching disorders from teacher and student perspective]. *Zeitschrift Für Entwicklungspsychologie Und Pädagogische Psychologie, 48*, 171–183. https://doi.org/10.1026/0049-8637/a000159

Wisniewski, B., & Zierer, K. (2020). Entwicklung eines Online-Fragebogens zur Erhebung von Unterrichtsqualität durch Lernendenfeedback und erste Validierungsschritte [Development of an online questionnaire to assess instructional quality through learner feedback and initial validation steps]. *Psychologie Für Erziehung Und Unterricht, 67*, 138–155. https://doi.org/10.2378/peu2020.artnd

Wisniewski, B., Zierer, K., Dresel, M., & Daumiller, M. (2020). Obtaining students' perceptions of instructional quality – Two-level structure and measurement invariance. *Learning and Instruction*. https://doi.org/10.1016/j.learninstruc.2020.101303

Youmans, R. J., & Jee, B. D. (2007). Fudging the numbers: Distributing chocolate influences student evaluations of an undergraduate course. *Teaching of Psychology, 34*, 245–247. https://doi.org/10.1080/00986280701700318

Zell, E., & Krizan, Z. (2014). Do people have insight into their abilities? *Perspectives on Psychological Science, 9*, 111–125. https://doi.org/10.1177/1745691613518075