Educational Assessment, Evaluation and Accountability (2025) 37:35–71
https://doi.org/10.1007/s11092-024-09450-9

Can feedback from students to teachers improve different dimensions of teaching quality in primary and secondary education? A hierarchical meta-analysis

Sebastian Röhl[1] · **Hannah Bijlsma**[2] · **Martin Schwichow**[3]

Received: 29 August 2023 / Accepted: 26 November 2024 / Published online: 25 January 2025

Abstract

This meta-analysis summarizes the available evidence on the effectiveness of student feedback interventions in primary and secondary schools on different aspects of teaching quality. It aims to examine indications of success conditions for an effective use of student feedback in practice. The analysis included 23 studies with 314 effect sizes. We estimated an overall weighted mean effect size of $d = 0.27$ based on a three-level random-effects model. For effectiveness with respect to various dimensions of teaching quality, only minor and mostly nonsignificant differences were found. Further moderator analyses showed larger effect sizes for studies in which teachers received support for interpreting student feedback and implementing changes in their teaching. Moreover, studies in which teachers were encouraged to discuss the feedback with their students have larger effect sizes than studies in which teachers were not encouraged to do so. Item and answering scale characteristics also showed significant effects on the effectiveness of student feedback. Implications for further research and the use of student feedback in schools are discussed.

Keywords Student feedback · Teaching development · Teaching quality · Meta-analysis

✉ Sebastian Röhl
sebastian.roehl@ph-freiburg.de

[1] Institute for Education, University of Education Freiburg, Kunzenweg 21, 79117 Freiburg, Germany

[2] Radboud University, Behavioural Science Institute, P.O. Box 9104, NL-6500 HE Nijmegen, The Netherlands

[3] Department of Physics and Physics Education, University of Education Freiburg, Kunzenweg 21, 79117 Freiburg, Germany

1 Introduction

Based on almost one hundred years of feedback research in the field of education (Remmers, 1927; Stalnaker & Remmers, 1928), we know that feedback, when designed appropriately, can have a strong performance-enhancing effect (Hattie & Timperley, 2007; Ilgen et al., 1979; Kluger & DeNisi, 1996). Feedback can be understood as a communicative process "in which some sender […] conveys a message to a recipient. In the case of feedback, the message comprises information about the recipient" (Ilgen et al., 1979, p. 350). This information, when understood and accepted by the recipient, can be used to improve task performance (Kluger & DeNisi, 1996) or to initiate learning processes (Hattie & Timperley, 2007).

In education, however, teachers do not obtain much structured feedback on their performance. This issue is particularly critical because teachers with poorer performance tend to overestimate the quality of their own teaching (Wisniewski et al., 2022). One way for teachers to obtain feedback is through lesson observations (Lasagabaster & Sierrra, 2011). However, reliable and valid observation scores of teaching quality require multiple observations and multiple trained raters (Hill et al., 2012), which can make lesson observations expensive and time-consuming. Although teachers could use student performance as a way of feedback on the effectiveness of their lessons, an increase or decrease in student performance does not provide specific feedback: it does not say anything about the teacher's behavior or what he or she can improve.

Another way to obtain feedback is by collecting student perceptions of teaching. Student perceptions have been found to be a valid indicator of the quality of teaching, as they have been shown in many studies to have significant effects on various outcomes, such as learning achievement (e.g., Arens and Möller, 2016; Baumert & Kunter, 2013; Fauth et al., 2014; Ferguson, 2012; Kuhfeld, 2017; Stahns et al., 2020; Wang et al., 2022) and learning engagement (e.g., Maulana & Helms-Lorenz, 2016; van der Grift et al., 2014). Moreover, student perceptions show a much higher predictive validity on many learning outcomes compared to teacher ratings of their own teaching quality (e.g., Seidel & Shavelson, 2007). In addition, research shows that students' ratings vary primarily as a function of the teachers' teaching skills (Benton & Cahsin, 2012; Richardson, 2005). Additionally, student perceptions showed significant correlations with ratings by trained classroom observers (e.g., Kuhfeld, 2017; van der Scheer et al., 2019). Further, since students spend many hours in a teacher's classroom, student feedback arguably mitigates the need for additional time-consuming classroom observations. Therefore, student ratings can be an effective way to provide teachers with feedback on what impact their teaching has and can serve as a valuable source of information for teachers that encourages them to reflect on their teaching practices, which has the potential to lead to an improvement process of the quality of teaching through improvement-oriented actions (Muijs, 2006; Peterson et al., 2000).

Student feedback to teachers about their teaching quality has been studied for almost 100 years, starting with the first experiments in higher education in the

1920s (Stalnaker & Remmers, 1928) and later in schools (Porter, 1942; Remmers, 1934). At the beginning of the 1960s, several studies were conducted that used student perceptions of the teaching or the teacher as feedback for the teacher (e.g., Gage, 1960; Ryan, 1974; Veldman & Peck, 1969). The findings of these studies showed the value of student ratings in stimulating a teacher's desire to improve: "It provides the compass check for 'come alive' teaching. Certain actions are observable and appraisal of these actions by students can point the way to master teaching" (Williams, 1962, p. 284). Since then, interventions that use student ratings for instructional improvement have been studied multiple times in the research on teaching effectiveness (Bartel, 1970; Bijlsma et al., 2019; Tozoglu, 2006), learning environments (e.g., Bell & Aldridge, 2014; den Brok et al., 2006; Fraser et al., 1982), and teaching quality (e.g., Ditton & Arnold, 2004; Ferguson, 2012; Tuckman & Olivier, 1968). Additionally, student feedback is used during the practical phases of teacher training as an improvement and reflection tool for pre-service teachers (Göbel et al., 2021; Ryan, 1974).

In higher education, using feedback from students to give information to teachers on their instructional practices is well established as part of performance appraisal (*Student Evaluations of Teaching [SET]*; Lang & Kersting, 2007; Penny & Coe, 2004). In this field, many meta-analyses with a long tradition demonstrate a positive effect of student feedback on the quality of teaching and lecturing (Cohen, 1980; L'Hommedieu et al. 1990; Penny & Coe, 2004). Concerning K-12 education, until now, only one meta-analysis exists (Röhl, 2021) that reports a positive effect ($d = 0.21$) of student feedback on teaching quality. The existing meta-analyses on student feedback in higher and K-12 education have used only the average effect size per study, ignoring the existing heterogeneity of effect sizes reported for different aspects of teaching quality within studies or intervention groups. For practical use, however, it is relevant to know whether student feedback interventions are equally suitable for the improvement of all teaching areas. This might not be the case, for example, if student perceptions show different qualities for different aspects of teaching or if they require various efforts to improve (cf. 2.2). Moreover, the effect of student feedback might also be influenced by several other factors that have, so far, received less attention in the literature, such as whether student feedback was used in practical phases of teacher training, the different characteristics of the feedback intervention (such as the questionnaire used), the frequency of the feedback, or the experimental conditions of different intervention groups.

Therefore, this meta-analysis addresses the questions of whether student feedback proves equally beneficial for different aspects of teaching, and which characteristics of student feedback interventions emerge as particularly effective. First, we describe how teachers collect, interpret, and use feedback from students to improve their quality of teaching based on the framework of the Process Model of Student Feedback on Teaching (SFT; Röhl et al., 2021). This model provides indications of possible factors influencing the effectiveness of student feedback interventions. Based on this theoretical model, we subsequently discuss empirical findings regarding the quality and usefulness of student feedback in improving teaching quality.

2 Theoretical background

2.1 The student feedback process

When using student feedback for the development of teaching and teachers, it is assumed that student perceptions and assessments of teaching provide valuable information for improving teaching. The teacher (or a third person) asks the students in a class about their perceptions of the teaching using a questionnaire. The teacher identifies information contained in the feedback regarding areas of improvement for his or her teaching and implements appropriate strategies and methods to address the areas of improvement, which may then lead to a higher quality of teaching.

In the Process Model of Student Feedback on Teaching (SFT; Fig. 1; Röhl et al., 2021), this process and its influencing conditions are described in more detail, summarizing various findings from organizational and school feedback research (among others, Ilgen et al., 1979; Kahmann & Mulder, 2011; Kluger & DeNisi, 1996; Smither et al., 2005). The first step is to obtain information about teaching and classes from the students' perspective. Thereby, the nature and quality of the survey instrument play an important role, including such factors as its theoretical foundation, validity, and item formulations, as well as the aspects of teaching that are being surveyed (questionnaire characteristics). In addition, student and class characteristics may have an impact on the information obtained (e.g., the subject taught or the socioeconomic background of the students). In the next step, the collected information must be understood and interpreted by the teacher, which not only relates to their cognition but also involves their affective reactions. This can lead to teachers having better knowledge of their own teaching and how it is perceived by

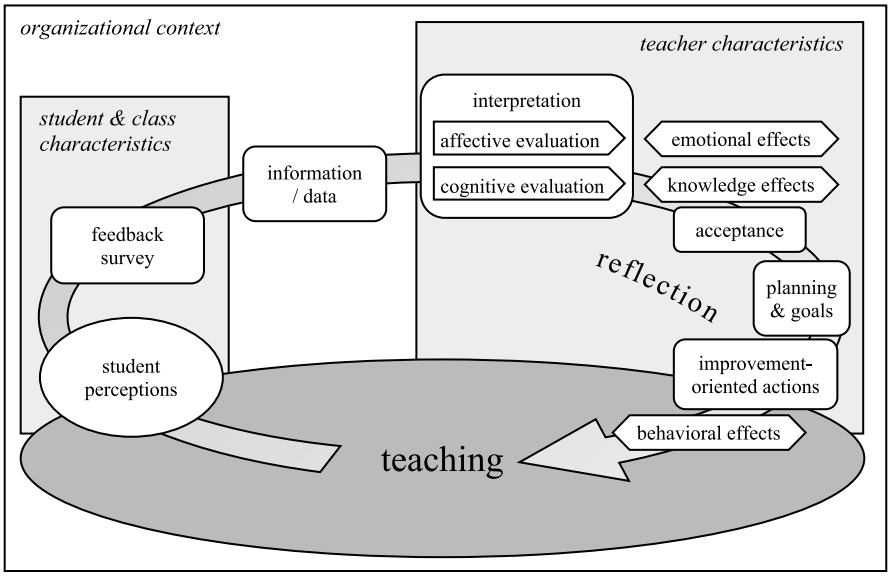

Fig. 1 Process model of student feedback on teaching (SFT; Röhl et al., 2021)

their students. To actually bring about changes in teacher behavior, this information must first be seen by the teacher as valid and relevant, or in other words, accepted. Furthermore, teachers must develop the intention to improve their teaching and set goals for themselves, which, in turn, allows for improvement-oriented actions to take place, such as increased attention to improvement aspects in the preparation of lessons or during teaching (Röhl et al., 2021), discussions for collaborative improvement with students about the obtained feedback (Gaertner, 2014), participation in special professional learning courses (Balch, 2012), or adapting the teaching more to the students' needs (Gaertner, 2014).

For such an intervention process to have measurable effects on teaching (i.e., to be perceived by the students or external observers, or to show a higher learning achievement), a large number of processing steps are necessary. The model points particularly to the important role of the teacher as the "bottleneck" of the feedback process. On the one hand, the individual characteristics of the teacher are relevant conditions for a successful teaching-improvement process. These include, for example, data literacy, self-efficacy expectations, attributional tendencies, knowledge of content, mastery goal orientation, and the relevance and importance that teachers attribute to student feedback (for an overview, see Röhl & Gärtner, 2021a). On the other hand, these individual characteristics of teachers can be influenced by favoring circumstances in the organization or the design of the intervention. These include, for example, helping to support and supervise teachers in the feedback and development process, encouraging specific instructional changes, the provision of optimal professional learning, or the design of feedback reports.

2.2 Teaching quality

Central to student feedback interventions is the improvement of teaching quality. Therefore, in the following, we present the theoretical background of conceptualizations of teaching quality and its improvement.

2.2.1 Conceptions and aspects of teaching quality

In order to assess the quality of teaching, theoretical assumptions are required about which characteristics of teaching can be considered conducive to learning. In addition, student assessments of these features form the basis for feedback to the teacher.

The definition of teaching quality can be based on scientifically determined characteristics of effective lessons (teacher effectiveness research), as the study of effective classroom practices has been central to the measurement of teaching quality (Reynolds et al., 2014). Within this field of research, a variety of teacher behaviors that promote student learning have been identified (Creemers, 1994; Fauth et al., 2014; Hattie, 2009; Muijs et al., 2014; Pianta & Hamre, 2006; Reynolds et al., 2014; Sammons et al., 1995; van de Grift, 2007). Some authors list factor-analytically obtained dimensions of teaching quality. According to Day et al. (2008), for example, effective teaching approaches can be categorized into three dimensions: creating a supportive classroom climate, having a well-organized and structured lesson (classroom

management), and having a clear instructional approach. Other frameworks partly overlap with this or add other aspects, e.g., the three generic dimensions of teaching quality that are very widespread in Germany (Praetorius et al., 2018), supportive climate, classroom management, and cognitive activation, or the five teaching dimensions of van de Grift (2007), efficient classroom management, safe and stimulating learning climate, clear instructions, adaptation of teaching, and teaching learning strategies. Two approaches to unify the various categorizations of teaching quality aspects are the synthesis of classroom observation system frameworks by Praetorius and Charalambous, (2018) and the teacher effectiveness characteristics by van de Grift et al. (2014). In this meta-analysis, these two frameworks were used as the basis for the teaching quality categorization scheme and extended (cf. 3.2).

2.2.2 Improving teaching quality aspects

The heterogeneity of teaching quality conceptualizations in different studies (2.2.1) indicates that this is not a uniform construct with precisely defined characteristics of teaching that are considered positive for learning. In most of the student feedback intervention studies, several and varying effect sizes are reported for the different teaching quality aspects. This means that an intervention aimed at improving teaching quality might not be as effective for one aspect as it is for another. One possible explanation for this could be that some teaching quality aspects require the improvement of complex teaching skills, for example, differentiated instruction (van de Grift et al., 2014). This is supported by the stage-wise professional development of the pedagogical didactic skills of teachers, which starts with the development of more basic teaching quality aspects, such as creating a safe atmosphere, and ends with the development of more complex teaching quality aspects, such as clear instruction, differentiation, and asking questions that cognitively activate students. In addition, Kyriakides et al. (2009) report a similar cumulative order, using student observations of primary education teachers. Van der Lans and colleagues (2018, 2019) also found that effective teaching strategies of secondary school teachers can be ordered in terms of their difficulty in implementation in the classroom and that not every teaching quality aspect improves to a similar degree.

Another reason for differential effectiveness of student feedback interventions on different aspects of instruction could be due to students' idiosyncratic perceptions of teaching (Göllner et al., 2018). Students usually do not have the same knowledge about modern and professional methodological and didactic approaches as trained external observers. Therefore, when it comes to the effect measurements of an intervention, students may perceive less improvement in some aspects of teaching quality compared to what an external observer, or "expert," might see (Strong et al., 2011).

2.3 Possible moderators in the student feedback process

Based on the SFT model, we discuss possible relevant characteristics of student feedback interventions and studies that may have an influence on the effect and are therefore considered in our meta-analysis.

2.3.1 Measurement of teaching quality: characteristics of questionnaires and feedback reports

A valid and appropriate feedback questionnaire provides an important basis for the feedback process. In the item response process model, Tourangeau et al. (2000) distinguish four necessary steps that students also go through when responding to item-based feedback questionnaires: Students have to *comprehend* the item and to *retrieve* the relevant and available information. Subsequently, they have to *judge* the retrieved information with regard to the question and *respond* by choosing an option. Therefore, item formulations and response scales in feedback questionnaires can influence the quality and the meaningfulness of the included information.

For example, item formulations used in feedback questionnaires can differ with respect to the *item referent*, which means the subject to which an item refers (den Brok et al., 2006; Fauth et al., 2020). Items may be formulated in reference to the teaching or the class in general ("In math class, the lesson is often disrupted"), the teacher ("The teacher explains things very clearly"), the students ("Students who work faster than others move on to the next topic"), or to the individual address ("I have enough time to work on new things I learn"). In the item response process, different item referents could trigger students to retrieve different information and, therefore, lead to different rating tendencies (for an overview, see Fauth et al., 2020; Göllner et al., 2020).

Additionally, feedback questionnaires differ in the applied *response scale*. Thus, they offer specific and limited options for the student response process, which could possibly influence the way the information retrieved is judged by the students. Many instruments are based on agreement scales (e.g., "strongly disagree" to "strongly agree"), but instruments also use frequency scales ("never" to "always"), grading scales ("poor" to "very good"), or even a comparison to other teachers ("below average" to "above average"). Saris et al. (2010) showed that, for many countries, responses to agreement scales have significantly lower reliability compared to responses to item-specific scales like frequency scales. For example, agreement scales are regarded as more inferential than frequency scales (Tourangeau et al., 2000; Wagner, 2008) and show a higher risk for acquiescence bias (Saris et al., 2010).

Moreover, the wording of the response categories may also influence teachers' acceptance and use of the feedback. Theories on the effectiveness and use of feedback indicate that critical feedback is particularly effective when the feedback recipient, who is the teacher, considers the cause to be variable and changeable (Hattie & Timperley, 2007; Kluger & DeNisi, 1996). Therefore, different item referents in the questionnaire might elicit different reactions from the teacher as the feedback recipient; for example, using the teacher as the item referent could lead teachers to understand this as feedback related to their own person, which is seen as stable and difficult to change; they are therefore more likely to react dismissively to the feedback (Kluger & DeNisi, 1996). In contrast, critical feedback from items related to students might be interpreted by teachers as not being changeable, such that they are not motivated to improve their teaching (Bertrand & Marsh, 2015; Weiner, 1985). In addition, when using grading or comparison answering scales, negative feedback

may also be more likely interpreted as feedback on the person level (Hattie & Timperley, 2007; Kluger & DeNisi, 1996). Responses on frequency scales might be perceived as more objective information, which furthermore can be discussed with the students as feedback givers. Overall, it can be presumed that more usable information for instructional improvement can be obtained when using agreement and frequency scales instead of grading or comparison scales.

Whether questionnaires and feedback to teachers on teaching quality should be based on single items or on scales is controversially discussed. Single items, if formulated appropriately, provide teachers with more concrete indications for the improvement of teaching (Röhl & Rollett, 2020). For example, rather negative feedback on an item stating, "The exercise tasks challenge me to think" is easier to translate into an improvement action than a low scale average of the teaching quality aspect "cognitive activation." Scale averages, in contrast, have the advantage of conforming more to psychometric criteria and of presenting the results of very extensive questionnaires more clearly (Wisniewski et al., 2020). However, this is only important if items can be summarized into scales representing a meaningful construct. As some studies show, theoretically postulated dimensions of teaching quality can in many cases be separated statistically only to a limited extent (e.g., Kuhfeld, 2017; Wallace et al., 2016). Since scale-based questionnaires often have more items and are therefore more time-consuming to answer, the benefit of reporting scale averages instead of single item results seems to be disappearing for practical use in schools (Nelson et al., 2014; Röhl & Rollett, 2020).

2.3.2 Support for teachers in working with the received feedback

In the next steps of the student feedback process, it is necessary for the teacher to understand and interpret the information received, accept it, and use it to set goals for change. This means that teachers do not automatically improve their teaching quality sustainably when they receive feedback (Bijlsma et al., 2019). According to Kember et al. (2002), intrinsic motivation, positive encouragement from school leaders, and coaching on how to effectively utilize the teaching quality data are considered important factors in the effective use of feedback. In line with this, findings of L'Hommedieu et al. (1990) and Röhl (2021) point to the enhancing effect of support for teachers on the effectiveness of student feedback interventions.

2.3.3 Discussion of the received feedback with students

An important first improvement-oriented action can be to discuss the obtained feedback with the students in class, which enables teachers to understand the feedback in more detail and to ask directly about specific conspicuous aspects and how these results are to be interpreted from the class's point of view. For example, the information potential inherent in the feedback can be exploited by identifying specific classroom situations that are the source of negative student ratings and clarifying misunderstandings (Röhl et al., 2021). Students also appreciate a discussion about the feedback (Schmidt, 2018), which might lead to more thoughtful participation of students when giving feedback the next time. Based on teacher surveys, the findings

from Gaertner (2014) confirmed the significance of the scope and constructiveness of teachers' discussion of the received feedback with the class for subsequent improvement-oriented actions. For intervention studies that measure changes using student perceptions of teaching, it can be assumed that teachers' communication about planned improvement actions could lead to stronger effects because the students are more sensitized to possible changes. However, the results of a qualitative interview study with teachers also point to the extensive subsequent improvement-oriented actions of teachers without communication about the received feedback (Röhl & Rollett, 2021b).

2.3.4 Teacher characteristics

With regard to the study sample, teacher characteristics might affect the study outcomes as well. For example, the effect of student feedback on teaching quality might differ if the study was conducted on fully trained teachers compared to pre-service teachers because young teachers' expertise grows during the first years of professional experience, and therefore, pre-service teachers have more room for improvement (Stigler & Miller, 2018). It can also matter whether teachers decide to participate in the study voluntarily or whether they are obliged to participate by, for example, their school leader (due to extrinsic motivational factors; see Kember et al., 2002).

2.3.5 Intervention and study design characteristics

The outcome of intervention studies can also be influenced by the characteristics of the intervention. In the case of a student feedback intervention, the time between pre- and post-measurement seems to be particularly relevant: if this time span differs, the effect on teaching quality may be larger or smaller because the teachers have had respectively more or less time to work on improving the quality of their teaching. Additionally, the frequency with which a teacher receives feedback from students during the intervention period can have an influence. For example, some studies indicate a higher effectiveness of high-frequency feedback (Schiepek et al., 2015; Schmidt, 2018). Alternatively, the frequent giving of feedback could also lead to a fatigue effect on the students, such that attrition could appear.

Regarding the study design, the effect size of studies utilizing a control group (who did not receive any student feedback) might conceivably be a more accurate reflection of the impact of the intervention because this controls for retest effects.

3 Research questions

As this meta-analysis has included more publications than the previous of Röhl (2021) and uses a more sophisticated statistical approach, we first examine whether our overall analysis leads to the same results.

1. What is the average effect size of student feedback interventions in K-12 on teaching quality estimated by a hierarchical meta-regression model?

Furthermore, the modeling allows for a differentiated examination of the effects of student feedback interventions in terms of their effectiveness on different aspects of teaching quality:

2. Does this effect differ with regard to different aspects of teaching quality?

Following the outlined factors influencing the student feedback process, we furthermore address the following research questions regarding possible moderators:

3. Does the effect of student feedback interventions vary with respect to

   - Feedback questionnaire characteristics (item referent, answering scales, use of scales or single items),
   - The support provided to teachers in using feedback information to improve instruction,
   - Whether a discussion of the results with the class is part of the feedback use process,
   - The characteristics of the teacher in the sample (fully trained or pre-service, study participation voluntary or obligatory), and,
   - The characteristics of the intervention (duration, frequency of obtaining feedback) and study (with or without control group, published in a peer-reviewed journal, publication year)?

4 Methods

4.1 Literature search

4.1.1 Search and criteria for the inclusion and exclusion of studies

We used a procedure described by Littell et al. (2008) to efficiently plan a systematic literature search. The steps of this procedure include searching in bibliographic and scientific databases by using terms and strings, searching for sources of unpublished articles and dissertations, and asking personal contacts. We include not only studies published in peer-review journals but also gray literature including relevant dissertation and master theses to avoid publication bias, which can appear if unpublished studies, that often have smaller effect sizes, are excluded from meta-analyses (Lipsey & Wilson, 2001). We searched for studies published in English, German, and Dutch for three reasons: (1) we can read studies published in these languages; (2) there is an abundance of research on student feedback, particularly in the Netherlands and Germany; and (3) to reduce a regional bias that could occur if we limit our meta-analysis to studies published in English, as these studies predominantly come from English-speaking countries. In October 2020, we searched three international databases (ERIC, PsycInfo, Open dissertation), as well as one German database (FIS Bildung) and one Dutch database (Narcis) using several search terms. A detailed description of the search process can be found in

Supplement S1. Additionally, we contacted fellow researchers from three different countries, who all do research in the field of teacher evaluation or student feedback. Further sources of studies included the reference lists in relevant studies as well as the forward citation history of relevant articles in Google Scholar. To keep track of the search results, the PRISMA reporting guidelines were followed (Liberati et al., 2009).

We included studies published in English, German, or Dutch in all types of sources—including peer-reviewed and non-peer-reviewed journal articles as well as dissertations, reports, and unpublished articles. We did not apply any exclusion criteria based on publication year because we wanted to include a broad variety of empirical evidence (Littell et al., 2008).

We specified the characteristics of population, intervention, context, and outcome, as follows.

Population The participants were fully trained teachers or teachers in training in primary and secondary education. No exclusion was made based on the subject matter taught by the teacher or on the context of country, school, or classes.

Intervention The studies investigated the effect of student feedback on teaching quality. Studies that measured teaching quality by means of student perceptions but did not include an intervention (e.g., giving at least one piece of feedback to teachers before a second measure) were excluded. Studies in which teaching quality was measured but the intervention itself did not focus on teaching development were also excluded.

Context The treatment and control groups were comparable with respect to pretest measures or initial teaching quality. Studies were only excluded when significant differences in initial teaching quality were reported.

Reporting of outcomes of interest All studies included focused on teaching quality aspects, which are given in Table 1. The quantitative data necessary for calculating the effect size are reported. If such data were not reported, we contacted the authors to ask for the data. If we did not obtain sufficient data for calculating effect sizes, those studies were excluded.

The studies found were discussed among the authors of the study to ensure they met our inclusion criteria.

4.2 Study coding

The coding procedure was separated into two phases. In the first phase, we coded general information about the publication, intervention, and sample. To determine the objectivity and reliability of this coding, the three authors of this article first coded two studies together (these studies were extreme in regard to publication year and research design). Disagreement between coders was resolved by discussion. After coding and discussing these two studies, the authors divided the remaining

Table 1 Coding scheme of teaching quality aspects

Aspect	Description	Example item
Socio-emotional support [a,b,d]	Supporting students' social relatedness to other students, the relationship with the teacher, and the teacher-student interaction	The teacher created a good classroom climate
Classroom management [a,b]	Maximize productive instruction time by preventing and correcting the loss of time, attention, and undesired behavior by means of rules, routines, correction, and encouragement	The teacher made sure we worked well during the lesson
Motivating the students [a,d]	Keeping students motivated and being interested in the students, e.g., by explaining the relevance of the learning content and why they are going to learn it	What I learn in this class is relevant for my day-to-day life/the teacher keeps me interested
Content-related learning		
Clear instruction [a,b,d]	Explaining the subject matter in such a way that students understand it well	My teacher explains things clearly
Cognitive activation [a]	Making sure students think deeply about the learning content	The teacher asked questions that make me think deeply about the learning content
Practicing [a]	Giving useful exercises and enough time to practice the learning content on their own	The teacher gave us enough time to work on the assignments in the lesson
Formative assessment and feedback [a]	Formatively assessing the students' understanding about the content and giving helpful feedback	If I answered a question incorrectly, the teacher explained why my answer was wrong
Differentiation and adaption [a,b]	Adapting the learning contents, goals, and/or assignments to the performance level of different students in class	I'm able to work at the speed that suits my ability
Self-regulated learning [b,d]	Stimulating students to use effective strategies to achieve their learning goals, by, e.g., making them check their own learning progress	My teacher helps me make plans for how I'll do my work
Summative assessment [c]	Explaining the assessment criteria and/or grading system of the test(s)	The teacher explained the test criteria clearly to me

Table 1 (continued)

Aspect	Description	Example item
Teacher's habits and characteristics [c]		
Teacher's professional habits [c]	Teacher's professional appearance	The teacher has good grooming and hygiene habits
Teacher's subject-related characteristics [c]	Teacher's knowledge and understanding of his/her teaching field and knowledge about the material and course content	The teacher knows the material and course content very well
Overall [c]	Students' overall rating or indication of how informative/instructive the lesson has been	What is your overall evaluation of your teacher's effectiveness?
Other [c]	Other and only very rarely mentioned aspects	

Origins of teaching quality categories: [a]Praetorius and Charalambous, (2018), [b]van de Grift et al. (2014), [c]inductively developed category. [d]Adapted or unified designation

studies among themselves so that each study was coded by one person. The coders met regularly to discuss ambiguous decisions. During this coding, the following information was extracted from the studies.

- *Publication type*: peer-reviewed journal articles, book chapters, dissertations, research reports, and conference proceedings
- *Study design*: country of the study, experimental vs. quasi-experimental design, pre-post vs. control-experimental group, and theoretical background (how were the studies theoretically motivated?)
- *Teacher sample*: number of teachers, subject taught by the teachers, method for obtaining/recruiting the sample, and fully trained or pre-service
- *Student sample*: number of students, mean age, and grade
- *Feedback intervention*: number of items of obtained feedback, duration between the first and last items of feedback, and kind of information given to the teachers (e.g., scale means and distribution of student response on single item)
- *Provided support*: the way the teachers were informed or trained to deal with the feedback, support for utilizing the feedback, discussion with students, and selection of an area of improvement
- *Feedback questionnaire*: digital/paper-based, answering scale (e.g., frequency or agreement), and referent used in the item wording

Please see the Suplement S1 for detailed coding information.

As described above, a synthesis of the two frameworks of Praetorius and Charalambous, (2018) and van de Grift et al. (2014) was developed for the coding scheme for the *aspects of teaching quality*. This newly developed coding scheme was tested on a subset of the studies. In this context, the identification of teaching characteristics that were not covered resulted in the need for the addition of new, inductively derived categories as well as a more explicit clear separation of certain categories. The coding scheme finally used with the origin of the categories and sample items can be found in Table 1. The final coding scheme was used for recoding the teaching quality aspects of every effect size by two independent raters. The resulting substantial interrater agreement of weighted $\kappa = 0.70$ indicated the appropriateness of the modified coding scheme (Landis & Koch, 1977). Any discrepant coding was resolved by discussion afterward.

4.3 Calculation of effect sizes and study variance

Effect sizes are calculated as Cohen's *d*. Most of the studies provided means and standard deviations for pre- and post-measurements. For studies including control groups, the differences of the pre-post changes of intervention and control group were calculated using a group-size-adjusted standard deviation of the pretest, as suggested by Carlson and Schmidt (1999), using the following formula:

$$d = \frac{\left(M_{2\mathrm{IG}} - M_{1\mathrm{IG}}\right) - \left(M_{2\mathrm{CG}} - M_{1\mathrm{CG}}\right)}{\mathrm{SD}_{\mathrm{prePooled}}} \quad (1)$$

with $SD_{prePooled}$ being the group-size-adjusted standard deviation of the pretest. Since this meta-analysis also considers longitudinal studies without a control group, we decided to calculate pre-post effect sizes for these studies using the standard deviations of the premeasurement accordingly. In addition, for control purposes, we calculated the pre-post effect sizes of the control and intervention groups of the other studies separately. This could help in identifying a possible bias relating to study design. For four of the older studies, the articles did not report the standard deviations, and it was not possible to obtain this information from the authors. In this case, we calculated effect sizes using the means and the included *t* or *F* statistics (Lipsey & Wilson, 2001). Effect size variances were estimated following Lipsey and Wilson (2001, pp. 48–49) using Hedge's formula,

$$SE^2 = \frac{n_{G1} + n_{G2}}{n_{G1} n_{G2}} + \frac{ES^2}{2(n_{G1} + n_{G2})} \tag{2}$$

4.4 Investigating publication bias

To investigate potential publication bias, we used a funnel plot (Borenstein et al., 2009) in combination with an Egger test to get an objective measure for potential publication error. A significant regression coefficient in the Eggers test can be evidence of publication bias (Egger et al., 1997). However, the standard error and effect size estimations correlate with each other as the effect size is part of the second term of the standard error (see Formula 2). To avoid an increased error type 1 (false positive results) of the Eggert test because of this correlation, we followed the recommendation of Pustejovsky and Rodgers (2019) to modify the measure of precision by using only the first term of Formula 2 $\left(SE = \sqrt{\frac{n_{G1}+n_{G2}}{n_{G1} n_{G2}}}\right)$ as a predictor in the Eggert test. We deal with the dependency between effect sizes in the Eggert test by using the multi-level meta-regression models instead of simple meta-regressions (Fernández-Castilla et al., 2021; Rodgers & Pustejovsky, 2021).

4.5 Data analysis with a multi-level meta-regression

To analyze the data, we utilized a multi-level meta-regression analysis, which is similar to a regular multi-level regression model but with the difference that the effect sizes (dependent variable) are weighted by the inverse study or intervention group variance to consider that they are based on studies with different sample sizes (Borenstein et al., 2009). As in regular multi-level analysis, the underlying assumption of a multi-level meta regression is that the effect sizes included in the analysis are a random sample from a population of effect size parameters originating from all relevant studies that have been conducted or will be conducted in the future (Assink & Wibbelink, 2016; Konstantopoulos, 2011). Using such multi-level models to analyze meta-analytical data is necessary in many cases because effect sizes come from different studies and thus differ systematically from each other and not just because of random sampling error. In many cases, the data structure is even more

Table 2 Fit comparison of different hierarchical model specifications

Parameter	Model A: 4-level		Model B: 3-level w. intervention group level		Model C: 3-level with study level		Model D: 2-level	
	Estim	*p*	Estim	*p*	Estim	*p*	Estim	*p*
d	0.269	0.006	0.270	0.004	0.254	0.006	0.131	< 0.001
σ^2 between studies	0.061	0.333	-	-	0.126	< 0.001	-	-
σ^2 between intervention groups	0.095	< 0.001	0.159	< 0.001	-	-	-	-
σ^2 within	0.007	< 0.001	0.007	< 0.001	0.009	< 0.001	0.041	< 0.001
AIC	225.55		224.49		251.03		403.71	
BIC	240.54		235.73		262.27		411.20	

complicated, as there is not only systematic variance between studies but also within studies if multiple effect sizes are extracted from single studies. The data of this meta-analysis has such a nested structure because we calculated several effect sizes with regard to different aspects of teaching quality within every intervention group. Additionally, multi-level models enable to estimate portions of variance explained by moderators on all analysis levels (Hox et al., 2018).

Some studies included more than one intervention group with different treatment conditions but with the same control group. These different intervention groups mostly differ in the extent of support provided to teachers but use the same feedback questionnaire and mostly share the same characteristics (school type, level, country, education). Thus, the within study variance of our data might further be divided into two levels, as in the same study, there are dependencies both within the same intervention group and between different intervention groups. Therefore, the data structure of our meta-analysis breaks down into four levels: (1) between study variance, (2) variance between intervention groups within the same study, (3) variance within intervention groups between different effect sizes, and (4) random sampling error. However, only a small number of studies included multiple intervention groups, so it is uncertain whether the fourth level is relevant for this meta-analysis. To identify which levels are relevant for our dataset, we used a likelihood-ratio testing (LRT, see Table 2) to compare (A) a four-level model, (B) a three-level model with intervention groups as the upper level, (C) a three-level model with studies as the upper level, and (D) a two-level model only considering random sampling error and between study variance (Assink & Wibbelink, 2016). If these model comparisons show that some levels are non-relevant, this means that there is not enough variability between effect sizes on that level. This happens if the number of effect sizes on a level is small compared to the number of effect sizes on other levels (Konstantopoulos, 2011). The identified model configuration with all relevant levels will be utilized in all subsequent analysis.

To answer the first research question regarding the mean effect sizes overall studies, we calculated a regression model including only an intercept and no moderator variable. In order to address the remaining research questions regarding moderator

effects of study and treatment characteristics, we added dummy coded moderator variables as predictors to the meta-regression model. We calculated mean weighted effect sizes and confidence intervals for every category based on the intercept and regression coefficients (Borenstein et al., 2009). The dependencies between effect sizes were also considered when estimating confidence intervals and hypothesis testing by utilizing cluster robust variance estimation with bias-reduced linearization and small sample adjustment as proposed by Pustejovsky and Tipton (2018). This approach showed a high level of robustness and accuracy in simulations for hierarchical meta-analyses even with a low number of clusters (> 15; Pustejovsky & Tipton, 2018).

Regarding continuous study characteristics, such as the number of teachers in the sample, the studies were split at the median, as these study characteristics were very unevenly distributed and, in most cases, linear effects could not be assumed. For example, it cannot be expected that the effort required to implement the intervention increases linearly with the number of participants, or that studies improve continuously over time. Since not all studies reported information for all moderators tested, the resulting missing values were listwise excluded from the corresponding moderator analyses. Additionally, meta-regressions including several predictors at once were used to analyze possible confounding between the effects of different moderators. Estimation of the overall and moderator effect sizes and confidence intervals was done using REML estimator within the r packages *metafor* (Viechtbauer, 2010) and *clubSandwich* (Pustejovsky, 2021).

5 Results

5.1 Results of the literature search and overview of included studies

A total of 1484 publications were identified through the database search. Two more studies were found in references lists, and nine more studies were found by other searches (e.g., asking fellow researchers). This results in a total of 1495 potentially relevant studies. After removing duplicates and screening abstracts and titles, 37 articles that presumably met our inclusion criteria remained. Based on a detailed reading of the full text of these papers, we omitted 14 further studies because they did not meet our inclusion criteria; thus, 23 studies remained in our analysis (see Fig. 2).

An overview of the characteristics of the 23 studies identified is presented in the supplemental Table S2. The majority of the studies were conducted in the USA (9), Australia (5), and Germany (4). Two studies took place in the Netherlands and one each in Austria, Great Britain, and Turkey.

Of the studies, 19 referred to fully trained in-service teachers and another four to pre-service teachers in training. The interventions were conducted in grades 5 to 13. While seven interventions were limited to exactly one grade level, the other studies involved classes from different levels. Some studies focused on only a single subject matter for a better comparability of the classes (e.g., Novak, 1972a, b on

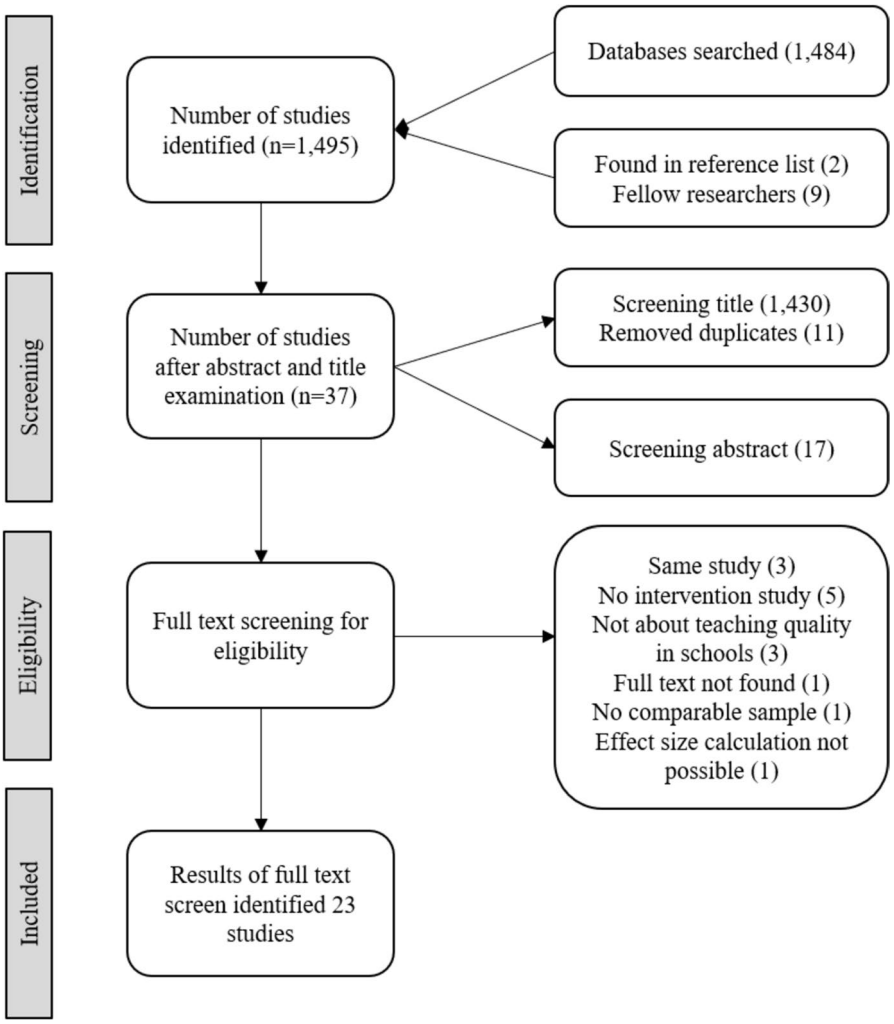

Fig. 2 Flowchart of the literature search according to PRISMA guidelines

biology, Rösch, 2017 on physics, and Bijlsma et al., 2019 on mathematics), while others included multiple subjects.

The duration of the intervention in the studies differs to a large extent, from 1 month to 1 year. During these periods, the number of feedback reports provided varied from mostly one up to 10. In one case (Bijlsma et al., 2019), teachers were allowed to obtain feedback as often as they wished during the intervention period of four months; this was used up to 16 times, with an average of 6.7 times. The feedback questionnaires used also cover the various teaching quality aspects in very different ways; for example, one study only focused on the aspect of collaboration (Schmidt, 2018), while other studies covered a wide range of aspects (e.g.,

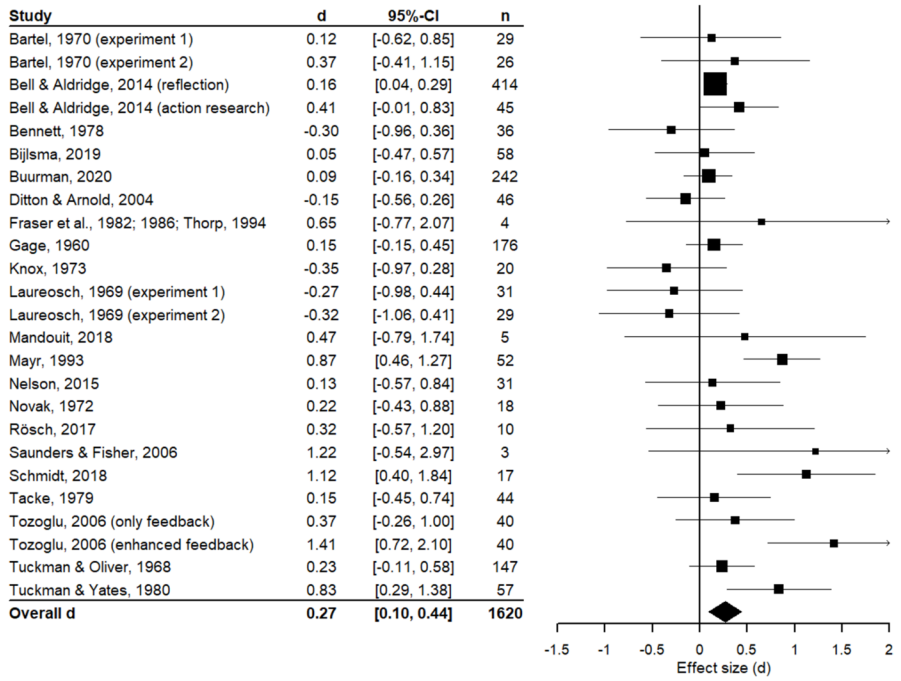

Fig. 3 Forest plot of all included studies. Note. A study is a dataset consisting of one sample. Some papers report data from more than one sample and thus appear multiple times in the plot. n = number of classes

Mandouit, 2018; Tuckman and Olivier, 1968). The forest plot in Fig. 3 shows a large variation between the mean effect sizes of the included studies.

5.2 Model comparison

The number of effect sizes within intervention groups varied from only one (Schmidt, 2018) up to 29 (Mandouit, 2018). The model characteristics and variance proportions at the different levels are shown in Table 2.

The model comparisons indicate the lowest fit for the two-level model D. This means that the variability between effect sizes cannot be attributed exclusively to random sampling error and between-study variance. The three-level model C with nesting in studies as the top level has a significantly better fit ($LRT(1) = 154.68$, $p < 0.001$) compared to model D. This is in turn inferior to both the four-level model A ($LRT(1) = 27.48$, $p < 0.001$) and the intervention group three-level model B, whereby no significant difference in model fit ($LRT(1) = 0.94$, $p = 0.333$) emerged between model A and B. However, the four-level model A does not show a significant amount of variance at the highest level ($\sigma^2 = 0.061$, $p = 0.333$), which is reasonable because only four studies do include multiple intervention groups. Thus, we choose the three-level model B with intervention groups as the highest level of clustering for further analyses. For the studies including only one intervention group,

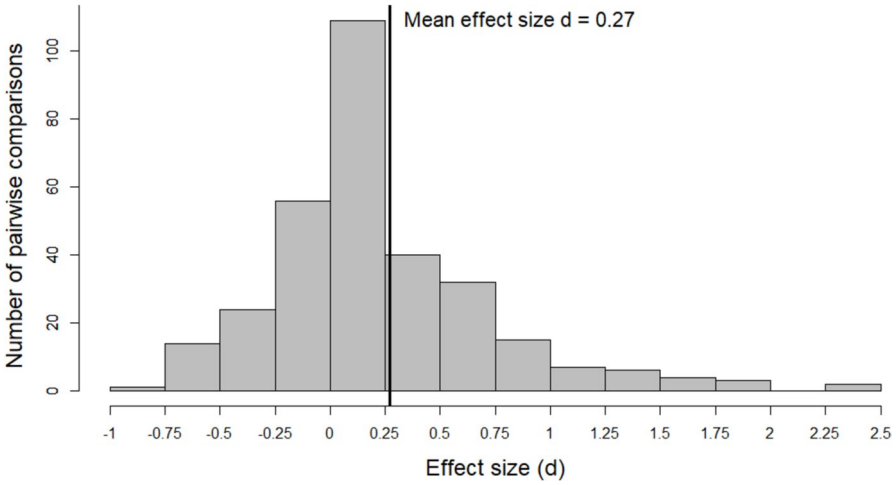

Fig. 4 Histogram of the effect sizes of all pairwise comparisons ($k = 313$ effect sizes)

the intervention group level equals the study level. The variances of this model at the between-intervention group level ($\sigma^2 = 0.159$, $p < 0.001$) and the within-intervention group level ($\sigma^2 = 0.009$, $p < 0.001$) were significant. The model showed high heterogeneity between the included studies ($Q(313) = 540.3$, $p < 0.001$) and portions of variance of 74.93% between intervention groups, 21.63% within intervention groups, and 3.44% for sampling error. These estimates support both the specification of random effect models and the assumption of the existence of possible moderators to explain the variance between studies and individual effect sizes (Hunter & Schmidt, 2007).

5.3 Meta-analyses of overall intervention effects

Overall, 313 effect sizes were extracted from the 23 included studies. These effect sizes vary from − 0.90 to 2.44, while most effect sizes are in the range between 0 and 0.25 (see Fig. 4). The mean weighted effect size of all studies considering studies with a pre-post and treatment–control-group design is $d = 0.27$ ($p = 0.004$, 95%-C.I. 0.098 to 0.441). If the mean effect size is estimated only as the change between pre- and post-measurement without considering the control group effects, the result is a slightly higher effect size of $d = 0.29$ ($p = 0.002$, 95%-C.I. 0.116 to 0.463). The weighted mean pre-post change in the control groups is $d = 0.021$ ($p = 0.621$, 95%-C.I. − 0.073 to 0.114) and not significantly different from zero.

5.4 Meta-regression and analyses of moderators

For the subsequent meta-regressions, the effect sizes for studies with a control group were controlled by the corresponding control group effects. Since only minimal changes were observed in the control groups, the simple pre-post change was used for studies without a control group.

Table 3 Results of moderator analyses I: teaching quality aspects and questionnaire characteristics

Moderator	*n*	*k*	*d*	(SE)	95%-C.I		Signif. of moderator
Teaching quality aspects							
Socio-emotional support	19	90	0.234	(0.089)	[0.050; 0.418]	*	0.214
Classroom management	12	34	0.295	(0.123)	[0.037; 0.553]	*	0.804
Motivation	16	27	0.303	(0.085)	[0.126; 0.481]	**	0.265
Content learning (summarized)	20	103	0.307	(0.094)	[0.112; 0.501]	**	0.221
Clear instruction	15	35	0.306	(0.109)	[0.080; 0.533]	*	0.363
Cognitive activation	12	24	0.416	(0.109)	[0.184; 0.648]	**	0.121
Practicing	5	7	0.197	(0.097)	[−0.019; 0.413]	†	0.142
Formative assessment and feedback	7	12	0.307	(0.088)	[0.107; 0.508]	**	0.395
Adaption and differentiation	13	21	0.288	(0.090)	[0.100; 0.476]	**	0.662
Self-regulated learning	3	4	0.223	(0.091)	[−0.061; 0.507]	†	0.300
Summative assessment	2	2	0.173	(0.217)	[−1.700; 2.050]		0.695
Teacher characteristics	7	19	0.265	(0.107)	[0.030; 0.500]	*	0.887
Overall	8	10	0.122	(0.080)	[−0.052; 0.296]		**0.001**
Questionnaire and feedback report characteristics							
Item referent							
Teacher	21	209	0.253	(0.087)	[0.073; 0.433]	**	0.320
Students	6	12	0.584	(0.086)	[0.368; 0.800]	***	**0.007**
Class	9	33	0.289	(0.110)	[0.056; 0.522]	*	0.722
I/we/you	6	17	0.194	(0.088)	[0.001; 0.387]	*	**0.046**
Type of answering scale							
Agreement	8	114	0.197	(0.093)	[0.023; 0.418]	†	0.478
Bipolar	2	30	0.224	(0.588)	[−0.927; 1.378]		0.941
Comparison	2	34	−0.118	(0.173)	[−0.862; 0.627]		0.121
Frequency	7	92	0.518	(0.138)	[0.185; 0.852]	**	**0.035**
Grading	4	40	0.172	(0.135)	[−0.172; 0.516]		0.479
Scale or item use							
Scale	9	90	0.494	(0.163)	[0.123; 0.866]	*	0.068
Item use	14	224	0.138	(0.083)	[−0.038; 0.315]		

n, number of studies; *k*, number of effect sizes; *d*, effect size as Cohen's *d*; *95%-CI*, 95% confidence interval of effect size; †$p < 0.10$, *$p < 0.05$, **$p < 0.01$, ***$p < 0.001$

5.4.1 Effects on different aspects of teaching quality

Overall, the mean effect sizes for almost all analyzed sub-aspects of teaching were significantly different from zero (Table 3). The mean average effect sizes of *practicing*, *self-regulated learning*, and *summative assessment* missed the two-tailed level of significance although only a few effect sizes concerning these aspects could be identified. Therefore, it can be assumed that the use of student feedback has a positive effect independent from the aspect of teaching considered. A notable exception appears for the effect of items and scales that

measure the *overall* impression or assessment of teaching. Here, the analyses indicated that no stable effect different from zero can be detected for this area, and at the same time this aspect showed a significantly smaller effect ($d = 0.122$, $t(5.71) = 5.99$, $p = 0.001$).

5.4.2 Characteristics of feedback questionnaires

For the choice of referent in item wording, the analyses indicated significant moderator effects (Table 3). Thus, using "students" as item referent resulted in a significant increase in effect size up to $d = 0.584$ ($t(1.47) = 25.1$, $p = 0.007$), whereas the use of "I," "we," or "you" decreased the effect down to $d = 0.194$ ($t(1.63) = 5.71$, $p = 0.046$). However, these results should be considered critically as only a few studies with these item characteristics could be identified, and therefore the degrees of freedom for the *t*-tests are very low. In addition, studies with frequency answering scales showed a significantly higher intervention effect compared to the other scale types ($d = 0.518$, $t(7.59) = 2.57$, $p = 0.035$). Furthermore, studies using scale-constructs in questionnaires ($d = 0.494$) tended to show a slightly higher effect than studies with feedback instruments based on single items ($d = 0.138$, $t(17.81) = 1.95$, $p = 0.068$).

5.4.3 Support of teachers and discussion of feedback with the class

Highly significant effects were found with respect to the characteristics and degrees of support for teachers as well as for discussing the feedback received with the class. Thus, there was a significantly large difference in the effectiveness of the intervention when individual teachers were offered suggestions for possible areas of improvement as well as guidance for improving their teaching ($d = 0.568$) compared to teachers without support or only with hints as to statistical interpretation ($d = 0.050$, $t(20.9) = 3.20$, $p = 0.004$). In addition, we found a significant moderator effect for individual and groupwise counseling sessions ($d = 0.452$) compared with no support ($d = 0.082$, $t(21.5) = 2.08$, $p = 0.049$). Additional analysis using multiple meta-regression with both support characteristics revealed that these two variables were highly confounded and that the type of support was more relevant for the intervention effect than the way of providing support.

There was a significant increase in the effect when the instruction for the participating teachers included encouragement or even an obligation to discuss the received feedback with the class ($d = 0.556$) compared with studies without this kind of instruction ($d = 0.169$, $t(10.49) = 2.28$, $p = 0.045$). However, six from the eight studies that instructed their participants on this use of the feedback also included extensive support; hence, this effect was confounded with the support type, with a slightly stronger effect for the latter ($B = 0.463$, $p = 0.030$ vs. $B = 0.203$, $p = 0.261$), while only extensive support remained significant in multiple meta-regression analyses.

5.4.4 Teacher characteristics

Concerning the characteristics of the participating teachers, the estimated effect sizes of the subgroups pointed to the expected direction of a higher effect size for pre-service teachers and voluntarily recruited teachers (Table 4), but they were not significant. An analysis of the number of teachers in the intervention groups as a continuous moderator also showed no significant effect ($B = -0.001$, $p = 0.568$).

5.4.5 Intervention and study design characteristics

With regard to intervention characteristics, the analyses indicated that the duration and the number of feedback cycles conducted have no statistically relevant impact on the intervention effect (Table 4). Additionally, no significant moderator effects were found with regard to the analyzed study characteristics design and publication type (Table 4). Regarding the year of publication, studies published after 1990 showed a higher effect size ($d = 0.436$) than older studies ($d = 0.093$, $t(23.4) = 2.16$, $p = 0.041$).

5.5 Multivariate meta-regression

To address a possible confounding of the moderator effects found, we conducted a meta-regression including all characteristics that were significant when considered individually (Table 5). In Model 1, all characteristics to be considered statistically robust were integrated as independent variables. Model 2 additionally includes the item referent, as this effect has the potential to be weak due to the low degrees of freedom. Model 3 controls for the year of publication and the use of single items or scales by adding these variables.

The results of Model 1 indicate that only the moderator effects of type of support for teachers ($B = 0.421$, $p = 0.032$) and scales or items addressing overall teaching quality ($B = -0.109$, $p = 0.048$) stay significant. Including the characteristics of the item referent in Model 3 showed a significant effect of "student" as referent ($B = 0.298$, $p = 0.028$). This model explains 33.3% of the variance between intervention group level and 30.1% within. The inclusion of other study characteristics (Model 3) showed no further significant effects although the p-value of the moderator effects of teacher support and teaching quality aspects increased slightly ($p = 0.091$ and $p = 0.069$). However, the proportion of variance explained at between intervention level decreased compared to the previous model, so this effect should be considered critical. In summary, these additional analyses indicate that the moderator effects of teacher support and the use of items that survey the "overall" quality of teaching are relatively stable and reliable.

Table 4 Results of moderator analyses II: study and intervention characteristics

Moderator	*n*	*k*	*d*	(SE)	95%-C.I	Signif. of moderator
Support characteristics and discussion with students						
Type of support for teachers						
No or statistical interpretation only	14	179	0.050	(0.061)	[−0.081; 0.182]	**0.004**
Improvement areas & developmental guidance	13	135	0.568	(0.150)	[0.235; 0.901]**	
Way of support for teachers						
No support/ Not mentioned	14	154	0.082	(0.057)	[−0.042; 0.206]	**0.049**
Support meetings (individual or collective)	13	155	0.452	(0.169)	[0.079; 0.825]*	
Discussion of received feedback with students						
Not mentioned	16	214	0.169	(0.980)	[−0.036; 0.375]	**0.045**
Encouraged or binding	8	100	0.556	(0.139)	[0.217; 0.895]**	
Teacher characteristics						
Sample						
Fully trained teachers	19	239	0.255	(0.083)	[0.081; 0.429]**	0.774
Pre-service teachers	4	75	0.350	(0.303)	[−0.507; 1.206]	
Recruitment of sample						
Obligatory	6	82	0.125	(0.195)	[−0.356; 0.606]	0.332
Voluntary	15	208	0.348	(0.101)	[0.134; 0.562]**	
Intervention characteristics						
Duration						
Duration < 2 months	10	89	0.382	(0.100)	[0.157; 0.607]**	0.276
Duration ≥ 2 months	13	225	0.208	(0.119)	[−0.045; 0.461]	
Number of feedback reports						
Feedback reports = 1	17	248	0.277	(0.100)	[0.067; 0.487]*	0.911
Feedback reports > 1	6	66	0.255	(0.156)	[−0.149; 0.660]	
Study characteristics						
Design						
Pre-post only	12	158	0.358	(0.124)	[0.084; 0.632]*	0.394
(Quasi-)experimental	11	156	0.209	(0.122)	[−0.043; 0.462]†	
Publication type						
Peer-reviewed journal	9	111	0.325	(0.089)	[0.120; 0.530]**	0.589
Other	14	203	0.242	(0.122)	[−0.017; 0.501]†	
Publication year						
Published before 1990	11	147	0.093	(0.992)	[−0.125; 0.310]	**0.041**
Published after 1990	12	167	0.436	(0.159)	[0.165; 0.706]**	

Note. *n*, number of studies; *k*, number of effect sizes; *d*, effect size as Cohen's *d*; *95%-CI*, 95% confidence interval of effect size; †$p < 0.10$, *$p < 0.05$, **$p < 0.01$, ***$p < 0.001$

Table 5 Results of meta-regression ncluding all moderators showing significant effects in single moderator analyses

	Model 1				Model 2				Model 3			
	Est	(SE)	*p*		Est	(SE)	*p*		Est	(SE)	*p*	
Intercept	−0.008	(0.072)	0.919		−0.025	(0.075)	0.744		−0.080	(0.301)	0.797	
Discussion with students encouraged or binding	0.150	(0.182)	0.435		0.212	(0.183)	0.286		0.119	(0.321)	0.721	
Support with improvement areas and developmental guidance	0.421	(0.176)	0.032	*	0.435	(0.180)	0.030	*	0.450	(0.238)	0.091	†
Answering scale: frequency	0.233	(0.149)	0.168		0.270	(0.160)	0.151		0.212	(0.189)	0.311	
Teaching quality aspect "overall"	−0.109	(0.043)	0.048	*	−0.097	(0.048)	0.092	†	−0.108	(0.046)	0.069	†
Item referent "students"					0.298	(0.035)	0.028	*	0.299	(0.032)	0.025	*
Item referent "I/we/you"					−0.035	(0.027)	0.345		−0.034	(0.025)	0.327	
Published after 1990									0.161	(0.255)	0.550	
Use of scales									0.002	(0.279)	0.993	
σ^2 between	0.107				0.106				0.116			
σ^2 within	0.008				0.005				0.005			

Note. †$p < 0.10$, *$p < 0.05$, **$p < 0.01$, ***$p < 0.001$

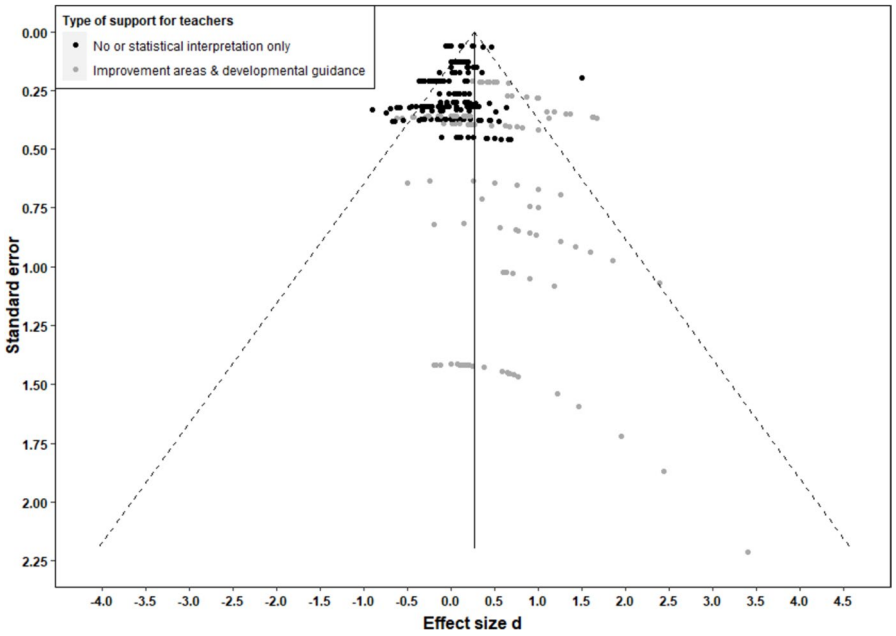

Fig. 5 Funnel plot showing the mean weighted effect size in relation to the corresponding standard error. Note. The dotted line indicates the 95% confidence interval

5.6 Publication bias

An inspection of the funnel plot in Fig. 5 shows an asymmetric distribution of the effect sizes, with more studies in the lower right than in the lower left part of the funnel. However, the results of the Eggert test were non-significant ($b = 0.30$; $p = 0.11$), which speaks against a potential publication bias. A possible explanation for the slightly asymmetric funnel plot could be a correlation between effect sizes and sample size based on the fact that studies offering support to the teachers, which lead to a higher effect size (Table 2), have significantly fewer participants ($M = 19.92$) than studies without support ($M = 94.78$; $t(312) = 8.16$, $p < 0.001$). Moreover, Fig. 5 shows that studies with support are overrepresented in the lower-right area of the funnel plot because they have higher effect sizes and lower sample sizes compared to studies with no support.

6 Discussion

6.1 Discussion of the findings

This meta-analysis summarizes the available evidence on the effectiveness of student feedback interventions in schools and examines indications of possible success conditions for their beneficial use in practice. In the multilevel meta-analysis

of the effects of student feedback interventions on student-perceived teaching quality in schools, 23 studies with 314 effect sizes were included. Due to the different study designs (some with multiple intervention groups within a study), different model specifications with two to four levels were compared, and a three-level model with clustering by intervention groups was found to fit best, resulting in a weighted mean effect size of $d = 0.27$. Although this effect seems to be relatively small, it is significant and lies in a range similar to those of meta-analyses of student feedback interventions in higher education (Cohen, 1980; L'Hommedieu et al. 1990).

Concerning the effectiveness of student feedback for different aspects of teaching quality, the findings indicated stable and important improvements for all analyzed aspects except for those for which only very few effect sizes have been identified. Only items and scales related to the overall assessment of teaching (e.g., "What is your overall evaluation of your teacher's effectiveness?") showed significantly lower effect sizes. This could be due to teachers generally being asked to focus their professional development on only a few specific aspects of teaching that were particularly in need of improvement, as teachers deliberately practice specific skills to, at the end, become better overall teachers. Moreover, these items and scales related to the overall assessment of teaching do not contain any information on *what* to improve specifically, and arguably, therefore, no improvement of the overall quality of teaching can be expected.

Moreover, we cannot confirm findings from the literature (e.g., by van der Lans et al., 2019) that different aspects of instructional quality require different amounts of effort for improvement. However, it should be noted that only those studies with a very small sample (Fraser & Fisher, 1986; Fraser et al., 1982; Thorp et al., 1994) provided information on which areas the teachers focused on in the development process. Although these studies showed significant improvements explicitly for these aspects, no further conclusions can be drawn about the developmental areas of the other studies.

As presented in the SFT model (Röhl et al., 2021), the student feedback process starts with obtaining the feedback using an appropriate feedback questionnaire. Our findings confirm the relevance of the questionnaire design, as item and scale characteristics showed significant effects on the effectiveness of student feedback. For example, the analyses indicate that stronger improvements in teaching may result in item formulations with the "class" as the referent, as opposed to "I" or other wordings. Whether this use of item referent tends to result in more accurate measures of instructional quality—or, alternatively, in greater teacher acceptance of feedback as it could lead to a more neutral attribution of causes for critical feedback than using the referent "teacher"—cannot be conclusively answered based on the available information. Similarly, the causality behind the tendency toward a higher effect when a frequency scale is used cannot be conclusively assessed. Overall, there seems to be more favorable effects for survey instruments that ask for frequency assessments of the teaching aspects rather than evaluations and individual judgments. However, both effects should be considered with caution as the number of studies included with these characteristics is small. The higher effect of scale compared to item use in feedback questionnaire seems to be confounded with the publication year of studies,

as older studies used scales less often in the feedback questionnaires (see below for further considerations on the differences between older and younger studies).

Following the SFT model, the next step in the student feedback process requires an intense perception and understanding of the feedback received, acceptance of the information contained, and goal setting for improvement-oriented actions to have an instructional improvement effect. In this regard, moderator analysis showed an increase in effect size to $d = 0.55$ for additional support for the participating teachers, especially if the teachers received support while identifying and developing possible improvement areas. This support was mostly provided by the researchers themselves or by coaches and supervisors commissioned for this purpose. This finding is also in line with findings for college and university teachers (Penny & Coe, 2004) and also with the less comprehensive meta-analysis of Röhl (2021). Regarding the way teachers are given support, we found that both individual and collective support seem to be particularly effective for inducing changes in teaching through student feedback. However, because the support type and way it is given are confounded in primary studies, the separate effect of type and way could not be estimated.

Another significant enhancement of the effect emerges for intervention groups in which teachers were encouraged to discuss the obtained feedback with the students who gave the feedback. This finding is in line with positive correlations between feedback use and discussion with the class in studies that were not included here due to a lack of pre-post measures of teaching quality (Gaertner, 2014; Kane & Maw, 2005). However, this finding seems to be confounded with support type because these interventions also integrated a high level of support for teachers. Talking with students about the feedback received might also show a commitment to clarifying understanding and signal to the students that the teacher has the intention to listen and improve based on their feedback. Moreover, discussing the feedback with students in class about improvement-oriented action promotes *student voice*: the voice of students in their education and in the instructional development process (Barker, 2018; Beattie, 2012).

Regarding teacher characteristics, we found no evidence that the effect of feedback differs between in-service teachers and teachers in practical training or between teachers who participated voluntarily and those who were obliged to participate in the feedback intervention. However, these findings must be interpreted with caution as the comparisons are based on a small number of studies.

Concerning study characteristics, we found that earlier studies (before 1990) tend to show lower effects than more recent studies. However, as mentioned above, this effect is confounded with the use of single items in earlier studies and scales in more recent studies. The lower effect of the older studies could be related to the fact that some of the studies still contain less developed conceptions of teaching quality, such as items on the teacher's "odor" and appearance (Knox, 1973) or on whether the teacher shares "amusing happenings and experiences" (Lauroesch et al., 1969). Furthermore, negative feedback on such items could easily be interpreted by teachers as an attack on the person and thus significantly reduce the willingness to change (Hattie & Timperley, 2007). Apart from this, however, there are also older studies with methodologically sophisticated designs (such as the additional use of tape recordings of lessons to measure change in Novak, 1972a, b, so that these studies also provide meaningful findings of high scientific quality. Furthermore, another explanation for this finding might

be the accumulation of evidence regarding more effective feedback, which was then used to design better interventions. The intervention duration and feedback frequency showed no significant influences on the intervention effect. In addition, our analyses of possible publication bias did not reveal any evidence of systematic bias in the findings.

6.2 Limitations of the study and suggestions for future research

In all studies that we included in our meta-analysis, the same student questionnaires were utilized to gather pre- and post-measures of teaching quality and—in most of the studies—as a tool for collecting student feedback. In other words, the test questionnaires are part of both the assessment and of the intervention. This can lead to validity issues because, as a consequence of an intervention, the meaning of items can change, meaning that identical items can potentially measure different things in pre- and post-tests. For example, if students who sometimes did not finish assignments in time have to answer an item such as "the pace of the lesson was good for me," they must decide for themselves whether they ran out of time because the teacher did not give enough time or because they were wasting time by not focusing on the task. If this student's teacher receives the student feedback that the pace of the teaching is not ideal for task completion, the teacher might address this feedback by asking students how much time they need. This in turn might lead to the students' impression that it is actually the fault of the teacher that they need more time and not their fault because they are wasting time. Thus, students' interpretation of what is "good pace for me" might be different at the two measurement time points. We do not have any evidence that such the assumption of measurement invariance is violated in research on student feedback. However, this is only due to existing studies not investigating or reporting this issue and not necessarily because all instruments are not measurement invariant. This issue is not a problem restricted to our meta-analysis but a problem to all research on the effects of student feedback on teaching quality. Although recent research suggests that measurement invariance may be less relevant when comparing group means than previously assumed (Robitzsch & Lüdtke, 2023), further research in this field should address this issue as it can be addressed with low effort based on the data typically gathered in research on student feedback.

A further limitation to our analysis is the low number of studies, which limits the statistical power of the moderator analyses. Unfortunately, this also applies specifically to student feedback intervention studies in the context of the practical phases of teacher training so that the findings presented here have only limited significance in this regard. Additionally, most current studies (except the study by Novak, 1972a, b) exam intervention effects solely from the students' perspective but do not utilize other ways of surveying teaching quality, such as video ratings by external observers or measures of students' learning achievement. The serious issues regarding the validity of student ratings thus are not addressed in existing intervention studies; however, many related findings from other studies exist that point towards their predictive validity regarding several teaching outcomes (e.g., Ferguson, 2012; Kuhfeld, 2017). To clearly determine the effect of discussing the feedback received with the class for clarification and for the further teaching development process, studies are needed that

include this intervention characteristic as mandatory and vary the strength of support for teachers. Moreover, related effects of the use of student feedback, such as increased student ownership of teaching and their learning, motivational effects, or effects on teacher competence in assessing the quality of their own teaching (Bastian, 2010; Zierer & Wisniewski, 2019), have received little attention so far.

6.3 Conclusion and implications for practical implementation in schools

The findings of the present study suggest that student feedback is a suitable instrument for the improvement of all aspects of teaching quality—at least from the students' perspective, whereby many studies indicate a correlation between students' perception of teaching quality and their learning achievement and motivation. In addition, interview studies also point to the usefulness of this tool from the teacher's perspective (overview: Röhl, 2021). However, it seems essential to carefully design feedback questionnaires and to provide supportive measures for teachers in its interpretation and in the subsequent teaching development steps. For example, the feedback obtained by student perceptions should be at least based on what we know about effective teaching. This requires highly validated questionnaires (Kane, 2006) to make sure that the questionnaire measures what you want to measure and will, thus, provide relevant information upon which to base adjustments and improvements to teaching practice. Moreover, although some student perception questionnaires might seem effective in providing more precise information about what a teacher does into the classroom, such information does not automatically translate into information about how teachers might develop and improve their teaching practice (Van der Lans, 2017).

Obtaining student feedback is relatively inexpensive and has been greatly simplified in recent years by the development of digital tools with validated questionnaires (e.g., Bijlsma et al., 2019; Wisniewski et al., 2020). However, as teachers need support in identifying areas for improvement and implementing changes in their teaching, an effective student feedback intervention requires time and monetary resources—providing yet more evidence for why teacher support should be carefully designed and implemented. This finding is in line with findings on other methods of getting feedback on teaching—such as classroom observations by peers or students' achievement data, which also appear to be effective only for teaching development when certain preconditions are met, such as working with data teams in schools when considering achievement data (Schildkamp et al., 2013), or sufficient time for observations and collaboration for instructional improvement in peer observations (Lasagabaster & Sierrra, 2011). In addition, these methods focus on other information than student feedback (e.g., cognitive outcomes and experts' opinion) and could conceivably, therefore, have different effects on the teaching quality.

Our findings showed no difference between the effectiveness of individual and collective forms of support. This suggests that support can happen in groups or be integrated as an additional element of other forms of teacher collaboration, such as *professional learning communities* (Lomos et al., 2011; Rosenholtz et al., 1986).

A further potential method to promote the effective use of feedback is to discuss the received feedback with the class for clarification and collaboration on ways to improve the quality of the classroom teaching. Such discussions might convey to students that the teacher both values and utilizes their feedback for his or her developmental processes, which in turn might help students to see the value in their own feedback and thus motivate them to give feedback. Additionally, the use of feedback items that refer to the class or teaching situation rather than the teacher as well as the use of frequency scales are also easily implemented design characteristics that could have a positive impact on feedback utilization. Thereby, the feedback items are aimed at facilitating a constructive cognitive processing of the feedback by the teacher and the accompanying affective reactions, which are both necessary for a teacher's goal setting and development of improvement-oriented actions, as proposed in the SFT model (Röhl et al., 2021).

Overall, the findings presented here indicate that student feedback is an effective tool for the professional development of teachers and their teaching, especially when teachers receive support and guidance on how to use it. Further research in this regard has the potential to be fruitful.

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/s11092-024-09450-9.

Author contribution The initial idea for this review with meta-analysis was developed by the first author, Sebastian Röhl. All authors contributed to the further conceptualization and design of the meta-analytic review. Literature search was conducted by Hannah Bijlsma and Sebastian Röhl. All authors conducted the coding of study and effect size information. Data analysis was conducted by Martin Schwichow and Sebastian Röhl. All authors wrote parts of the first draft and commented, revised, and approved the final manuscript.

Funding Open Access funding enabled and organized by Projekt DEAL.

Data availability The datasets generated and analyzed in the current study are not publicly available because some of them are unpublished data provided by the authors of the included primary studies, but they can be obtained from the corresponding author upon request.

Declarations

Conflict of interest The authors declare no competing interests.

References

Studies included in the meta-analysis are marked with an asterisk*

Arens, A. K., & Möller, J. (2016). Dimensional comparisons in students' perceptions of the learning environment. *Learning and Instruction, 42*, 22–30.

Assink, M., & Wibbelink, C. J. M. (2016). Fitting three-level meta-analytic models in R: A step-by-step tutorial. *The Quantitative Methods for Psychology, 12*(3), 154–174. https://doi.org/10.20982/tqmp.12.3.p154

Atlay, C., Tieben, N., Fauth, B., & Hillmert, S. (2019). The role of socioeconomic background and prior achievement for students' perception of teacher support. *British Journal of Sociology of Education, 40*(7), 970–991. https://doi.org/10.1080/01425692.2019.1642737

Balch, R. T. (2012). *The validation of a student survey on teacher practice [Dissertation, Vanderbilt University, Nashville, TN]*. COinS.

Barker, S. (2018). *Student voice to improve instruction: Leading transformation of a school system. Electronic Theses and Dissertations: Vol. 112*. Digital Commons @ ACU.

*Bartel, B. W. (1970). The effectiveness of student feedback in changing teacher classroom image [Dissertation]. University of Minnesota, Minneapolis, MN.

Bastian, J. (2010). Feedbackarbeit in Lehr-Lern-Prozessen. *Gruppendynamik Und Organisationsberatung, 41*(1), 21–37. https://doi.org/10.1007/s11612-010-0097-4

Baumert, J., & Kunter, M. (2013). The effect of content knowledge and pedagogical content knowledge on instructional quality and student achievement. 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. 175–205). US: Springer. https://doi.org/10.1007/978-1-4614-5149-5_9

Beattie, H. (2012). Amplifying student voice: The missing link in school transformation. *Management in Education, 26*(3), 158–160.

*Bell, C. A., & Aldridge, J. M. (2014). Investigating the use of student perception data for teacher reflection and classroom improvement. Learning environment research, 17, 371-388. https://doi.org/10.1007/s10984-014-9164-z

*Bennett, C. R. (1978). A developed and field-tested experiment to test the effect of increased student feedback on specific teacher performance behaviors [Dissertation]. Iowa State University, Ames, IA.

Benton, S. L., & Cahsin, W. E. (2012). *Student ratings of teaching: A summary of the research and literature*. (IDEA Paper No. 50). Retreived Oct. 15th 2024, from https://ideacontent.blob.core.windows.net/content/sites/2/2020/01/PaperIDEA_50.pdf

Bertrand, M., & Marsh, J. A. (2015). Teachers' sensemaking of data and implications for equity. *American Educational Research Journal, 52*(5), 861–893. https://doi.org/10.3102/0002831215599251

*Bijlsma, H. J., Visscher, A. J., Dobbelaer, M. J., & Veldkamp, B. P. (2019). Does smartphone-assisted student feedback affect teachers' teaching quality? *Technology, Pedagogy and Education*, *28*(2), 217–236. https://doi.org/10.1080/1475939X.2019.1572534

Borenstein, M., Hedges, L. V., Higgins, J. P. T., & Rothstein, H. R. (2009). *Introduction to meta-analysis*. Wiley.

*Buurman, M., Delfgaauw, J., Dur, R., & Zoutenbier, R. (2020). When do teachers respond to student feedback? Evidence from a field experiment. *Labour Economics*, *65*. https://doi.org/10.1016/j.labeco.2020.101858

Carlson, K. D., & Schmidt, F. L. (1999). Impact of experimental design on effect size: Findings from the research literature on training. *The Journal of Applied Psychology, 84*(6), 851–862. https://doi.org/10.1037/0021-9010.84.6.851

Cohen, P. A. (1980). Effectiveness of student-rating feedback for improving college instruction: A meta-analysis of findings. *Research in Higher Education, 13*(4), 321–341. https://doi.org/10.1007/BF00976252

Creemers, B. M. P. (1994). *The effective classroom*. Cassall.

Day, C., Sammons, P., & Kington, A. (2008). *Effective classroom practice (ECP): A mixed-method study of influences and outcomes*. University of Nottingham.

Den Brok, P., Brekelmans, M., & Wubbels, T. (2006). Multilevel issues in research using students' perceptions of learning environments: The case of the Questionnaire on Teacher Interaction. *Learning Environments Research, 9*(3), 199–213. https://doi.org/10.1007/s10984-006-9013-9

*Ditton, H., & Arnold, B. (2004). Wirksamkeit von Schülerfeedback zum Fachunterricht. [Effectiveness of student feedback on subject teaching.]. In J. Doll & M. Prenzel (Eds.), *Bildungsqualität von Schule. Lehrerprofessionalisierung, Unterrichtsentwicklung und Schülerförderung als Strategien der Qualitätsentwicklung* (pp. 152–172). Waxmann.

Egger, M., Smith, G. D., Schneider, M., & Minder, C. (1997). Bias in meta-analysis detected by a simple, graphical test. *BMJ, 315*(7109), 629–634. https://doi.org/10.1136/bmj.315.7109.629

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

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

Ferguson, R. (2012). Can student surveys measure teaching quality? *Phi Delta Kappan, 94*(3), 24–28.

Fernández-Castilla, B., Declercq, L., Jamshidi, L., Beretvas, S. N., Onghena, P., & van den Noortgate, W. (2021). Detecting selection bias in meta-analyses with multiple outcomes: A simulation study. *The Journal of Experimental Education, 89*(1), 125–144. https://doi.org/10.1080/00220973.2019.1582470

*Fraser, B. J., Seddon, T., & Eagleson, J. (1982). Use of student perceptions in facilitating improvement in classroom environment. *Australian Journal of Teacher Education*, *7*(1). https://doi.org/10.14221/ajte.1982v7n1.3

Fraser, B. J., & Fisher, D. L. (1983). Development and validation of short forms of some instruments, measuring student perceptions of actual and preferred classroom learning environment. *Science Education, 67*(1), 115–131.

Fraser, B. J., & Fisher, D. L. (1986). Using short forms of classroom climate instruments to assess and improve classroom psychosocial environment. *Journal of Research in Science Teaching, 23*(5), 387–413. https://doi.org/10.1002/tea.3660230503

Gaertner, H. (2014). Effects of student feedback as a method of self-evaluating the quality of teaching. *Studies in Educational Evaluation, 42*, 91–99. https://doi.org/10.1016/j.stueduc.2014.04.003

*Gage, N. L. (1960). *Equilibrium theory and behavioral change: An experiment in feedback from pupils to teachers.* Urbana: Illinois University, Urbana.

*Gage, N. L. (1963). A method for "improving" teacher behavior. *Journal of teacher education, 14*, 261–266.https://doi.org/10.1177/002248716301400306

Göbel, K., Wyss, C., Neuber, K., & Raaflaub, M. (2021). Student feedback as a source for reflection in practical phases of teacher education. 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* (pp. 173–189). Springer Nature. https://doi.org/10.1007/978-3-030-75150-0_11

Göllner, R., Wagner, W., Eccles, J. S., & Trautwein, U. (2018). Students' idiosyncratic perceptions of teaching quality in mathematics: A result of rater tendency alone or an expression of dyadic effects between students and teachers? *Journal of Educational Psychology, 110*(5), 709–725. https://doi.org/10.1037/edu0000236

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

Hattie, J. (2009). *Visible learning: A synthesis of over 800 meta-analyses relating to achievement*. Routledge.

Hattie, J., & Timperley, H. (2007). The power of feedback. *Review of Educational Research, 77*(1), 81–112. https://doi.org/10.3102/003465430298487

Hill, H. C., Charalambous, C. Y., & Kraft, M. A. (2012). When rater reliability is not enough: Teacher observation systems and a case for the generalizability study. *Educational Researcher, 41*(2), 56–64.

Hox, J. J., Moerbeek, M., & van de Schoot, R. (2018). *Multilevel analysis* (3rd ed.). UK: Routledge. https://doi.org/10.4324/9781315650982

Hunter, J. E., & Schmidt, F. L. (2007). *Methods of meta-analysis: Correcting error and bias in research findings* (2nd ed.). SAGE.

Ilgen, D. R., Fisher, C. D., & Taylor, M. S. (1979). Consequences of individual feedback on behavior in organizations. *Journal of Applied Psychology, 64*(4), 349–371. https://doi.org/10.1037/0021-9010.64.4.349

Kahmann, K., & Mulder, R. H. (July 2011). Feedback in organizations: A review of feedback literature and a framework for future research (Research Report No. 6). Regensburg. Institute for Educational Science, Universität Regensburg.

Kane, R. G., & Maw, N. (2005). Making sense of learning at secondary school: Involving students to improve teaching practice. *Cambridge Journal of Education, 35*(3), 311–322.

Kane, M. T. (2006). Validation. In R. L. Brennan (Ed.), *Educational Measurement* (4th ed., pp. 17–64). American Council on Education & Praeger Publishers.

Kember, D., Leung, D. Y., & Kwan, K. (2002). Does the use of student feedback questionnaires improve the overall quality of teaching? *Assessment & Evaluation in Higher Education, 27*(5), 411–425. https://doi.org/10.1080/0260293022000009294

Kluger, A. N., & DeNisi, A. (1996). The effects of feedback interventions on performance: A historical review, a meta-analysis, and a preliminary feedback intervention theory. *Psychological Bulletin, 119*(2), 254–284.

*Knox, J. T. (1973). *Student feedback as a means for improving teacher effectiveness* [Dissertation]. Wayne State University, Detroit, MI.

Konstantopoulos, S. (2011). Fixed effects and variance components estimation in three-level meta-analysis. *Research Synthesis Methods, 2*(1), 61–76. https://doi.org/10.1002/jrsm.35

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

Kyriakides, L., Creemers, B. P. M., & Antaniou, P. (2009). Teacher behavior and student outcomes: Suggestions for research on teacher training and professional development. *Teaching and Teacher Education, 25*, 12–23.

L'Hommedieu, R., Menges, R. J., & Brinko, K. T. (1990). Methodological explanations for the modest effects of feedback from student ratings. *Journal of Educational Psychology, 82*(2), 232–241. https://doi.org/10.1037/0022-0663.82.2.232

Landis, J. R., & Koch, G. G. (1977). The measurement of observer agreement for categorical data. *Biometrics, 33*(1), 159–174.

Lang, J. W., & Kersting, M. (2007). Regular feedback from student ratings of instruction: Do college teachers improve their ratings in the long run? *Instructional Sciences, 35*(3), 187–205.

Lasagabaster, D., & Sierrra, J. M. (2011). Classroom observation: Desirable conditions established by teachers. *European Journal of Teacher Education, 34*, 449–463.

*Lauroesch, W. P., Pereira, P. D., & Ryan, K. A. (1969). *The use of student feedback in teacher training* [Dissertation]. University of Chicago, Chicago,IL.

Liberati, A., Altman, D. G., Tetzlaff, J., Murrow, C., Gøtzsche, P. C., Ioannidis, J. P. A., & Moher, D. (2009). The PRISMA statement for reporting systematic reviews and meta-analyses of studies that evaluate health care interventions: Explanation and elaboration. *PLoS Med, 6*(7), W65. https://doi.org/10.1371/journal.pmed.1000100

Lipsey, M. W., & Wilson, D. B. (2001). *Practical meta-analysis*. *Applied Social Research Methods Series: Vol. 49*. SAGE.

Littell, J. H., Corcoran, J., & Pillai, V. (2008). *Systematic reviews and meta-analysis*. Oxford University Press.

Lomos, C., Hofman, R. H., & Bosker, R. J. (2011). Professional communities and student achievement – A meta-analysis. *School Effectiveness and School Improvement, 22*(2), 121–148. https://doi.org/10.1080/09243453.2010.550467

*Mandouit, L. (2018). Using student feedback to improve teaching. *Educational Action Research*.https://doi.org/10.1080/09650792.2018.1426470

Maulana, R., & Helms-Lorenz, M. (2016). Observations and student perceptions of the quality of preservice teachers' teaching behaviour: Construct representation and predictive quality. *Learning Environments Research, 19*(3), 335–357. https://doi.org/10.1007/s10984-016-9215-8

Maulana, R., Helms-Lorenz, M., & van de Grift, W. (2015). Development and evaluation of a questionnaire measuring pre-service teachers' teaching behaviour: A Rasch modelling approach. *School Effectiveness and School Improvement, 26*(2), 169–194. https://doi.org/10.1080/09243453.2014.939198

*Mayr, J. (1993). *Mitarbeit und Störung im Unterricht: Beschreibung und Evaluierung eines Konzepts zur Verbesserung pädagogischen Handelns*. Linz. Pädagogische Akademie der Diözese Linz.

Muijs, D. (2006). Measuring teacher effectiveness: Some methodological reflections. *Educational Research and Evaluation, 12*(1), 53–75.

Muijs, D., Kyriakides, L., van der Werf, G., Creemers, B. M. P., Timperley, H., & Earl, L. (2014). State of the art – Teacher effectiveness and professional learning. *School Effectiveness and School Improvement, 25*, 231–256.

Nelson, P. M., Demers, J. A., & Christ, T. J. (2014). The Responsive Environmental Assessment for Classroom Teaching (REACT): The dimensionality of student perceptions of the instructional environment. *School Psychology Quarterly: The Official Journal of the Division of School Psychology, American Psychological Association, 29*(2), 182–197. https://doi.org/10.1037/spq0000049

Nelson, P. M., Ysseldyke, J. E., & Christ, T. J. (2015). Student perceptions of the classroom environment: Actionable feedback to guide core instruction. *Assessment for Effective Intervention, 41*(1), 16–27. https://doi.org/10.1177/1534508415581366

*Novak, J. H. (1972a). *A study of the effects of the use of a pupil response instrument on the behaviors of biological science teachers.* National Center for Educational Research and Development.

*Novak, J. H. (November 1972b). *A study of the effects of the use of a pupil response instrument on the behaviors of biological science teachers*. Final Report. Pittsburgh, PA.

Penny, A. R., & Coe, R. (2004). Effectiveness of consultation on student ratings feedback: A meta-analysis. *Review of Educational Research, 74*(2), 215–253. https://doi.org/10.3102/00346543074002215

Peterson, K. D., Wahlquist, C., & Bone, K. (2000). Student surveys for school teacher evaluation. *Journal of Personnel Evaluation in Education, 14*(2), 135–153.

Pianta, R. C., & Hamre, B. K. (2006). Conceptualization, measurement, and improvement of classroom processes: Standardized observation van leverage capacity. *Educational Researcher, 38*(2), 109–119.

Porter, W. A. (1942). Pupil evaluation of practice teaching. *The Journal of Educational Research, 35*(9), 700–704.

Praetorius, A.-K., & Charalambous, C. Y. (2018). Classroom observation frameworks for studying instructional quality: Looking back and looking forward. *ZDM, 50*(3), 535–553. https://doi.org/10.1007/s11858-018-0946-0

Praetorius, A. K., Klieme, E., Herbert, B., & Pinger, P. (2018). Generic dimensions of teaching quality: The German framework of Three Basic Dimensions. *ZDM, 50*(3), 407–426.

Pustejovsky, J. E., & Rodgers, M. A. (2019). Testing for funnel plot asymmetry of standardized mean differences. *Research Synthesis Methods, 10*(1), 57–71. https://doi.org/10.1002/jrsm.1332

Pustejovsky, J. E., & Tipton, E. (2018). Small-sample methods for cluster-robust variance estimation and hypothesis testing in fixed effects models. *Journal of Business & Economic Statistics, 36*(4), 672–683. https://doi.org/10.1080/07350015.2016.1247004

Pustejovsky, J. E. (2021). *clubSandwich: Cluster-robust (sandwich) variance estimators with small-sample corrections* (Version 0.5.3) [Computer software]. https://CRAN.R-project.org/package=clubSandwich

Remmers, H. H. (1927). The Purdue rating scale for instructors. *Educational Administration and Supervision, 6*, 399–406.

Remmers, H. H. (1934). Reliability and halo effect of high school and college students' judgments of their teachers. *Journal of Applied Psychology, 18*(5), 619–630. https://doi.org/10.1037/h0074783

Reynolds, D., Sammons, P., De Fraine, B., van Damme, J., Towsend, T., Teddlie, C., & Stringfield, S. (2014). Educational effectiveness research (EER): A state-of-the-art review. *School Effectiveness and School Improvement, 25*(2), 197–230.

Richardson, J. T. E. (2005). Instruments for obtaining student feedback: A review of the literature. *Assessment and Evaluation in Higher Education, 30*, 387–415. https://doi.org/10.1080/02602930500099193

Robitzsch, A. (2020). Why ordinal variables can (almost) always be treated as continuous variables: Clarifying assumptions of robust continuous and ordinal factor analysis estimation methods. *Frontiers in Education, 5*, 177. https://doi.org/10.3389/feduc.2020.589965

Robitzsch, A., & Lüdtke, O. (2023). Why full, partial, or approximate measurement invariance are not a prerequisitefor meaningful and valid group comparisons. *Structural Equation Modeling: A Multidisciplinary Journal, 30*(6), 859–870.

Rodgers, M. A., & Pustejovsky, J. E. (2021). Evaluating meta-analytic methods to detect selective reporting in the presence of dependent effect sizes. *Psychological Methods, 26*(2), 141–160. https://doi.org/10.1037/met0000300

Röhl, S. (2021). Effects of student feedback on teaching and classes: An overview and meta-analysis on intervention studies. In W. Rollett, H. J. E. Bijlsma, & S. Röhl (Eds.), *Student feedback in schools: Using perceptions for the development of teaching and teachers.* Springer Nature: Cham, CH.

Röhl, S., & Rollett, W. (2020). Alles nur die Nettigkeit der Lehrkraft? Die Communion der Lehrkräfte als Erklärung für den Halo-Bias in Schülerbefragungen zur Unterrichtsqualität. *Empirische Pädagogik, 34*(1), 30–45.

Röhl, S., & Gärtner, H. (2021a). 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.*

Röhl, S., & Rollett, W. (2021b). Jenseits von Unterrichtsentwicklung: Intendierte und nicht-intendierte Nutzungsformen von Schülerfeedback durch Lehrpersonen. In K. Göbel, C. Wyss, K. Neuber, & M. Raaflaub (Eds.), *Quo vadis Forschung zu Schülerrückmeldungen zum Unterricht: Konzeptionelle Überlegungen und empirische Befunde zu Chancen und Herausforderungen* (pp. 167–189). Springer VS. https://doi.org/10.1007/978-3-658-32694-4_9

Röhl, S., Bijlsma, H. J. E., & Rollett, W. (2021). The Process Model of Student Feedback on Teaching (SFT): a theoretical framework and introductory remarks. 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* (pp. 1–14). Springer Nature. https://doi.org/10.1007/978-3-030-75150-0_1

*Rösch, S. (2017). *Wirkung und Wirkmechanismen von regelmäßigem Schülerfeedback in der Sekundarstufe: Eine explorative Untersuchung im Physikunterricht* [Effect and impact mechanisms of regular student feedback in secondary education: An exploratory study in physics classrooms. Dissertation]. Universität Basel, Basel (CH).

Rosenholtz, S. J., Bassler, O., & Hoover-Dempsey, K. (1986). Organizational conditions of teacher learning. *Teaching and Teacher Education, 2*(2), 91–104. https://doi.org/10.1016/0742-051X(86)90008-9

Ryan, K. A. (1974). The use of feedback from students in the preservice training of teachers. American Educational Research Association. Annual Meeting of the American Educational Research Association, Chicago, IL.

Sammons, P., Hillman, J., & Mortimore, P. (1995). *Key characteristics of effective schools: A review of school effectiveness research*. Institute of Education.

Saris, W., Revilla, M., Krosnick, J. A., & Shaeffer, E. M. (2010). Comparing questions with agree/disagree response options to questions with item-specific response options. *Survey Research Methods, 4*(1), 61–79. https://doi.org/10.18148/srm/2010.v4i1.2682

*Saunders, K. J., & Fisher, D. L. (2006). An action research approach with primary pre-service teachers to improve university and primary school classroom environments. In D. L. Fisher & M. S. Khine (Eds.), *Contemporary approaches to research on learning environments: Worldviews* (pp. 247–272). World Scientific. https://doi.org/10.1142/9789812774651_0010

Schiepek, G., Eckert, H., Aas, B., Wallot, S., & Wallot, A. (2015). *Integrative psychotherapy: A feedback-driven dynamic systems approach*. Hogrefe Publishing.

Schildkamp, K., Lai, M. K., & Earl, L. (Eds.). (2013). *Data-based decision making in education*. Netherlands: Springer. https://doi.org/10.1007/978-94-007-4816-3

*Schmidt, J.-E. (2018). *Verborgene Kräfte im Klassenzimmer wecken* [Unleashing hidden power in the classroom. Dissertation]. Eberhard Karls Universität, Tübingen.

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. https://doi.org/10.3102/0034654307310317

Smither, J. W., London, M., & Reilly, R. R. (2005). Does performance improve following multisource feedback? A theoretical model, meta-analysis, and review of empirical findings. *Personnel Psychology, 58*(1), 33–66. https://doi.org/10.1111/j.1744-6570.2005.514_1.x

Stahns, R., Rieser, S., & Hußmann, A. (2020). Können Viertklässlerinnen und Viertklässler Unterrichtsqualität valide einschätzen? *Ergebnisse Zum Fach Deutsch. Unterrichtswissenschaft, 48*(4), 663–682. https://doi.org/10.1007/s42010-020-00084-6

Stalnaker, J. M., & Remmers, H. H. (1928). Can students discriminate traits associated with success in teaching? *The Journal of Applied Psychology, 12*(6), 602–610. https://doi.org/10.1037/h0070372

Stigler, J. W., & Miller, K. F. (2018). Expertise and expert performance in teaching. In A. M. Williams, A. Kozbelt, K. A. Ericsson, & R. R. Hoffman (Eds.). *The Cambridge Handbook of Expertise and Expert Performance* (2nd ed., pp. 431–452). Cambridge University Press. https://doi.org/10.1017/9781316480748.024

Strong, M., Gargani, J., & Hacifazlioğlu, Ö. (2011). Do we know a successful teacher when we see one? Experiments in the identification of effective teachers. *Journal of Teacher Education, 62*(4), 367–382. https://doi.org/10.1177/2F0022487110390221

*Tacke, G., & Hofer, M. (1979). Behavioral changes in teachers as a function of student feedback: A case for the achievement motivation theory? *Journal of school psychology, 17*, 172–180.https://doi.org/10.1016/0022-4405(79)90025-6

*Thorp, H. S., Burden, R. L., & Fraser, B. J. (1994). Assessing and improving classroom environment. *School Science Review*, *75*, 107–113.

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

*Tozoglu, D. (2006). *Effects of student ratings feedback on instructional practices, teaching effectiveness, and student motivation*. Dissertation: Florida State University.

Tuckman, B. W., & Olivier, W. F. (1968). Effectiveness of feedback to teachers as a function of source. *Journal of Educational Psychology, 59*, 297–301.

Tuckman, B. W., & Yates, D. (1980). Evaluating the student feedback strategy for changing teacher style. *Journal of Educational Research, 74*(2), 74–77. https://doi.org/10.1080/00220671.1980.10885286

van de Grift, W. J. C. M. (2007). Quality of teaching in four European countries: A review of the literature and application of an assessment instrument. *Educational Research and Evaluation, 49*(2), 127–152.

van de Grift, W. J. C. M., Helms-Lorenz, M., & Maulana, R. (2014). Teaching skills of student teachers: Calibration of an evaluation instrument and its value in predicting student academic engagement. *Studies in Educational Evaluation, 43*, 150–159.

van der Lans, R. M., Van de Grift, W. J. C. M., & van Veen, K. (2018). Developing an instrument for teacher feedback: Using the Rasch model to explore teachers' development of effective teaching strategies and behaviors. *The Journal of Experimental Education, 86*(2), 247–264. https://doi.org/10.1080/00220973.2016.1268086

van der Lans, R. M., de Grift, W. J., & van Veen, K. (2019). Same, similar, or something completely different? Calibrating student surveys and classroom observations of teaching quality onto a common metric. *Educational Measurement: Issues and Practice, 38*(3), 55–64. https://doi.org/10.1111/emip.12267

van der Scheer, E. A., Bijlsma, H. J. E., & Glas, C. A. W. (2019). Validity and reliability of student perceptions of teaching quality in primary education. *School Effectiveness and School Improvement, 30*(1), 30–50. https://doi.org/10.1080/09243453.2018.1539015

Veldman, D. J., & Peck, R. F. (1969). Influences on pupil evaluations of student teachers. *Journal in Educational Psychology, 60*(2), 103–108.

Viechtbauer, W. (2010). Conducting meta-analyses in R with the metafor package. *Journal of Statistical Software, 36*(3), 1–48.

Wagner, W. (2008). *Methodenprobleme bei der Analyse der Unterrichtswahrnehmung aus Schülersicht – am Beispiel der Studie DESI* [Dissertation, Universität Koblenz-Landau, Koblenz-Landau]. RIS. http://kola.opus.hbz-nrw.de/volltexte/2008/234

Wallace, T. L., Kelcey, B., & Ruzek, E. A. (2016). What can student perception surveys tell us about teaching? Empirically testing the underlying structure of the tripod student perception survey. *American Educational Research Journal, 53*(6), 1834–1868. https://doi.org/10.3102/0002831216671864

Wang, F., Liu, Y., & Leung, S. O. (2022). Disciplinary climate, opportunity to learn, and mathematics achievement: An analysis using doubly latent multilevel structural equation modeling. *School Effectiveness and School Improvement, 33*(3), 479–496. https://doi.org/10.1080/09243453.2022.2043393

Weiner, B. (1985). An attributional theory of achievement motivation and emotion. *Psychological Review, 92*(4), 548–573.

Williams, G. D. (1962). Your students can help you be a better teacher. *Illinois Education, 50*, 280–289.

Wisniewski, B., Zierer, K., & Hattie, J. (2020). The power of feedback revisited: A meta-analysis of educational feedback research. *Frontiers in Psychology, 10*, 3087. https://doi.org/10.3389/fpsyg.2019.03087

Wisniewski, B., Röhl, S., & Fauth, B. (2022). The perception problem: A comparison of teachers' self-perceptions and students' perceptions of instructional quality. *Learning Environments Research, 25*, 775–802. https://doi.org/10.1007/s10984-021-09397-4

Zierer, K., & Wisniewski, B. (2019). *Using student feedback for successful teaching*. Routledge.