Does teacher judgment accuracy matter? How judgment accuracy, teaching quality, and student achievement development are related

Kolovou, Dimitra ; Hochweber, Jan ; Praetorius, Anna-Katharina

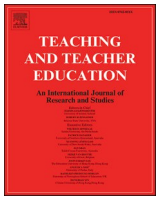
DOI: https://doi.org/10.1016/j.tate.2024.104555

Teaching and Teacher Education 144 (2024) 104555

Contents lists available at ScienceDirect

Teaching and Teacher Education

journal homepage: www.elsevier.com/locate/tate



Research paper

Does teacher judgment accuracy matter? How judgment accuracy, teaching quality, and student achievement development are related☆,☆☆

Dimitra Kolovou [a,b,*], Jan Hochweber [a], Anna-Katharina Praetorius [b]

[a] Institute of Educational Psychology, St. Gallen University of Teacher Education (PHSG), Notkerstrasse 27, 9000, St. Gallen, Switzerland
[b] Institute of Education, University of Zurich, Freiestrasse 36, 8032, Zurich, Switzerland

ARTICLE INFO

Keywords:
Judgment accuracy
Teacher judgments
Academic achievement
Teaching quality
Multilevel mediation

ABSTRACT

Teacher judgment accuracy is assumed to be positively related to student achievement; however, the empirical evidence for this assumption is inconclusive. Using two accuracy indicators measured at different levels (class/teacher- and student-level), we examined the theoretically hypothesized effects of judgment accuracy on German language achievement over a three-year period, and tested whether teaching quality mediates these effects. We applied multilevel mediation models and small sample methods to data from 35 language teachers and 646 students from 42 classes. While no mediating effects were found, the student-level indicator positively predicted achievement, suggesting that student-level accuracy measures should receive more attention.

Portions of these findings were presented at the 2023 conference of the European Association for Research on Learning and Instruction (EARLI), Thessaloniki, Greece. There are no known conflicts of interest to report.

Teachers make numerous routine judgments about how their students are learning every day to inform their teaching practice. The extent to which these judgments are accurate has been receiving growing attention from researchers (Urhahne & Wijnia, 2021) because the ability to accurately judge students is a prerequisite for being able to adapt teaching to their needs and strengths (Hoge & Coladarci, 1989; Wammes, Slof, Schot, & Kester, 2023). The underlying assumption is that this will lead to better student outcomes such as higher academic achievement (Meissel, Yao, & Meyer, 2022; Ready & Wright, 2011; Thiede, Oswalt, Brendefur, Carney, & Osguthorpe, 2019). This assumption, however, is not clearly supported by evidence. Studies have reported statistically significant positive, non-significant, or even significant negative effects of teacher accuracy on achievement (see Fig. 1A; Anders, Kunter, Brunner, Krauss, & Baumert, 2010; Förster, Humberg, Hebbecker, Back, & Souvignier, 2022). To determine why the results are so inconclusive, the mechanisms that operate between a teacher's judgment and student achievement must be further examined.

One popular hypothesis is that teaching quality mediates the effect of judgment accuracy (see Fig. 1B; Brunner, Anders, Hachfeld, & Krauss, 2013; Thiede et al., 2018; Urhahne & Wijnia, 2021). The causal sequence where teacher accuracy affects teaching quality which in turn affects student achievement is often a key argument to underline the importance of judgment accuracy. To date, however, there is little empirical support for this sequence and there have been no longitudinal studies of sufficient duration to pick up any long-term effects. Previous studies have focused on the short-term effects of judgment accuracy, collecting data over the course of a few lessons or one school year. However, in many countries students are taught by the same teacher for more than a year, so that it might be more informative to assess the effect of their accuracy over a longer time span. It has also been argued that studying long-term effects is the best way to uncover the true extent of teacher/school influences since short-term effects may quickly fade (Dimosthenous, Kyriakides, & Panayiotou, 2020).

Furthermore, researchers have focused on measuring judgment accuracy and teaching quality at the class-level. But there is some evidence that the way a specific student is viewed by their teacher affects the student's perceptions of this teacher, and in turn, their outcomes (Stang & Urhahne, 2016; Zhu, Urhahne, & Rubie-Davies, 2018). For similar reasons, there has been increased interest in the views of individual students by researchers in related fields (e.g., teaching quality; Göllner,

☆ For legal reasons, the data used in this article cannot be shared at this time. The authors may be able to arrange access to the data, but permission from the commissioner of the project from which the data were drawn would be required.☆☆ Special thanks to Alexander Naumann for helpful and inspiring discussions in the early stages of work on this article. We would also like to thank Ayse Yenal Vance for help with language and proofreading.

* Corresponding author. Institute of Educational Psychology, St. Gallen University of Teacher Education (PHSG), Notkerstrasse 27, 9000 St. Gallen, Switzerland.
E-mail address: dimitra.kolovou@phsg.ch (D. Kolovou).

https://doi.org/10.1016/j.tate.2024.104555
Received 3 September 2023; Received in revised form 13 February 2024; Accepted 12 March 2024
Available online 4 April 2024

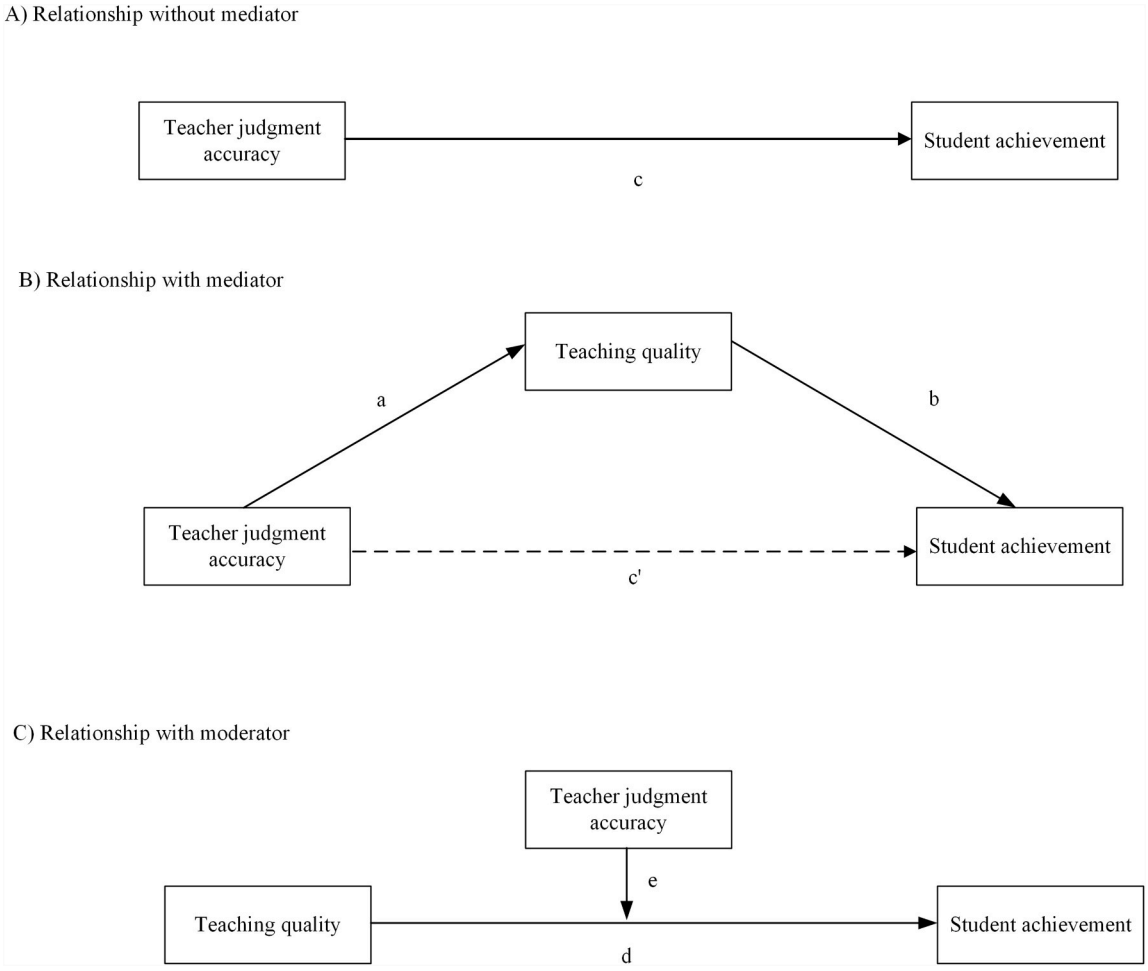


Fig. 1. Conceptual models and paths tested in previous studies.

Fauth, & Wagner, 2021). It is therefore important to measure judgment accuracy at both the student- and the class/teacher-level.

Finally, most studies on the effects of judgment accuracy have focused on mathematics rather than language. As language proficiency is evidently important for domain-specific learning, being the medium for both classroom communication and individual knowledge building, and also an indispensable resource for understanding test items and recalling learned material (Kempert, Schalk, & Saalbach, 2019; Peng et al., 2020; Zhu, 2022), further research is needed in this domain.

To address these issues, we examine how teachers' judgment accuracy affects students' German[1] language achievement by investigating the mediating role of four aspects of teaching quality measured using student ratings. We use data from a longitudinal study conducted over a three-year period, in which teacher accuracy and teaching quality were assessed both at the student- and the class/teacher-level.

1. Teacher judgment accuracy and how it relates to student achievement

Teacher judgment accuracy is the level of agreement between a teacher's judgment of a student characteristic, such as mathematics ability, and external scores from achievement tests or self-reported measurements (Hoge & Coladarci, 1989; Kaufmann, 2022). To operationalize judgment accuracy, researchers commonly use the rank component, which reveals how accurately teachers can rank order their students' achievement levels within class for the chosen criterion (Südkamp, Kaiser, & Möller, 2012; Urhahne & Wijnia, 2021). To estimate the rank component for student achievement, the correlation between teacher judgment and student test performance is calculated for each classroom or teacher (Kaufmann, 2020). According to meta-analyses by Hoge and Coladarci (1989; see also Kaufmann, 2020) and Südkamp et al. (2012), achievement is judged relatively accurately with a mean correlation between teacher judgments and student test results of $r^{o} = {}^{o}0.65$ (median = 0.66) and $r^{o} = {}^{o}0.63$ (median$^{o} = {}^{o}0.53$), respectively.

From a methodological perspective, a more suitable alternative to the correlation is to use the (random) slope for test performance (i.e., achievement test scores) when predicting teacher judgments using multilevel regression (Dollinger, 2013; Karst & Bonefeld, 2020).

While the rank component, operationalized as the correlation or the (random) slope, captures accuracy at the class/teacher-level, other, less commonly used accuracy indicators, focus on the student-level. These indicators capture how well the teacher judges each student, which is important in situations with an emphasis on providing support to individual students, especially those with lower ability (Begeny, Krouse, Brown, & Mann, 2011; Pielmeier, Huber, & Seidel, 2018; Wadmare, Nanda, Sabates, Sunder, & Wadhwa, 2022). To date, student-level

[1] In this study, German is assessed as the language of instruction.

accuracy measures have been based on the arithmetic difference[2] between teacher judgment and student achievement, calculated for each student. This difference is either used directly as an accuracy measure (Pielmeier et al., 2018) or to categorize students as over- or underestimated by the teacher (Stang & Urhahne, 2016; Urhahne, 2015).

Empirical evidence for the predictive power of teachers' student achievement judgment accuracy is sparse and inconclusive. Contrary to the theoretical expectation that high judgment accuracy positively affects achievement (cf. Fig. 1A), many studies using the rank component have reported no statistically significant effects (Brühwiler, 2017; Karing, Pfost, & Artelt, 2011; Schrader, 1989) or significant negative effects (Karst, Schoreit, & Lipowsky, 2014; Lingelbach, 1995). When significant positive effects were found, they were small (Anders et al., 2010; Thiede et al., 2018). Evidence is similarly inconsistent regarding other class/teacher-level measures (Gabriele, Joram, & Park, 2016; Hill & Chin, 2018; Karing et al., 2011).

Studies using student-level measurements found that teachers tend to overestimate their students (Bates & Nettelbeck, 2001; Thiede et al., 2018; Urhahne, 2015). The impact on student outcomes has been rarely studied, but preliminary evidence indicates that underestimation of student ability is associated with poorer learning outcomes (Bergold & Steinmayr, 2023; Meissel et al., 2022). Specifically, students whose skills had been underestimated tended to have less favorable perceptions of teacher behavior (accessibility, learning support, grading fairness) than overestimated students, which in turn had a negative effect on their motivational and cognitive outcomes (Stang & Urhahne, 2016; Urhahne, 2015). As far as we can ascertain, no study has investigated whether learning outcomes differ between students who have been judged accurately or inaccurately or whether the accuracy of teacher judgments is related to individual differences in students' perceptions of teaching.

Besides mixed results, previous studies share another characteristic: their focus on short-term effects (i.e., effects evaluated after a school year or few lessons; e.g., Brühwiler, 2017). For effects of teacher judgment accuracy to be measurable, teachers may need to be working with their students for a longer time. The more time, the more opportunities teachers have to use their (more or less) accurate knowledge about their students in their teaching effectively. Research has shown that an effect of teacher judgments can become apparent up to four school years later for judgments of academic abilities (Hinnant, O'Brien, & Ghazarian, 2009) and even longer for judgments of intelligence (Alvidrez & Weinstein, 1999; Fischbach, Baudson, Preckel, Martin, & Brunner, 2013).

Summarizing, previous research using class/teacher-level measures of judgment accuracy reported highly inconsistent and often statistically non-significant effects on student achievement necessitating investigation of the underlying mechanisms. Investigating indirect effects can be useful even when the total effect is not statistically significant, as non-significance of the total effect does not necessarily imply that teacher's accuracy has no relevance for achievement (e.g., in situations where mediators operate in contrary directions, or predictor and mediator variables differ in measurement precision; cf. Hayes, 2009; O'Rourke & MacKinnon, 2018; Rucker, Preacher, Tormala, & Petty, 2011). Further, it is important to extend current knowledge about the effects of teacher accuracy to the student-level to unravel effect pathways that might otherwise be overlooked and to systematically compare different accuracy measures in terms of their predictive power (Karst et al., 2014).

1.1. Teaching quality and how it links teacher accuracy to achievement

Teaching quality has been suggested as a mediator for the relationship between teacher judgment accuracy and student achievement. The few studies which have actually explored this relationship can be divided into two groups. The first examined the effect of teacher judgment accuracy on teaching quality (Fig. 1B, path a). The second examined the relationships between accuracy, teaching quality, and achievement but proposed two different mechanisms for how they are related; while some investigated whether the *interaction* between teacher accuracy and teaching quality predicted achievement (Fig. 1C, path e), two others investigated whether teaching quality *mediates* the effect of accuracy on achievement (Fig. 1B, path a*b). A review of the studies also shows that there is no agreement on a definition of teaching quality, a complex phenomenon with many dimensions, with judgment accuracy having a different meaning for each (Behrmann & Souvignier, 2013; Charalambous & Praetorius, 2020).

Researchers have developed several frameworks and models to conceptualize and measure teaching and its quality. Teaching is viewed as a social practice co-constructed by teachers and students around content to facilitate student learning of specific learning goals (Alp Christ, Capon-Sieber, Grob, & Praetorius, 2022; Charalambous et al., 2021; Praetorius, Klieme, Herbert, & Pinger, 2018). Teaching quality refers to the kind of teaching that creates learning opportunities which increase the likelihood of desired student outcomes (Charalambous et al., 2021; Praetorius et al., 2018). Accordingly, empirical evidence shows that teaching quality is associated with the development of cognitive (e.g., achievement) and non-cognitive outcomes (e.g., self-efficacy; Alp Christ, Capon-Sieber, Grob, & Praetorius, 2022; Kunter et al., 2013; Praetorius et al., 2018).

1.2. The MAIN-TEACH model as a framework for analyzing teaching quality dimensions related to teachers' accuracy

Previous studies of teaching quality and judgment accuracy have looked at a variety of dimensions but have not referred to a specific teaching quality model or framework. This study is based on the MAIN-TEACH model (Charalambous & Praetorius, 2020, 2022; Praetorius et al., 2023), which stands for *m*ulti-l*a*yered and *i*ntegrated in conceptualizing the quality of *teach*ing. The model represents an up-to-date, systematic synthesis and further development of many existing international frameworks and models (e.g., CLASS, Berlin & Cohen, 2018; MQI, Charalambous & Litke, 2018; Three Basic Dimensions; Klieme, Schümer, & Knoll, 2001) and integrates both generic and subject-specific aspects of teaching quality. MAIN-TEACH understands teaching as an interplay of learning opportunities designed by teachers and the use of these learning opportunities by students, which in turn can lead to specific effects on students. The teaching quality dimensions are structured according to their function for students' learning processes and are accordingly located on three layers (see Fig. 2). The fundamental dimension is *adaptation* because adapting teaching to students' individual learning needs is a crucial basis and requirement for all other dimensions. The dimensions that directly support the learning process, and are therefore closely linked to learning outcomes, are located at the model's center: (1) *selection and implementation of content, learning objectives and (subject -specific) methods*, (2) *cognitive activation*, (3) *support* for *consolidation*, and *(4) assessment and feedback*. They are arranged to map the support of a prototypical learning process moving from providing content that is subject-appropriate, structured, aligned with the learning objective, and recognizably relevant to students, to encouraging deep engagement with content and supporting its consolidation, to assessment and associated feedback on students' individual learning performance. In between these two layers sit three dimensions which only support the learning process indirectly, their effects being mediated by the four central dimensions: *classroom management* (i.e. preventing or intervening in case of disturbances in order to maximize learning time); *social support*, (i.e., ensuring positive relationships between teachers and students and between students); and *support for self-responsibility of learning* (i.e., scaffolding the extent to which students

[2] Student-level residuals from regression models have also been used to identify discrepancies between teacher judgment and student achievement. This approach is common in teacher expectations research (Hollenstein, 2020) and has its own shortcomings compared to the use of absolute differences (see Bergold & Steinmayr, 2023).

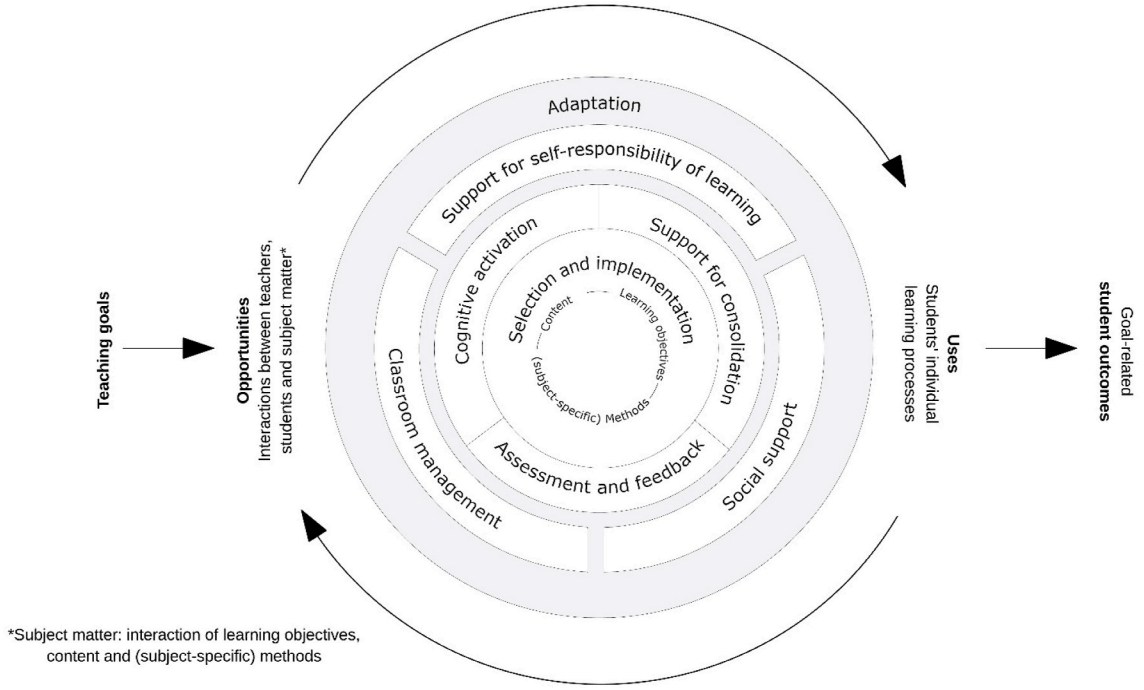


Fig. 2. *The MAIN-TEACH Model*
Note. MAIN-TEACH Model, Version 2.0 (by Praetorius et al., 2023).

receive opportunities for self-regulated/self-responsibility for their learning; for details see Charalambous & Praetorius, 2020; Praetorius et al., 2023).

Table 1 organizes various aspects of teaching quality found in the literature on judgment accuracy into the dimensions of the MAIN-TEACH model. Researchers found statistically significant effects of teacher judgment accuracy for aspects which fit in the *adaptation* or *cognitive activation* dimensions. Some studies reported interaction effects between teacher accuracy and teaching quality on student outcomes for aspects of the dimensions *adaptation*, *selection and implementation of content, learning objectives and (subject -specific) methods,* and *assessment and feedback*. There are also initial indications that *cognitive activation* mediates the effect of judgment accuracy on achievement. By contrast, judgment accuracy seems less relevant to *classroom management*. The review of previously reported effects also reveals that, to date, studies have only examined the relationships between accuracy, teaching quality, and achievement at the class/teacher-level and over a period of no more than one school year.

These findings do not suggest obvious mechanisms by which teaching quality might play an important role in how accuracy affects achievement. Statistically significant results have been reported for interaction and mediation effects and for the impact of accuracy on specific dimensions of teaching quality, which could signify that judgment accuracy acts as a mediator. Specifically, teacher accuracy was found to be associated with those teaching quality dimensions that the MAIN-TEACH model considers to be directly related to student learning. However, the results vary by study, subject, and method used to measure accuracy and teaching quality, with no discernible pattern.

Given that the existing empirical evidence is so inconclusive, statistical decisions about how to model the relationships between judgment accuracy, teaching quality, and achievement should be based on robust reasoning. We believe that there are three reasons why mediation is the most rational choice of mechanism for modeling these relationships: First, the importance of teacher judgment accuracy is often justified by implicitly or explicitly citing the causal sequence that teacher accuracy affects teaching (quality), which then affects student achievement, even though the empirical evidence to support this assertion is scant. Second, the small effects of judgment accuracy on achievement previously found suggest that intermediate variables may play a role. Third, the mediation hypothesis can be based on established theoretical models or frameworks, particularly opportunity-use models for teaching effectiveness (Brühwiler & Blatchford, 2011; Vieluf, Praetorius, Rakoczy, Kleinknecht, & Pietsch, 2020). Opportunity-use-models provide an overview of elements that influence the efficacy of student learning in class. Teacher professional competence affects teaching quality (Brühwiler & Blatchford, 2011; Helmke, 2014), which in turn influences student outcomes (Fauth et al., 2019; Kunter et al., 2013). Since teacher accuracy is considered an important aspect of professional competence (Meissel et al., 2022; Ready & Wright, 2011), reflecting teachers' knowledge of their students (Hill & Chin, 2018), these models suggest mediation through teaching quality (Thiede et al., 2018). This notion is also supported by models of teaching quality that emphasize the importance of teachers' knowledge of their students' achievement in dimensions closely related to students learning processes and achievement development. Based on the MAIN-TEACH model, this applies to the dimension *adaptation* (see also Hardy, Decristan, & Klieme, 2019) and the four dimensions at the center of the model. Specifically, teachers need to consider student achievement levels when selecting content- and subject-specific teaching methods that build on existing student knowledge and use appropriate language and examples (dimension *selection and implementation of content, learning objectives and (subject -specific) methods*). Teachers' knowledge of their students enables them to pose questions that foster deep thinking (Pielmeier et al., 2018), provide them with challenging tasks that match their achievement level (Anders et al., 2010; Brunner et al., 2013), and prepare tasks that help them to consolidate newly acquired skills (dimensions *cognitive activation and support for consolidation*). Lastly, accuracy seems important for assessment, by enabling, among others, the selection of appropriate assessment tasks to verify student understanding and provide meaningful feedback, which should in turn contributes to student learning (dimension *assessment and feedback*; Hill & Chin, 2018; Pielmeier et al., 2018; Thiede et al., 2018). However, judgment accuracy regarding *student achievement* is not thought to play a role for dimensions indirectly related to learning processes: classroom management, social support, or

Table 1
MAIN-TEACH dimensions of teaching quality identified in studies that examine effects of teacher judgment accuracy.

Teaching quality dimension	Investigated teaching quality aspect	Domain	Grade	Role of teaching quality in analyses	Significant results			Citation
					Effect (path in Fig. 1)	Yes	No	
Adaptation	Individualized supportive contact (OR)	Mathematics	5–6	TJA × TQ → ACH	Interaction (1C, e)	R		Helmke and Schrader (1987)
	Individualisation (TR)	Mathematics	5–6	TJA × TQ → ACH	Interaction (1C, e[a)]		H, R	Karing et al. (2011)
	Individualisation (TR)	Reading	5–6	TJA × TQ → ACH	Interaction (1C, e[a)]	H	R	Karing et al. (2011)
	Differentiation (SR)	Mathematics	8	TJA → TQ	Predictor (1B, a)	L, P	R	Westphal, Gronostaj, Vock, Emmrich, and Harych (2016)
	Differentiation (SR)	Reading	8	TJA → TQ	Predictor (1B, a)		L, P, R	Westphal et al. (2016)
Classroom management	Classroom management (OR)	Reading	5, 6, 7	TJA × TQ → ACH	Interaction (1C, e)		R	Behrmann and Souvignier (2013)
	Classroom management (OR)	Reading strategies	5, 6, 7	TJA × TQ → ACH	Interaction (1C, e)		R	Behrmann and Souvignier (2013)
Selection and implementation of content, learning objectives and (subject-specific) methods	Structuring cues (OR)	Mathematics	5–6	TJA × TQ → ACH	Interaction (1C, e)	R		Helmke and Schrader (1987)
	Structuring cues (TR)	Mathematics	5–6	TJA × TQ → ACH	Interaction (1C, e[a)]		H, R	Karing et al. (2011)
	Structuring cues (TR)	Reading	5–6	TJA × TQ → ACH	Interaction (1C, e[a)]	H	R	Karing et al. (2011)
Cognitive activation	Cognitive activation potential of tasks (PCA)	Mathematics	9–10	TJA → TQ → ACH	Mediation (1B, a*b)	D	R	Anders et al. (2010)
	Teachers' use of student productions (OR-V)	Mathematics	4, 5	TJA → TQ	Predictor (1B, a)	P		Hill and Chin (2018)
	Remediation of student mistakes (OR-V)	Mathematics	4, 5	TJA → TQ	Predictor (1B, a)	P		Hill and Chin (2018)
	Elaborating teacher questions (OR-V)	Mathematics	8	TJA → TQ	Predictor (1B, a[b)]		R	Pielmeier et al. (2018)
Assessment and feedback	Feedback (OR)	Reading	5, 6, 7	TJA × TQ → ACH	Interaction (1C, e)	R		Behrmann and Souvignier (2013)
	Feedback (OR)	Reading strategies	5, 6, 7	TJA × TQ → ACH	Interaction (1C, e)		R	Behrmann and Souvignier (2013)
	Monitoring, evaluation, and feedback (SR)	Mathematics	4, 5	TJA → TQ	Predictor (1B, a)		P	Hill and Chin (2018)
	Dialogic feedback (OR-V)	Science	3–8	TJA × TQ → ACH	Interaction (1C, e)		H	Kuhn (2015)
	Dialogic feedback (OR-V)	Science	3–8	TJA → TQ → ACH	Mediation (1B, a*b)		H	Kuhn (2015)
	Supportive teacher feedback (OR-V)	Mathematics	8	TJA → TQ	Predictor (1B, a[b)]		R	Pielmeier et al. (2018)

Note. Relations between the variables were examined at the class/teacher-level in all 8 reported studies. Studies are represented multiple times where multiple domains/teaching quality aspects/effect paths were examined. Aspects were assigned to the MAIN-TEACH dimensions by this paper's authors. The sub-category *Effect* notes the conceptual models (paths) referred to: Interaction = effect of the interaction between teacher judgment accuracy and teaching quality when predicting achievement, Predictor = effect of teacher judgment accuracy on teaching quality, Mediation = effect of teacher accuracy on achievement through teaching quality. *Yes* means effect/path was statistically significant; *No*, not significant. The letters represent: OR = observer rating; TJA = teachers' judgment accuracy; TQ = teaching quality; ACH = achievement; R = rank component; TR = teacher rating; H = task-specific hit rate; SR = student rating; L = level component; P = percentage of accurately judged student; PCA = joint evaluation of the cognitive activation potential of all classwork over a school year; D = accuracy in judging difficulty levels of tasks; OR-V = video analysis. Studies examining interaction or mediation used a longitudinal design (except for Kuhn, 2015) and focused on short-term effects of teachers' accuracy on student achievement (i.e., over one (school) year at most). Studies examining whether teachers' accuracy predicted teaching quality (see subcolumn *Effect*) had cross-sectional designs except for Hill and Chin (2018) who used longitudinal data. We excluded the study by Brühwiler (2017), which examined only bivariate correlations between teachers' judgment accuracy and the following teaching quality aspects: pressurised teaching, student participation, explaining quality, pupil interest in instruction.

[a] The authors of the cited study assumed that teaching quality moderates the relationship between teachers' judgment accuracy and student achievement, in contrast to the other studies which assumed that teachers' accuracy moderates the effect of teaching quality on achievement.

[b] In this study, the teaching quality aspects of elaborating teacher questions and supportive teacher feedback were predicted by student characteristics, teachers' judgment accuracy as well as the interaction between teachers' accuracy and students' prior achievement. We only report the main effects of teachers' accuracy because they are of primary interest.

support for self-responsibility of learning. These dimensions have been shown to be dependent on other cognitive and motivational aspects of professional competence (e.g., teachers' pedagogical knowledge, self-efficacy; Fauth et al., 2019). However, accurate knowledge of other student characteristics may indeed be relevant for some of these dimensions (e.g., judgment accuracy concerning students' self-regulated learning skills for support for self-responsibility of learning; Karlen, Bäuerlein, & Brunner, 2023).

1.3. Student ratings of teaching quality

Teaching quality can be assessed from different perspectives, including classroom observations by external observers, teacher self-reports and student ratings. Increasing attention is being paid to student perceptions and many studies use student ratings to measure teaching quality as they have been shown to be a valid, reliable, cost- and time-efficient approach (Fauth, Decristan, Rieser, Klieme, & Büttner, 2014; Lucksnat et al., 2024; Senden, Nilsen, & Teig, 2023).

Compared to other approaches (e.g., observer ratings), student ratings are considered to have the following advantages: They provide a more general, long-term view of teaching because they are based on students' daily classroom experiences (Clausen, 2002; Praetorius, Pauli, Reusser, Rakoczy, & Klieme, 2014) and inform about teaching quality at both the class (students' shared perception) and the student-level. At the class level, psychometric properties of the class average perceptions of teaching quality are comparable to those of observation measures (Clausen, 2002; Maulana & Helms-Lorenz, 2016). At the student-level, student ratings provide unique insight into the individual student's learning experience within class (Göllner et al., 2021). Finally, student ratings have been shown to predict student outcomes and, for some outcomes (e.g., student engagement), to be even more predictive than observer ratings (Clausen, 2002; Maulana & Helms-Lorenz, 2016).

1.4. The study

The purpose of this longitudinal study was to investigate whether selected teaching quality aspects of the lower and upper layer of the MAIN-TEACH model (see Fig. 2), measured using student ratings mediated the effect of teacher judgment accuracy on student achievement in German language. The study looked at the impact of teacher accuracy on achievement over a three-year interval, modelling accuracy at both student and class/teacher-level. The following research questions (RQ) and hypotheses (H) were addressed.

RQ1. Does teacher judgement accuracy have an impact on student achievement in German language?

Teacher accuracy at class/teacher-level (H1a) and at student-level (H1b) should have a positive effect on achievement.

RQ2. Are the effects of teacher judgment accuracy on achievement mediated by teaching quality?

Teacher accuracy at class/teacher-level should have an indirect effect on student achievement via *student perception* (i.e., student-level measure; H2a) and via *students' shared perception* (i.e., class/teacher-level; H2b) of the four selected aspects of teaching quality dimensions. Furthermore, we expect teacher accuracy at student-level to have an indirect effect on achievement via *student perception* of the teaching quality aspects (H2c).

2. Methods

2.1. Design and sample

We used data from a research project with a longitudinal quasi-experimental design that studied the development of German language and mathematics achievement in 18 public lower secondary schools in the German-speaking Swiss Canton of Zurich (for project details, see Hochweber et al., 2020). Data were collected at four time points: T1: the beginning of 7th grade (2016/17 school year); T2: toward the end of 7th grade; T3: end of 8th grade (2017/18 school year); and T4: end of 9th grade (2018/19 school year). For each subject, students completed computerized curriculum-based tests at each time point and rated teaching quality at T1, T3, and T4. Online questionnaires were used to collect all other data, including student and teacher demographics, student ratings of teaching quality, and teacher ratings of student test performance.

Overall, 1687 students participated at least once while the project was running. For this study, we used data from T1 and T4 for German language only and limited our analyses to students who were taught by the same teachers in the same classes for the duration of the study (time-stable sample). We also excluded: 1. Students or teachers for whom information about teacher or class assignment was missing; 2. Teachers or classes with $\leqq 5$ time-stable students; 3. Teachers or classes with $\leqq 60\%$ time-stable student body. The resulting sample comprised 646 students (53% female; average age at T4 = 15.9, $SD = 0.5$) from 42 German language classes and their 35 teachers (60% female; age: 9% 25–29 years old, 42% 30–39 years old, 33% 40–49 years old, 9% 50–59 years old; 6% 60 years or older; teaching experience: 30% 1–5 years, 36% 6–15 years, 18% 16–25 years, and 15% up to 25 years; 5.7% missing data).

Student test data was collected at T1 in September 2016 and at T4 (denoted as T2 for this study) in May/June 2019. Teacher judgments of student achievement were collected at T1 over a period of four weeks in December/January 2017. Students rated teaching quality at T1 from November to February 2017. Since all classes were newly formed at the beginning of 7th grade, teachers and students had to be given time to become familiar with each other.

2.2. Measures

2.2.1. Student rating of teaching quality

Scales for *individual support*, *comprehensibility and clarity*, *cognitive activation*, and *consolidation*, rated on a four-point Likert scale (1 = totally disagree, 4 = totally agree), were selected from the student questionnaire and used to measure a subset of four MAIN-TEACH dimensions: *adaptation*, *selection and implementation of content, learning objectives and (subject -specific) methods*, *cognitive activation*, *support for consolidation*. Only those scales were selected for which a mediating role could be assumed based on our theoretical considerations (cf. Section 1.3). The scales were modified for this study because the project's item set needed to be updated to reflect current definitions of teaching quality aspects and inconsistent forms of address in their wording that can have serious consequences when measuring teaching quality (Jaekel, Wagner, Trautwein, & Göllner, 2022). Therefore, we only included items that were both appropriate according to current definitions of the studied aspects and made it clear to the responding student that they were to consider themselves as the target of the teacher's behavior (e.g., "When I don't understand something in German class, I get tips from my teacher that really help me."). Consistent implementation of the second criterion was challenging because for some items the addressee was not clearly defined (responding student vs. whole class; e.g., "It is important to my teacher that claims are well-founded."). We opted for items that the responding students were likely to rate based on their own experiences and interactions with their teacher. The resulting scales were shorter and more content-selective than the originals. Table 2 shows the scales, items, the corresponding dimensions of teaching quality, and scale reliabilities.

To probe the scales' psychometric properties, a multidimensional confirmatory factor analysis (CFA) where each teaching quality aspect/scale was represented by one factor was computed using the lavaan package in R (v0.6-5; Rosseel, 2012; v3.6.0; R Core Team, 2019). Because of small data nonnormality[3] we used maximum likelihood (ML) estimation with robust standard errors and full information ML (FIML) to deal with missing data (Lai, 2018). Cluster-robust standard errors were used to account for the hierarchical data structure (Level 1: students; Level 2: classes; Muthén & Satorra, 1995). The model's goodness of fit was acceptable ($\chi^2 = 79.258$, $df = 38$, $p < .001$; RMSEA = 0.043, p-close of RMSEA = 0.82, SRMR = 0.031, CFI = 0.973, TLI = 0.961; Bühner, 2011; Hu & Bentler, 1999). Standardized factor loadings ranged

[3] We tested for nonnormality using MVN (v5.9; Korkmaz, Goksuluk, & Zararsiz, 2014) following Lai (2018). The means of the univariate kurtosis and skewness of the eleven items were 0.33 and −0.63, respectively. According to Lai (2018), these values and especially the kurtosis value, which is particularly important for covariance structure analysis, indicate small nonnormality.

Table 2
Scales and reliability values for teaching quality aspects.

Dimension	Scale and Items	α	ICC (1)/ ICC(2)
Adaptation	Individual support	0.76	0.17/ 0.74
	My teacher gives me the opportunity to learn at my own pace.		
	When I don't understand something in German class, I get tips from my teacher that really help me.		
	If I need help, I get it from my German teacher.		
Cognitive activation	Cognitive activation	0.36	0.06/ 0.48
	My teacher gives us tasks where I have to think thoroughly.		
	It is important to my teacher that claims are also well substantiated		
Selection and implementation of content, learning objectives and (subject-specific) methods	Comprehensibility and clarity	0.74	0.21/ 0.79
	My teacher is good at explaining.		
	My teacher gives good examples so that I understand the material better.		
	My German teacher expresses herself/himself clearly and comprehensibly.		
Support for consolidation	Consolidation	0.68	0.07/ 0.53
	In German class (spelling and grammar), I can practice and repeat what I have learned until it sinks in.		
	When we practice in German class (spelling and grammar), I can apply what I learn to other things (e.g., my own text products).		
	In German lessons (spelling and grammar) I have enough time to practice until I can do something new well.		

Note. α = Cronbach's alpha; ICC(1) = intraclass correlation coefficient indicating the proportion of variance attributed to the class/teacher-level; ICC(2) = intraclass correlation coefficient representing the reliability of the aggregated student perceptions at the class/teacher-level.

from 0.36 to 0.78 ($M = 0.65$)[4] and were statistically significant ($p < 0.05$). The between-factor correlations were all positive and statistically significant, ranging from $r = 0.65$ to 0.79, indicating that the factors were psychometrically separable.

2.2.2. Student achievement in German language

Student achievement in German language was assessed using standardized tests covering three domains of the common curriculum for German-speaking Switzerland: reading comprehension, listening comprehension, and language in focus. Each domain was assessed with 25 (at T1) and 13–15 (at T2) dichotomously scored items. Language in focus is a rather complex construct capturing covering several aspects. The items used in this study captured primarily the aspects "grammar terms" and "spelling rules".

In the canton of Zurich, lower secondary school students are assigned to one of three achievement-based streams. To account for the resulting large differences in achievement, a multi-matrix test booklet design was used. Three test booklets were created, one for each level, for each domain at both time points. The booklets varied in their average item difficulty. All test booklets contained a subset of identical items (anchor-items) to ensure that test performance was comparable across achievement levels and time points.

Longitudinal scaling was conducted using unidimensional Rasch models to capture the single construct "German language achievement" at both time points. All models were estimated using the TAM R package (v3.3-10; Robitzsch, Kiefer, & Wu, 2019). Measurement invariance was tested with differential item functioning (DIF) analysis using the sirt R package (v3.9-4; Robitzsch, 2020). At both time points, the weighted mean square statistic indicated acceptable fit to the Rasch model for all items ($0.70 \leqq \text{WMNSQ} \leqq 1.30$). Moderate to large DIF, as defined by the Educational Testing Service classification system (Monahan, McHorney, Stump, & Perkins, 2007), was found between T1 and T2 in five of 29 common items. These were not used as anchor items.

To establish a common scale at T1 and T2, we used the fixed parameters calibration longitudinal linking method. Weighted-likelihood estimates (WLE) were obtained for both time points and used in the subsequent analyses. WLE reliability at T1 and T2 was 0.86 and 0.89, respectively (EAP/PV reliability: T1 = 0.80; T2 = 0.81).

2.2.3. Teacher judgments

Teacher judgments of student test performance in reading comprehension, listening comprehension, and language in focus were assessed by three items with a 10-point Likert response scale. For reading comprehension, for example, the following instruction was given: "For each student, please tick the box that indicates how well, in your opinion, he or she has performed relative to all other students in the reading comprehension part of the test given at the beginning of seventh grade in Canton Zurich." The rating scale labels were "0–10%, in the lowest 10% of students" through to "90–100%, in the highest 10% students". Our goal was to have teachers focus on individual students and not rely on in-class comparisons, as research has provided evidence of reference group effects (i.e., effects of the class context such as class-average achievement) on teacher judgments and their accuracy (Bergold, Weidinger, & Steinmayr, 2022; Trautwein & Baeriswyl, 2007). Judgment scores for each of the three domains of the test were averaged for each student (and rounded to the nearest whole number) to obtain the teacher's overall judgment score for the student's German language achievement. This should be valid given the high correlations between the domains ($0.81 \leqq r \leqq 0.88$).

2.2.4. Control variables

We controlled for student achievement at T1 to account for initial differences between students. *Socioeconomic status* (SES) was measured by using a generalized partial credit model (GPCM; Muraki, 1992) to create an index based on variables from the student questionnaire: parents' highest educational attainment, number of books at home, and cultural possessions at home (e.g., books on art; for details see Hochweber et al., 2020). *Student gender* was included as a dummy variable (1 = female). Finally, the project providing our data used a quasi-experimental design in which schools were assigned to treatment or control conditions. For this study, we primarily used data from T1 (student ratings of teaching quality, teacher judgments) and were not interested in effects of the treatment, which only targeted variables that were not part of our study. Therefore, we controlled for attending a *treatment* school using a dummy variable, "treatment" (1 = treatment group).

[4] One item from the cognitive activation scale had a loading clearly below usual standards (value of 0.36; ≦ 0.60). We retained this item because its specific content seemed essential to the construct of cognitive activation. An item from the comprehensibility and clarity scale with similarly low loading was excluded.

2.3. Analyses

2.3.1. Teachers' judgment accuracy

Teacher accuracy at class/teacher-level (rank component) was operationalized using Pearson's correlation, calculated for each class/teacher[5] between student test performance in German language (i.e., WLEs) at T1 and teachers' overall judgments of student German language ability (i.e., the mean of the three judgments). At student-level, teacher accuracy was operationalized using difference scores. First, each student's test score was ranked to one of the following percentiles based on all students' test performance: 0–10th, 0–20th, 20–30th, 30–40th, 40–50th, 50–60th, 60–70th, 70–80th, 80–90th, 90–100th, and labelled 1 to 10 to correspond to the teacher's rating scale. Then the difference scores were calculated by subtracting each student's test performance from their teacher's overall judgment of their performance. Students were classified as under-/overestimated based on their difference scores ($M = 0.38$; $SD = 2.51$). Following Urhahne (2015), half the standard deviation of the difference score was used as the cut-off value (i.e., $2.51/2 \approx 1.25$; underestimated students: difference score < -1.25; overestimated students: difference score >1.25). Finally, two dummy variables were formed for the underestimated (n = 129) and overestimated (n = 166) students, with the correctly estimated students (n = 267) as reference.

2.3.2. Variables and statistical models

We created a series of multilevel regression models (MLM) with students (level 1; L1) nested within classes/teachers (level 2; L2), following a manifest variable approach, to test our hypotheses.

To examine the effect of teacher accuracy on student achievement we specified a MLM with T2 achievement as the L1 outcome variable (see H1a, H1b). Teacher accuracy at L1 (i.e., dummy variables reflecting under- and overestimation) and L2 (i.e., rank component) and the control variables (L1: T1 achievement, SES, gender; L2: treatment) were entered as predictor variables.

To test H2a and H2b, we conducted cross-and unique cluster-level mediation analyses (Pituch & Stapleton, 2012), where the predictor of interest was measured at L2 (rank component), while the mediator (teaching quality aspect) was measured at both L1 and L2. This type of mediation analysis tests the indirect effect of a L2 predictor on an L1 outcome through an L1 mediator (cross-level mediation) as well as the effect of the L2 predictor on the L1 outcome as mediated by the class/teacher mean of the same mediator (unique cluster-level mediation). To answer H2c, we tested a 1-1-1 mediation in which the predictor and mediator of interest were all measured at L1. The conceptual models of the hypotheses tested are illustrated in Fig. 3.

To estimate mediation paths, we applied Krull and MacKinnon's approach of using a series of univariate MLMs (2001; see also Tofighi & MacKinnon, 2011).[6] We first estimated four MLMs to predict each teaching aspect (mediator) from our predictor variables, then estimated four MLMs to predict the outcome variable from each teaching aspect (mediation) and the predictor variables. Based thereon, point estimates for each indirect effect were calculated following Tofighi and MacKinnon (2011).

Given the small L2 sample size, we opted for multilevel regression modelling with manifest variables rather than multilevel structural equation modelling (MSEM; see Preacher, Zyphur, & Zhang, 2010) with latent variables for our mediation analysis. Using appropriate methods (see below), MLMs tend to have better small sample performance than MSEM (McNeish, 2017b). All models were estimated using restricted maximum likelihood (REML) and the Kenward-Roger correction for fixed effect standard errors following McNeish (2017a; for an application to mediation see Kuhn, Schwenk, Souvignier, & Holling, 2019). The MLMs were run using the R package lme4 (v1.1-21; Bates, Mächler, Bolker, & Walker, 2015), then the correction was implemented using the *summary* function (see lmerTest package; Kuznetsova, Brockhoff, & Christensen, 2017). Cases with missing data for outcome and/or predictor variables were excluded by the software.[7] Between 135 and 172 students and seven classes/teachers were excluded from all MLMs. Missing data were present for all variables, the proportion ranging from 2.6% to 13.0% for variables at L1 and from 2.4% to 14.3% at L2 (see Table 3). For five of 35 teachers, data were not available for at least one variable (T1/T2 achievement, teaching quality ratings). Since the formation of the accuracy measures requires two variables, any missing value led to missing values for accuracy. Therefore, the largest missing data proportion is seen for teachers' accuracy.

To test the indirect effects at both L1 and L2, 95% confidence intervals (CI) were estimated using the distribution of product method by applying the *medci* function from the RMediation package (v1.1.4; Tofighi & MacKinnon, 2011). The null hypothesis that no indirect effect is present is rejected if zero is not included in the CI.

To facilitate the interpretation of the results, variables were centered/standardized before the analyses. The dummy-coded over-/underestimation predictors were centered within class/teacher to capture the pure within-group effect (Enders & Tofighi, 2007). Student achievement at T1 and SES were standardized based on their overall mean and variance in the analysis sample. Achievement at T2 was standardized based on the mean and variance at T1 (see Hochweber & Vieluf, 2018, for a similar approach). The rank component and all mediators (teaching quality aspects at L1 and their class/teacher aggregates at L2) as well as categorical control variables (treatment, gender) were entered using the raw scores. Raw scale scores were used for the mediator variables to enable interpreting the absolute scale scores on these variables. It also allowed us to interpret the L2 effects of the mediators as their unique impact on the outcome over and above their effect at L1 (Pituch & Stapleton, 2012).

3. Results

3.1. Descriptive statistics

Table 3 displays basic descriptive statistics and correlations for each analysis variable. In line with previous findings, on average, the correlations between teacher judgments and student achievement (i.e., the rank component) are relatively large ($M = 0.59$, $SD = 0.31$; correlations varied from 0.07 to 0.89 between classes/teachers). However, the difference scores ($M = 0.38$; $SD = 2.51$, cf. Section 2.3) that resulted when student test performance was subtracted from teachers' judgments indicate a tendency towards overestimation at the student-level (i.e., $M > 0$).

[5] Six teachers taught two classes, the rank component was calculated for them at class level and not at teacher level as for the others who taught only one class.

[6] Because of the very unsatisfactory reliability of the cognitive activation scale (see Table 2), we estimated all models for both the scale and each of the two items used to form the scales. The results of the models with the scale and the individual items did not differ. To ensure comparability with the other scales, the results are therefore presented using the scale.

[7] Leading methods for dealing with missing data (e.g., FIML, multiple imputation [MI]) all have nontrivial shortcomings for small samples, while best-practice recommendations for multilevel data are not well established (McNeish, 2017a). We considered MI for our study, but small samples may lead to insufficient accuracy of the imputation model, as MI relies on the observed values and regression analyses to impute missing data. As a result, the model estimates may be biased (McNeish, 2017a). Furthermore, MI for *multilevel* data with small samples and missing data on independent and outcome variables at different levels is still an area of ongoing methodological research and requires further investigation.

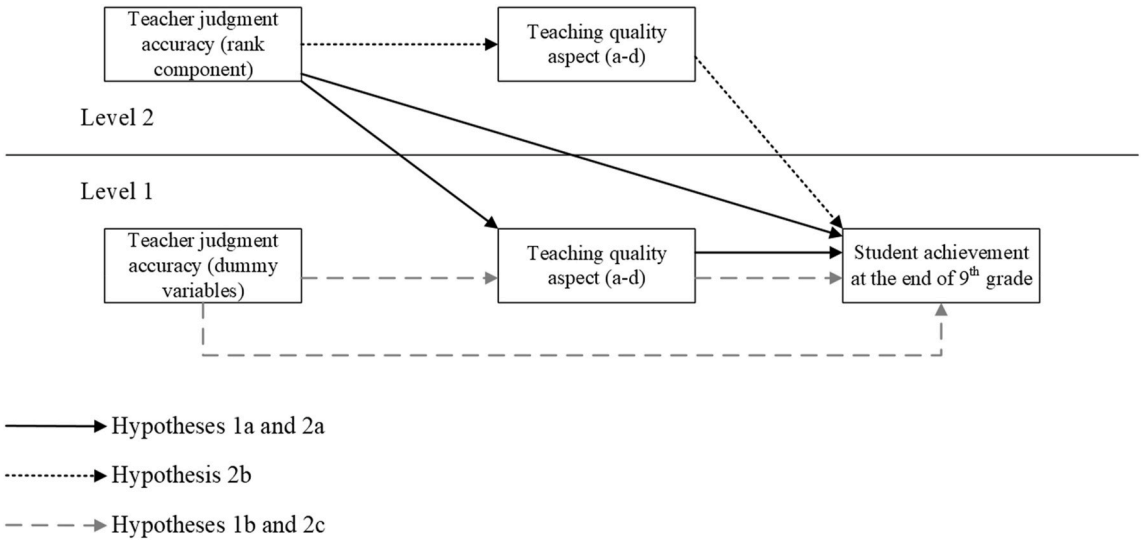


Fig. 3. Overview of Hypotheses and Illustration of Tested Conceptual Mediation Models Tested
Note. Separate mediation models were tested for each teaching quality aspect (a–d), (a) individual support, (b) cognitive activation, (c) comprehensibility and clarity, and (d) and consolidation. The dummy variables representing teacher judgment accuracy at Level 1 reflect overestimated and underestimated students, with accurately judged students treated as the reference group. The control variables (students' achievement at T1, SES of students' families, gender, treatment) included in all models are not depicted for the sake of clarity.

Table 3
Study variable descriptive statistics and correlations.

Variables	*M*	*SD*	*% Missing*	1	2	3	4	5	6	7	8	9	10
Student-level (n = 646)													
1. L1_D1 TJA: overestimation	0.30	0.46	13.00										
2. L1_D2 TJA: underestimation	0.23	0.42	13.00	−0.35***									
3. Individual support	3.13	0.62	10.06	0.08	−0.13**								
4. Cognitive activation	3.06	0.54	10.06	0.06	−0.10*	0.41***							
5. Comprehensibility and clarity	3.44	0.54	9.13	0.01	−0.09*	0.55***	0.44***						
6. Consolidation	3.05	0.55	10.37	−0.05	−0.05	0.49***	0.40***	0.48***					
7. Achievement T1 (WLE)	0.00	1.00	7.43	−0.46***	0.49***	−0.11**	−0.04	0.00	0.08				
8. Achievement T2 (WLE)	1.37	1.43	8.56	−0.26***	0.23***	−0.08	0.01	0.02	0.09*	0.69***			
9. SES	0.00	1.00	2.63	−0.12**	0.13**	−0.08*	−0.01	−0.01	0.05	0.39***	0.32***		
10. Gender: female	0.53	0.50	0.00	−0.03	0.04	−0.05	−0.09*	−0.06	−0.04	0.07	0.09*	0.09*	
11. Treatment: treatment group	0.55	0.50	0.00	0.18***	−0.13**	0.10*	0.14***	0.15***	0.11**	−0.12**	0.01	−0.05	0.00
Class/Teacher-level (n = 42)													
1. TJA L2 (rank component)	0.59	0.31	14.29										
2. Individual support	3.12	0.30	2.38	0.18									
3. Cognitive activation	3.05	0.21	2.38	−0.03	0.62***								
4. Comprehensibility and clarity	3.42	0.28	2.38	−0.05	0.86***	0.66***							
5. Consolidation	3.03	0.22	2.38	−0.13	0.71***	0.46**	0.71***						

Note. Student-level variables no. 7 to 9 are depicted after standardization. Teacher's judgment accuracy (TJA) at class/teacher-level (no. 1) was calculated by correlating teacher judgments and student achievement at T1 per class/teacher using Pearson's correlation. The correlations were Fisher-Z transformed, then averaged, and finally transformed back into a correlation coefficient, resulting in the mean reported in the table. Classroom level variables no. 2 to 5 were formed by aggregating student ratings of teaching quality to the class/teacher-level. L1_D1 TJA = teachers' judgment accuracy at Level 1 (student-level), dummy variable representing overestimated students; L1_D2 TJA = teachers' judgment accuracy at Level 1 (student-level), dummy variable representing underestimated students; T1 = first measurement point (beginning of the 7th grade); T2 = second measurement point (end of the 9th grade); WLE = weighted-likelihood estimate; SES = socioeconomic status of students' families; L2 TJA = teachers' judgment accuracy at Level 2 (class/teacher-level).
*$p < 0.05$, **$p < 0.01$, ***$p < 0.001$.

3.2. Effects of teachers' accuracy on student achievement

RQ1 aimed to investigate whether teachers' accuracy, measured at L1 (student-level) and L2 (class/teacher-level), has an impact on student achievement. The results in Table 5 (see M0) indicate that, contrary to H1a, teachers' accuracy at L2 (rank component) did not have a statistically significant positive effect on student achievement. However, partly consistent with H1b, students whose achievement was accurately judged at the beginning of 7th grade had a higher achievement at the end of 9th grade than those whose achievement was underestimated by their teacher ($B = -0.507$, $p < 0.001$). In contrast, students whose achievement was overestimated did not differ in their achievement development from students who were accurately judged.

3.3. Teaching quality as mediator

RQ2 focused on exploring the mediating role of teaching quality (see Fig. 3 for an illustration). To estimate the mediation paths, we used a series of MLM (see Variables and statistical models Section). In a first step, we predicted each mediator (teaching quality aspects) from the predictors (teachers' accuracy at L2 and L1) and the control variables. The results are presented in Table 4 (M1 to M4). As can be seen from M1 to M3, teachers' accuracy, measured at L2 and L1, did not significantly predict *individual* support, *cognitive activation*, or *comprehensibility and*

Table 4
Predicted mediation variables (aspects of teaching quality).

	Individual support (M1)	Cognitive activation (M2)	Comprehensibility and clarity (M3)	Consolidation (M4)
	B (*SE*)	*B* (*SE*)	*B* (*SE*)	*B* (*SE*)
Intercept	2.945 (0.167)***	3.013 (0.097)***	3.398 (0.152)***	3.093 (0.120)***
Student covariates (Level 1)				
L1_D1 TJA: overestimation[a]	0.087 (0.071)	0.041 (0.065)	0.007 (0.061)	−0.064 (0.066)
L1_D2 TJA: underestimation[a]	−0.090 (0.073)	−0.054 (0.067)	−0.080 (0.062)	−0.138 (0.068)*
Achievement T1	−0.033 (0.037)	−0.009 (0.031)	0.015 (0.031)	0.031 (0.033)
SES	−0.001 (0.029)	0.007 (0.026)	<0.001 (0.025)	0.033 (0.027)
Gender: female	−0.058 (0.052)	−0.099 (0.047)*	−0.054 (0.044)	−0.091 (0.048)
Treatment: treatment group	0.128 (0.103)	0.158 (0.059)*	0.141 (0.094)	0.120 (0.074)
Class/teacher covariates (Level 2)				
L2 TJA (rank component)	0.291 (0.24)	0.052 (0.138)	−0.014 (0.221)	−0.124 (0.172)
R^2 (MVP)	0.035	0.033	0.020	0.031
N	506	506	511	505

Note. L1_D1 TJA: = teachers' judgment accuracy at Level 1 (student-level), dummy variable for grouping overestimated students; L1_D2 TJA = teachers' judgment accuracy at Level 1 (student-level), dummy variable grouping for underestimated students; T1 = first measurement point (beginning of the 7th grade); SES = socioeconomic status of students' families; L2 TJA = teachers' judgment accuracy at Level 2 (class/teacher-level); R^2(MVP) = multilevel variance partitioning (LaHuis, Hartman, Hakoyama, & Clark, 2014).
*$p < 0.05$, **$p < 0.01$, ***$p < 0.001$.
[a] Accurately judged students were specified as the reference group.

clarity. However, a statistically significant effect of teachers' accuracy at L1 was found for *consolidation* (see M4). Students for whom teacher judgments were accurate viewed *consolidation* – how a teacher reinforced previously taught material – more positively than their underrated counterparts ($B = -0.138$, $p < 0.05$), while the perceptions of overestimated students did not differ from those of accurately rated ones.

In a second step, student achievement at T2 was predicted by the mediator, predictor, and control variables (see M1 to M4 in Table 5). In each model, one teaching quality variable was entered at L2 and L1, respectively, along with the predictor (teacher accuracy at L2 and L1) and control variables. No mediator was significantly associated with achievement at T2. Teacher accuracy at L2 also did not predict achievement. However, in all models, a statistically significant difference in achievement at T2 was found in favor of students who were accurately judged. Again, no such difference was found between overestimated and accurately judged students.

In a final step, the indirect effects of teacher accuracy on student achievement via teaching quality were estimated for all models. Contrary to our hypotheses (H2a - H2c), no indirect effect of teacher accuracy on student achievement was found via any of the four teaching quality aspects (see Table 6). The confidence intervals for all models included zero, both for the indirect effects of teacher accuracy on L2 via the aspects of teaching quality on L1 and L2 (H2a, H2b) and for the indirect effects of teacher accuracy on L1 via the mediators on L1 (H2c). As mentioned above, accurately judged students perceived the *consolidation* more positively than underestimated students. Still, students' perceptions of *consolidation* were not related to their achievement development (cf. Table 5), and no indirect effect for this teaching quality aspect could be demonstrated.

Table 5
Predicted outcome variables (student achievement at T2).

	Achievement T2 (M0)	Achievement T2 (M1)	Achievement T2 (M2)	Achievement T2 (M3)	Achievement T2 (M4)
	B (SE)	*B* (*SE*)	*B* (*SE*)	*B* (*SE*)	*B* (*SE*)
Intercept	1.319 (0.368)**	2.354 (1.262)	−2.093 (2.099)	1.799 (1.637)	1.493(1.830)
Student covariates (Level 1)					
L1_D1 TJA: overestimation[a]	0.197 (0.110)	0.186 (0.113)	0.177 (0.113)	0.190 (0.112)	0.175 (0.113)
L1_D2 TJA: underestimation[a]	−0.507 (0.114)***	−0.478 (0.116)***	−0.476 (0.117)***	−0.458 (0.115)***	−0.478 (0.117)***
Individual support		−0.069 (0.074)			
Cognitive activation			0.012 (0.081)		
Comprehensibility and clarity				0.008 (0.087)	
Consolidation					0.006 (0.079)
Achievement T1	0.967 (0.060)***	0.922 (0.062)***	0.930 (0.062)***	0.920 (0.061)***	0.923 (0.062)***
SES	0.044 (0.045)	0.055 (0.047)	0.062 (0.047)	0.058 (0.046)	0.063 (0.047)
Gender: female	0.065 (0.080)	0.034 (0.082)	0.039 (0.083)	0.028 (0.082)	0.035 (0.083)
Treatment: treatment group	0.094 (0.230)	0.121 (0.243)	−0.114 (0.255)	0.110 (0.253)	0.082 (0.246)
Class/teacher covariates (Level 2)					
L2 TJA (rank component)	−0.212 (0.541)	−0.107 (0.574)	−0.333 (0.537)	−0.271 (0.577)	−0.273 (0.568)
Individual support		−0.278 (0.421)			
Cognitive activation			1.163 (0.703)		
Comprehensibility and clarity				−0.135 (0.480)	
Consolidation					−0.042 (0.594)
R^2 (MVP)	0.403	0.389	0.395	0.378	0.382
N	503	474	475	479	474

Note. T2 = second measurement point (end of the 9th grade); *L1_D1 TJA* = teachers' judgment accuracy at Level 1 (student-level), dummy variable for grouping overestimated students; *L1_D2 TJA* = teachers' judgment accuracy at Level 1 (student-level), dummy variable grouping for underestimated students; T1 = first measurement point (beginning of the 7th grade); SES = socioeconomic status of students' families; L2 TJA = teachers' judgment accuracy at Level 2 (class/teacher-level). R^2(MVP) = multilevel variance partitioning (LaHuis, Hartman, Hakoyama, & Clark, 2014).
*$p < 0.05$, **$p < 0.01$, ***$p < 0.001$.
[a] Accurately judged students were specified as the reference group.

Table 6
Indirect effects of the predictor variables (teachers' judgment accuracy) in the mediation analyses.

Mediation models	*B*	*SE*	95% CIs
2-1-1 cross-level mediation			
L2 TJA (rank component) → L1 Individual support → Achievement T2	−0.020	0.033	[-0.082, 0.021]
L2 TJA (rank component) → L1 Cognitive activation → Achievement T2	0.001	0.012	[-0.018, 0.020]
L2 TJA (rank component) → L1 Comprehensibility and clarity → Achievement T2	<0.001	0.019	[-0.031, 0.031]
L2 TJA (rank component) → L1 Consolidation → Achievement T2	−0.001	0.017	[-0.028, 0.026]
2-2-1 unique cluster-level mediation			
L2 TJA (rank component) → L2 Individual support → Achievement T2	−0.081	0.173	[-0.404, 0.147]
L2 TJA (rank component) → L2 Cognitive activation → Achievement T2	0.060	0.191	[-0.225, 0.398]
L2 TJA (rank component) → L2 Comprehensibility and clarity → Achievement T2	0.002	0.111	[-0.173, 0.180]
L2 TJA (rank component)→ L2 Consolidation → Achievement T2	0.005	0.126	[-0.193, 0.213]
1-1-1 mediation student-level mediation			
L1_D1 TJA: overestimation → L1 Individual support → Achievement T2	−0.006	0.010	[-0.024, 0.006]
L1_D2 TJA: underestimation → L1 Individual support → Achievement T2	0.006	0.010	[-0.006, 0.025]
L1_D1 TJA: overestimation → L1 Cognitive activation → Achievement T2	<0.001	0.006	[-0.009, 0.011]
L1_D2 TJA: underestimation → L1 Cognitive activation → Achievement T2	−0.001	0.007	[-0.012, 0.010]
L1_D1 TJA: overestimation → L1 Comprehensibility and clarity → Achievement T2	<0.001	0.005	[-0.008, 0.009]
L1_D2 TJA: underestimation → L1 Comprehensibility and clarity → Achievement T2	−0.001	0.009	[-0.015, 0.013]
L1_D1 TJA: overestimation → L1 Consolidation → Achievement T2	<0.001	0.007	[-0.012, 0.011]
L1_D2 TJA: underestimation → L1 Consolidation → Achievement T2	−0.001	0.012	[-0.021, 0.019]

Note. L2 TJA = teachers' judgment accuracy at Level 2 (class/teacher-level); *L1 = Level 1 (student-level);* T2 = second measurement point (end of the 9th grade); L2 = Level 2 (class/teacher-level); *L1_D1 TJA* = teachers' judgment accuracy at Level 1 (student-level), dummy variable for grouping overestimated students; *L1_D2 TJA* = teachers' judgment accuracy at Level 1 (student-level), dummy variable grouping for underestimated students; 95% CI = 95% confidence interval estimates for indirect effect (Tofighi & MacKinnon, 2011).

4. Discussion

It is often assumed that teachers with high judgment accuracy can promote better student achievement over time. We investigated whether different aspects of teaching quality mediate the effects of teacher accuracy on German language achievement, supplementing a well-established class/teacher-level accuracy measure, the rank component, with a student-level measure to try to improve the predictive power of judgment accuracy. This also allowed us to examine a variety of mediating pathways. The study was designed to investigate the effects of accuracy over a longer time interval, 7th to 9th grade, than had been considered in previous studies.

While the rank component had no empirical relevance in any model, underestimation of individual students was related to lower achievement at the end of 9th grade and a less favorable view of teaching practices related to *consolidation*, suggesting that including student-level measures of judgment accuracy is helpful. However, none of the mediation pathways investigated were found to be statistically significant. Below, we discuss the benefits of using student-level measures and provide possible explanations for the results.

4.1. Effects of teacher judgment accuracy on student achievement

Our analyses showed no statistically significant effect of teacher accuracy on achievement in German language over a three-year time interval, as measured by the rank component. While this result contradicts our hypothesis, it is consistent with the results of some studies conducted over shorter time intervals (Brühwiler, 2017; Karing et al., 2011; Schrader, 1989). However, our results for student-level accuracy suggest that accurate teacher judgments play a role in long-term achievement development, especially for students who are underestimated by their teachers; accurately judged students showed higher achievement than underestimated students after three years of teaching by the same teacher. This demonstrates the importance of accurate teacher judgments of individual students and highlights that teachers need to be made aware of the possible negative consequences of underestimating their students. This is particularly relevant because teachers tend to underestimate students based on characteristics other than their achievement (e.g., an emotional and behavioral disorder diagnosis or class achievement level; Krämer & Zimmermann, 2021; Ready & Wright, 2011) and may be contributing to some students' relatively poor long-term achievement. Overestimating students, on the other hand, does not appear to confer any achievement (dis)advantage. Potentially, any detrimental effects from judgment inaccuracy were compensated for by other, positive, aspects of these students' learning environment. Some researchers have argued that slightly overestimating achievement may actually be beneficial to students (McElvany et al., 2009; Stang & Urhahne, 2016), as it encourages providing more challenging material without overwhelming students (Förster et al., 2022; but see Bergold & Steinmayr, 2023). While our data is not suitable for exploring this in greater detail, our results suggest that the student-level measure we used provides additional useful information and predictive power compared to the rank component. Finally, these results highlight the need to differentiate between levels (class/teacher- and student-level) when analyzing teacher accuracy.

4.2. Teaching quality as a mediator

Our findings do not confirm the proposed mediating role, linking teacher accuracy to student achievement, for any of the four selected aspects of teaching quality. This holds true for both indicators (at class/teacher- and student-level) and all mediation pathways (see Fig. 3). We found no mediation of effects for the rank component in line with results reported by Anders et al. (2010). While the rank component has been associated with some aspects of teaching quality, either correlatively or in interaction with other predictor variables (Behrmann & Souvignier, 2013; Brühwiler, 2017; Pielmeier et al., 2018), no longitudinal effects on teaching quality have been reported (Urhahne & Wijnia, 2021). Our results show that the same findings hold true even over a longer period, calling into question the relevance of the rank component to teacher behavior (for a similar discussion see Karst et al., 2014).

Our results indicate a connection between teacher accuracy and the *consolidation* aspect of teaching quality specific to the student-level. Accurately judged students perceived teaching more positively than underestimated students when it came to opportunities and sufficient time to practice targeted knowledge/skills. Although our results do not provide conclusive evidence, they suggest that the dimension *support for consolidation*, which has been neglected in research on teacher judgment accuracy, may be worth investigating in future research. They also once again point to the potential for extra insight gained by examining effects at the student-level.

There are several possible explanations for why we found no evidence for mediation. First, although teachers may be able to accurately judge their students, they may have difficulty using this information to adapt their teaching, as studies on data-based decision making (DBDM) have shown. Heritage, Kim, Vendlinski, and Herman (2009) showed that teachers are better able to assess students' levels of understanding from

formative assessment information than to use it when planning subsequent teaching. In their review of DBDM, Hoogland et al. (2016) reported that several kinds of professional knowledge (e.g., content knowledge) are needed to transform student-related information into appropriate teaching actions. When teachers do not have this knowledge, they tend to just follow the textbook without making adaptations.

Second, given the general lack of empirical evidence in support of the proposed causal sequence, the methods currently used to measure teacher accuracy may be inadequate. In fact, the appropriateness of using methods which are conceptually detached from actual teaching practices and which only refer to individual student characteristics, usually at the beginning of a school year, has increasingly been questioned (Glock, Krolak-Schwerdt, Klapproth, & Böhmer, 2013; Praetorius, Koch, Scheunpflug, Zeinz, & Dresel, 2017). While such measures may be useful in some diagnostic situations (e.g., when outlining content to be taught or creating groups of learners with different achievement levels), they are inadequate in others. In a study by Karst, Klug, and Ufer (2017), teachers were asked to recall situations in which they had judged their students' learning level. Some of these situations were planned well in advance. For example, a teacher judged his/her students' achievement level on a particular topic to check that they had sufficient knowledge before moving on to a new teaching phase. However, most situations were planned in the short-term and involved judging achievement based on a task, for example, to check whether new material had been understood, or to provide individual support to students as they worked on the task. Therefore, studies should systematically capture different diagnostic situations, conceptualizing which judgments are relevant to each situation, operationalizing them appropriately, and, in a next step, examining their relations to teaching and achievement. This would likely require a shift from currently favored purely quantitative study designs to multimethodological approaches. Also, there is increasing emphasis on viewing judgment accuracy not in isolation, but as part of a broader construct, teacher judgment competence (Heitzmann et al., 2019; Loibl, Leuders, & Dörfler, 2020), covering for example also aspects of content knowledge and other dispositions such as motivation. Doing so may help to identify variables with greater explanatory power than those currently available.

Third, the lack of mediation effects may be attributed to the challenges of measuring teaching quality (Charalambous et al., 2021; Mu, Bayrak, & Ufer, 2022). We assessed selected quality aspects using student ratings with very short scales, which meant that for some scales, reliability was not very satisfactory. This was especially significant for *cognitive activation*, for which low reliability of measures has been documented (e.g., Atlay, Tieben, Hillmert, & Fauth, 2019; Kunter et al., 2008). Another problem, not limited to our study, is that common measures of teaching quality are very general and not tailored to individual students and teacher-student interactions (see Ruelmann, Charalambous, & Praetorius, 2023). We were also unable to include an analysis of the effects of feedback quality because it was not measured during the original research project. This is unfortunate since teacher accuracy has been shown to be important for high quality feedback (Behrmann & Souvignier, 2013; Hill & Chin, 2018).

Finally, teacher accuracy may have a more complex influence on teaching quality and student achievement, necessitating the addition of multiple mediator variables such as student learning motivation and emotion and student learning processes, and/or moderator variables such as teacher motivation and beliefs regarding diagnostics (see Westphal, Zuber, & Vock, 2018). In this context, it would be particularly interesting to determine whether accurately and inaccurately judged students differ on the above-mentioned student variables and how these differences might affect their perceptions of teaching quality and achievement, and whether reasons for teacher inaccuracy (e.g., student characteristics and behaviours) play a role in how the variables of interest are related.

4.3. Limitations and future directions

Due to our limited sample size, we operationalized teacher accuracy as the correlation between teacher judgments and test performance for each class/teacher. This methodology has significant limitations when compared to using multilevel modelling and can lead to imprecise estimates at the class/teacher-level (Kolovou, Naumann, Hochweber, & Praetorius, 2021). Also, reference group effects can influence the size of the resulting correlations. Our student-level accuracy measure was based on difference scores, which also imposes implicit accuracy constraints on the data at different levels (Schönbrodt, Humberg, & Nestler, 2018). Alternative approaches to address this issue should be considered, such as response surface analysis (for an application to teachers' accuracy research, see Förster et al., 2022).

Limitations also arose from the teacher judgment measures used, which hardly correspond to daily assessment situations (Kaiser, Praetorius, Südkamp, & Ufer, 2017). Future research should pay closer attention to judgments teachers make during classroom teaching (e.g., students' understanding of content/methods, learning difficulties/misconceptions), and how these relate to teaching practices such as feedback and individual support (Alonzo & Kim, 2018; Klug, Bruder, Kelava, Spiel, & Schmitz, 2013). This also requires different data collection approaches – ideally a combination of naturalistic and simulation-based designs (Kaiser et al., 2017b).

The study was also limited by having to use data from an available item pool to capture relevant teaching quality aspects. The used multistep approach to select appropriate items resulted in highly selective scales that demonstrated only moderate reliability. This issue was prevalent across all scales, but it was particularly pronounced for cognitive activation. After item selection, this scale was reduced to merely two items and showed markedly low reliability. We cannot rule out that this may have compromised the validity of our mediation analysis results to some extent (Cole & Preacher, 2014). Although it is psychometrically challenging to measure some dimensions, such as *cognitive activation*, researchers should focus on measuring their selected aspects in as holistic and specific a fashion as possible.

Another limitation in measuring teaching quality is relying solely on student ratings. Their validity has been a subject of debate, with concerns about potential biases linked to student characteristics and the influence of teachers' popularity among students. While our study did not control for these factors, research suggests that biases in student ratings are typically relatively small, and these ratings can offer a valid evaluation of teaching quality (Benton & Cashin, 2012; Kunter & Baumert, 2006; Senden et al., 2023). However, it is widely recognized that any single method of evaluating teaching quality has its limitations. Ideally, a multifaceted approach that includes different perspectives is preferred. Additionally, student ratings may not be similarly suitable to capture all teaching quality dimensions. This is particularly true for the dimension *cognitive activation*, which may require nuanced subject-specific knowledge that transcends students' direct experience from participating in classroom teaching (Fauth et al., 2014; Göllner et al., 2021).

Another limitation concerns the operationalization of SES, which primarily reflects educational and cultural aspects, and only to a limited extent economic aspects, as we did not include the occupational status of parents/caregivers, contrary to common practice in many studies (e.g., PISA 2022; OECD, 2023).

Although we used statistical methods appropriate for small sample sizes, it is possible that missing data led to a decrease in statistical power. Strategies for multilevel imputation with small sample sizes are still poorly researched (McNeish, 2017a), so we refrained from applying this technique and instead used listwise deletion. Future studies should optimize sample size planning (see Maxwell, Kelley, & Rausch, 2008) when considering the specifics of the variables being studied (for teaching quality see Zitzmann et al., 2022).

Moreover, our study investigated effects over three years but only

used T1 measurements of judgment accuracy and teaching quality. To examine effects of teacher accuracy more comprehensively, it will be necessary to systematically examine its relationships with teaching quality and achievement over a variety of time spans, from very short periods of a few lessons to many years. Issues about the stability of variables over time must also be considered. For our study, we assumed accuracy and teaching quality were rather stable across the three years, but several factors might contribute to noticeable changes in these variables (e.g., changes in class composition).

Finally, due to the limitations mentioned above, we are unable to draw a conclusion as to whether our results concerning the underestimated students – who showed lower achievement and a less favorable view of support of consolidation –, can be attributed to cumulating effects of teachers teaching their classes over a long period of time. Whether this is the case requires further investigation, focusing on the specific mechanisms involved.

4.4. Conclusions

We examined the effects of teacher accuracy, measured at both the student and class/teacher-level, on achievement in German language achievement in secondary school students over a three-year period, and performed analyses to determine if these effects were mediated along multiple pathways by four aspects of teaching quality. While the inclusion of student-level measures of accuracy seems promising, as it led to the detection of effects on achievement that would otherwise have gone unnoticed, we did not find any of the hypothesized mediation effects. In light of previous studies, one might infer that the theorized relationships between teacher accuracy, teaching quality, and student achievement simply do not exist, but such a conclusion might be premature. Our broadly disappointing results on the impact of teacher accuracy could be due to the prevailing practice in this research area of collecting data on judgments which have little connection to regular daily assessment practices, using class/teacher-level accuracy measures, and evaluating teacher accuracy in isolation from other aspects of teachers' professional knowledge, not as part of broader constructs such as teacher diagnostic competence (Herppich et al., 2018). These practices all make it more difficult for research to uncover the mechanisms by which teacher judgment accuracy relates to teaching and learning.

CRediT authorship contribution statement

Dimitra Kolovou: Conceptualization, Formal analysis, Investigation, Methodology, Project administration, Writing – original draft, Writing – review & editing. **Jan Hochweber:** Conceptualization, Funding acquisition, Methodology, Project administration, Supervision, Writing – review & editing. **Anna-Katharina Praetorius:** Conceptualization, Supervision, Writing – review & editing.

Declaration of competing interest

There are no known conflicts of interest to report.

Data availability

For legal reasons, the data used in this article cannot be shared at this time. The authors may be able to arrange access to the data, but permission from the project commissioner would be required.

References

Alonzo, A. C., & Kim, J. (2018). Affordances of video-based professional development for supporting physics teachers' judgments about evidence of student thinking. *Teaching and Teacher Education, 76*, 283–297. https://doi.org/10.1016/j.tate.2017.12.008

Alp Christ, A., Capon-Sieber, V., Grob, U., & Praetorius, A.-K. (2022). Learning processes and their mediating role between teaching quality and student achievement: A systematic review. *Studies In Educational Evaluation, 75*, Article 101209. https://doi.org/10.1016/j.stueduc.2022.101209

Alvidrez, J., & Weinstein, R. S. (1999). Early teacher perceptions and later student academic achievement. *Journal of Educational Psychology, 91*(4), 731–746. https://doi.org/10.1037/0022-0663.91.4.731

Anders, Y., Kunter, M., Brunner, M., Krauss, S., & Baumert, J. (2010). Diagnostische Fähigkeiten von Mathematiklehrkräften und ihre Auswirkungen auf die Leistungen ihrer Schülerinnen und Schüler [Mathematics teachers' diagnostic skills and their impact on students' achievements]. *Psychologie in Erziehung und Unterricht, 57*, 175–193. https://doi.org/10.2378/peu2010.art13d

Atlay, C., Tieben, N., Hillmert, S., & Fauth, B. (2019). Instructional quality and achievement inequality: How effective is teaching in closing the social achievement gap? *Learning and Instruction, 63*, Article 101211. https://doi.org/10.1016/j.learninstruc.2019.05.008

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software, 67*(1), 1–48. https://doi.org/10.18637/jss.v067.i01

Bates, C., & Nettelbeck, T. (2001). Primary school teachers' judgements of reading achievement. *Educational Psychology, 21*(2), 177–187. https://doi.org/10.1080/01443410020043878

Begeny, J. C., Krouse, H. E., Brown, K. G., & Mann, C. M. (2011). Teacher judgments of students' reading abilities across a continuum of rating methods and achievement measures. *School Psychology Review, 40*(1), 23–38. https://doi.org/10.1080/02796015.2011.12087726

Behrmann, L., & Souvignier, E. (2013). The relation between teachers' diagnostic sensitivity, their instructional activities, and their students' achievement gains in reading. *Zeitschrift für Padagogische Psychologie, 27*(4), 283–293. https://doi.org/10.1024/1010-0652/a000112

Benton, S. L., & Cashin, W. E. (2012). Student ratings of teaching: A summary of research and literature. *IDEA Paper, 50*, 1–24.

Bergold, S., & Steinmayr, R. (2023). Teacher judgments predict developments in adolescents' school performance, motivation, and life satisfaction. *Journal of Educational Psychology, 115*(4), 642–664. https://doi.org/10.1037/edu0000786

Bergold, S., Weidinger, A. F., & Steinmayr, R. (2022). The "big fish" from the teacher's perspective: A closer look at reference group effects on teacher judgments. *Journal of Educational Psychology, 114*(3), 656–680. https://doi.org/10.1037/edu0000559

Berlin, R., & Cohen, J. (2018). Understanding instructional quality through a relational lens. *ZDM Mathematics Education, 50*, 367–379. https://doi.org/10.1007/s11858-018-0940-6

Brühwiler, C. (2017). Diagnostische und didaktische Kompetenz als Kern adaptiver Lehrkompetenz [Diagnostic and didactic competence as the core of adaptive teacher competence]. In A. Südkamp, & A.-K. Praetorius (Eds.), *Diagnostische Kompetenz von Lehrkräften. Theoretische und methodische Weiterentwicklungen* (pp. 123–134). Waxmann.

Brühwiler, C., & Blatchford, P. (2011). Effects of class size and adaptive teaching competency on classroom processes and academic outcome. *Learning and Instruction, 21*(1), 95–108. https://doi.org/10.1016/j.learninstruc.2009.11.004

Brunner, M., Anders, Y., Hachfeld, A., & Krauss, S. (2013). The diagnostic skills of mathematics teachers. 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: Results from the COACTIV project* (1st ed., pp. 229–248). Springer. https://doi.org/10.1007/978-1-4614-5149-5.

Bühner, M. (2011). *Einführung in die Test- und Fragebogenkonstruktion [Introduction to test and questionnaire construction]* (3rd ed.). Pearson Studium.

Charalambous, C. Y., & Litke, E. (2018). Studying instructional quality by using a content-specific lens: The case of the mathematical quality of instruction framework. *ZDM Mathematics Education, 50*, 445–460. https://doi.org/10.1007/s11858-018-0913-9

Charalambous, C. Y., Praetorius, A.-K., Sammons, P., Walkowiak, T., Jentsch, A., & Kyriakides, L. (2021). Working more collaboratively to better understand teaching and its quality: Challenges faced and possible solutions. *Studies In Educational Evaluation, 71*, Article 101092. https://doi.org/10.1016/j.stueduc.2021.101092

Charalambous, C., & Praetorius, A.-K. (2020). Creating a forum for researching teaching and its quality more synergistically. *Studies In Educational Evaluation, 67*, Article 100894. https://doi.org/10.1016/j.stueduc.2020.100894

Charalambous, Y. C., & Praetorius, A.-K. (2022). Synthesizing collaborative reflections on classroom observation frameworks and reflecting on the necessity of synthesized frameworks. *Studies In Educational Evaluation, 75*, Article Article101202. https://doi.org/10.1016/j.stueduc.2022.101202

Clausen, M. (2002). *Unterrichtsqualität: Eine Frage der perspektive? Empirische Analysen zur übereinstimmung, konstukt - und kriteriumsvalidität [Teaching quality: A question of perspective? Empirical analyses of agreement, construct and criterion validity]*. Waxmann.

Cole, D. A., & Preacher, K. J. (2014). Manifest variable path analysis: Potentially serious and misleading consequences due to uncorrected measurement error. *Psychological Methods, 19*(2), 300–315. https://doi.org/10.1037/a0033805

Dimosthenous, A., Kyriakides, L., & Panayiotou, A. (2020). Short- and long-term effects of the home learning environment and teachers on student achievement in mathematics: A longitudinal study. *School Effectiveness and School Improvement, 31*(1), 50–79. https://doi.org/10.1080/09243453.2019.1642212

Dollinger, S. (2013). *Diagnosegenauigkeit von erzieherInnen und LehrerInnen [Judgment accuracy of teachers]*. Springer VS.

Enders, C. K., & Tofighi, D. (2007). Centering predictor variables in cross-sectional multilevel models: A new look at an old issue. *Psychological Methods, 12*(2), 121–138. https://doi.org/10.1037/1082-989X.12.2.121

Fauth, B., Decristan, J., Decker, A., Büttner, G., Klieme, E., Hardy, I., et al. (2019). Teachers' professional competence, teaching quality, and student outcomes in elementary science education. *Teaching and Teacher Education, 86*, Article 102882. https://doi.org/10.1016/j.tate.2019.102882

Fauth, B., Decristan, J., Rieser, S., Klieme, E., & Büttner, G. (2014). Grundschulunterricht aus Schüler-, Lehrer- und Beobachterperspektive: Zusammenhänge und Vorhersage von Lernerfolg [Teaching quality in primary school from the perspective of students, teachers, and external observers]. *Zeitschrift für Padagogische Psychologie, 28*(3), 127–137. https://doi.org/10.1024/1010-0652/a000129

Fischbach, A., Baudson, T. G., Preckel, F., Martin, R., & Brunner, M. (2013). Do teacher judgments of student intelligence predict life outcomes? *Learning and Individual Differences, 27*, 109–119. https://doi.org/10.1016/j.lindif.2013.07.004

Förster, N., Humberg, S., Hebbecker, K., Back, M. D., & Souvignier, E. (2022). Should teachers be accurate or (overly) positive? A competitive test of teacher judgment effects on students' reading progress. *Learning and Instruction, 77*, Article 101519. https://doi.org/10.1016/j.learninstruc.2021.101519

Gabriele, A. J., Joram, E., & Park, K. H. (2016). Elementary mathematics teachers' judgment accuracy and calibration accuracy: Do they predict students' mathematics achievement outcomes? *Learning and Instruction, 45*, 49–60. https://doi.org/10.1016/j.learninstruc.2016.06.008

Glock, S., Krolak-Schwerdt, S., Klapproth, F., & Böhmer, M. (2013). Beyond judgment bias: How students' ethnicity and academic profile consistency influence teachers' tracking judgments. *Social Psychology of Education, 16*, 555–573. https://doi.org/10.1007/s11218-013-9227-5

Göllner, R., Fauth, B., & Wagner, W. (2021). Student ratings of teaching quality dimensions: Empirical findings and future directions. In W. Rolett, H. Bijlsma, & S. Röhl (Eds.), *Student feedback on teaching in schools* (pp. 111–124). Springer. https://doi.org/10.1007/978-3-030-75150-0_7.

Hardy, I., Decristan, J., & Klieme, E. (2019). Adaptive teaching in research on learning and instruction. *Journal for Educational Research Online, 11*(2), 169–191. https://doi.org/10.25656/01:18004

Hayes, A. F. (2009). Beyond Baron and Kenny: Statistical mediation analysis in the new millennium. *Communication Monographs, 76*(4), 408–420.

Heitzmann, N., Seidel, T., Opitz, A., Hetmanek, A., Wecker, C., Fischer, M., … Fischer, F. (2019). Facilitating diagnostic competences in simulations: A conceptual framework and a research agenda for medical and teacher education. *Frontline Learning Research, 7*(4), 1–24. https://doi.org/10.14786/flr.v7i4.384

Helmke, A. (2014). *Unterrichtsqualität und Lehrerprofessionalität.* In *Diagnose, Evaluation und Verbesserung des Unterrichts [Teaching quality and teacher professionalism. Diagnosis, evaluation and improvement of teaching]* (1st ed.) Klett-Kallmeyer.

Helmke, A., & Schrader, F.-W. (1987). Interactional effects of instructional quality and teacher judgment accuracy on achievement. *Teaching and Teacher Education, 3*(2), 91–98. https://doi.org/10.1016/0742-051X(87)90010-2

Heritage, M., Kim, J., Vendlinski, T., & Herman, J. (2009). From evidence to action: A seamless process in formative assessment? *Educational Measurement: Issues and Practice, 28*(3), 24–31. https://doi.org/10.1111/j.1745-3992.2009.00151.x

Herppich, S., Praetorius, A.-K., Förster, N., Glogger-Frey, I., Karst, K., Leutner, D., … Klug, J. (2018). Teachers' assessment competence: Integrating knowledge-, process-, and product-oriented approaches into a competence-oriented conceptual model. *Teaching and Teacher Education, 76*, 181–193. https://doi.org/10.1016/j.tate.2017.12.001

Hill, H. C., & Chin, M. (2018). Connections between teachers' knowledge of students, instruction, and achievement outcomes. *American Educational Research Journal, 55*(5), 1076–1112. https://doi.org/10.3102/0002831218769614

Hinnant, J. B., O'Brien, M., & Ghazarian, S. R. (2009). The longitudinal relations of teacher expectations to achievement in the early school years. *Journal of Educational Psychology, 101*(3), 662–670. https://doi.org/10.1037/a0014306

Hochweber, J., Brühwiler, C., Kolovou, D., Pham, G., Knöpfli, N., & Hochweber, A. C. (2020). *Aktive Lernzeit und Lernerfolg für ALLE. Schlussbericht zur Evaluation des Pilotprojekts* [Active learning time and learning success for ALL. Final report on the evaluation of the pilot project]. Bildungsdirektion Kanton Zürich https://www.zh.ch/content/dam/zhweb/bilder-dokumente/themen/bildung/bildungssystem/studien/Evaluation_ALLE_Schlussbericht_PHSG.pdf.

Hochweber, J., & Vieluf, S. (2018). Gender differences in reading achievement and enjoyment of reading: The role of perceived teaching quality. *The Journal of Educational Research, 111*, 268–283. https://doi.org/10.1080/00220671.2016.1253536

Hoge, R. D., & Coladarci, T. (1989). Teacher-based judgments of academic achievement: A review of literature. *Review of Educational Research, 59*(3), 297–313. https://doi.org/10.3102/00346543059003297

Hollenstein, L. (2020). *Leistungserwartungen: ein Aspekt professioneller Kompetenz von Lehrpersonen? Der Zusammenhang zwischen der Leistungserwartung von Lehrpersonen und den Schülerinnen- und Schülerleistungen im Fach Mathematik in der Primarstufe [Expectations: An Aspect of Teachers' Professional Competence? Relationship between Teachers' Expectations and Students' Mathematics Achievement in Primary School.* Kölner Universitäts Publikations Server. Doctoral dissertation, University of Cologne] http://kups.ub.uni-koeln.de/id/eprint/30038.

Hoogland, I., Schildkamp, K., Van der Kleij, F., Heitink, M., Kippers, W., Veldkamp, B., et al. (2016). Prerequisites for data-based decision making in the classroom: Research evidence and practical illustrations. *Teaching and Teacher Education, 60*, 377–386. https://doi.org/10.1016/j.tate.2016.07.012

Hu, L., & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling: A Multidisciplinary Journal, 6*(1), 1–55. https://doi.org/10.1080/10705519909540118

Jaekel, A.-K., Wagner, W., Trautwein, U., & Göllner, R. (2022). "The teacher motivates us – or me?" – the role of the addressee in student ratings of teacher support. *Contemporary Educational Psychology, 71*, Article 102120. https://doi.org/10.1016/j.cedpsych.2022.102120

Kaiser, J., Praetorius, A.-K., Südkamp, A., & Ufer, S. (2017). Die enge Verwobenheit von diagnostischem und pädagogischem Handeln als Herausforderung bei der Erfassung diagnostischer Kompetenz [The close interconnection of diagnostic and pedagogical practice as a challenge in the measurement of diagnostic competence]. In A. Südkamp, & A.-K. Praetorius (Eds.), *Diagnostische Kompetenz von Lehrkräften. Theoretische und methodische Weiterentwicklungen* (pp. 75–93). Waxmann.

Karing, C., Pfost, M., & Artelt, C. (2011). Hängt die diagnostische kompetenz von Sekundarstufenlehrkräften mit der Entwicklung der Lesekompetenz und der mathematischen kompetenz ihrer schülerinnen und schüler zusammen? [Is there a relationship between lower secondary school teacher judgment accuracy and the development of students' reading and mathematical competence? *Journal for Educational Research Online, 3*(2), 119–147. https://doi.org/10.25656/01:5626

Karlen, Y., Bäuerlein, K., & Brunner, S. (2023). Teachers' assessment of self-regulated learning: Linking professional competences, assessment practices, and judgment accuracy. *Social Psychology of Education*. https://doi.org/10.1007/s11218-023-09845-4

Karst, K., Klug, J., & Ufer, S. (2017). Strukturierung diagnostischer Situationen im inner- und außerunterrichtlichen Handeln von Lehrkräften [Structuring of diagnostic situations in the inner- and extracurricular activities of teachers]. In A. Südkamp, & A.-K. Praetorius (Eds.), *Diagnostische Kompetenz von Lehrkräften. Theoretische und methodische Weiterentwicklungen* (pp. 102–113). Waxmann.

Karst, K., Schoreit, E., & Lipowsky, F. (2014). Diagnostische Kompetenzen von Mathematiklehrern und ihr Vorhersagewert für die Lernentwicklung von Grundschulkindern [Diagnostic competences of mathematics teachers and their predictive value for the learning development of elementary school children]. *Zeitschrift für Padagogische Psychologie, 28*(4), 237–248. https://doi.org/10.1024/1010-0652/a000133

Karst, K., & Bonefeld, M. (2020). Judgment accuracy of preservice teachers regarding student performance: The influence of attention allocation. *Teaching and Teacher Education, 94*, Article 103099. https://doi.org/10.1016/j.tate.2020.103099

Kaufmann, E. (2020). How accurately do teachers judge students? Re-analysis of Hoge and Coladarci (1989) meta-analysis. *Contemporary Educational Psychology, 63*, Article 101902. https://doi.org/10.1016/j.cedpsych.2020.101902

Kaufmann, E. (2022). Lens model studies: Revealing teachers' judgements for teacher education. *Journal of Education for Teaching, 49*(2), 236–251. https://doi.org/10.1080/02607476.2022.2061336

Kempert, S., Schalk, L., & Saalbach, H. (2019). Sprache als werkzeug des lernens: Ein überblick zu den kommunikativen und kognitiven funktionen der Sprache und deren bedeutung für den fachlichen wissenserwerb [language as a tool of learning: An overview of communicative and cognitive functions of language and their role in knowledge acquisition]. *Psychologie in Erziehung und Unterricht, 66*(3), 176–195. https://doi.org/10.2378/PEU2018.art19d

Klieme, E., Schümer, G., & Knoll, S. (2001). Mathematikunterricht in der Sekundarstufe I: "Aufgabenkultur" und Unterrichtsgestaltung [Mathematics teaching in lower secondary schools. "Task culture" and instructional design In: Bundesministerium für Bildung und Forschung (BMBF). In *TIMSS -Impulse für Schule und Unterricht: Forschungsbefunde, Reforminitiativen, Praxisberichte und Video-Dokumente* (pp. 43–57). BMBF. http://docplayer.org/12825280-Timss-impulsefuer-schule-und-unterricht.html.

Klug, J., Bruder, S., Kelava, A., Spiel, C., & Schmitz, B. (2013). Diagnostic competence of teachers: A process model that accounts for diagnosing learning behavior tested by means of a case scenario. *Teaching and Teacher Education, 30*, 38–46. https://doi.org/10.1016/j.tate.2012.10.004

Kolovou, D., Naumann, A., Hochweber, J., & Praetorius, A.-K. (2021). Content-specificity of teachers' judgment accuracy regarding students' academic achievement. *Teaching and Teacher Education, 100*. https://doi.org/10.1016/j.tate.2021.103298. Advance online publication.

Korkmaz, S., Goksuluk, D., & Zararsiz, G. (2014). Mvn: An R package for assessing multivariate normality. *The R Journal, 6*(2), 151–162.

Krämer, S., & Zimmermann, F. (2021). Students with emotional and behavioral disorder and teachers' stereotypes – effects on teacher judgments. *The Journal of Experimental Education, 89*, 1–22. https://doi.org/10.1080/00220973.2021.1934809

Krull, J. L., & MacKinnon, D. P. (2001). Multilevel modeling of individual and group level mediated effects. *Multivariate Behavioral Research, 36*(2), 249–277. https://doi.org/10.1207/S15327906MBR3602_06

Kuhn, M. A. (2015). *Do teacher judgment accuracy and teacher feedback predict student achievement in elementary and middle school science? [Doctoral DissertationUniversity of Northern Iowa]. Dissertations and Theses @ UNI.198*. https://scholarworks.uni.edu/etd/198.

Kuhn, J. T., Schwenk, C., Souvignier, E., & Holling, H. (2019). Arithmetische Kompetenz und Rechenschwäche am Ende der Grundschulzeit. Die Rolle statusdiagnostischer und lernverlaufsbezogener Prädiktoren [Arithmetic skills and mathematical learning difficulties at the end of elementary school. The role of summative and formative predictors]. *Empirische Sonderpädagogik, 11*(2), 95–117. https://doi.org/10.25656/01:17773

Kunter, M., & Baumert, J. (2006). Who is the expert? Construct and criteria validity of student and teacher ratings of instruction. *Learning Environments Research, 9*, 231–251. https://doi.org/10.1007/s10984-006-9015-7

Kunter, M., Klusmann, U., Baumert, J., Richter, D., Voss, T., & Hachfeld, A. (2013). Professional competence of teachers: Effects on instructional quality and student development. *Journal of Educational Psychology, 105*(3), 805–820. https://doi.org/10.1037/a0032583

Kunter, M., Tsai, Y.-M., Klusmann, U., Brunner, M., Krauss, S., & Baumert, J. (2008). Students' and mathematics teachers' perceptions of teacher enthusiasm and instruction. *Learning and Instruction, 18*(5), 468–482. https://doi.org/10.1016/j.learninstruc.2008.06.008

Kuznetsova, A., Brockhoff, P. B., & Christensen, R. H. B. (2017). lmerTest package: Tests in linear mixed effects models. *Journal of Statistical Software, 82*(13), 1–26. https://doi.org/10.18637/jss.v082.i13

Lai, K. (2018). Estimating standardized SEM parameters given nonnormal data and incorrect model: Methods and comparison. *Structural Equation Modeling: A Multidisciplinary Journal, 25*(4), 600–620. https://doi.org/10.1080/10705511.2017.1392248

Lingelbach, H. (1995). *Unterrichtsexpertise von grundschullehrkraften [Teaching Expertise of Elementary School Teachers]*. Dr. Kovac.

Loibl, K., Leuders, T., & Dörfler, T. (2020). A framework for explaining teachers' diagnostic judgements by cognitive modeling (DiaCoM). *Teaching and Teacher Education, 91*, Article 103059. https://doi.org/10.1016/j.tate.2020.103059

Lucksnat, C., Richter, E., Henschel, S., Hoffmann, L., Schipolowski, S., & Richter, D. (2024). Comparing the teaching quality of alternatively certified teachers and traditionally certified teachers: Findings from a large-scale study. *Educational Assessment, Evaluation and Accountability, 2024*. https://doi.org/10.1007/s11092-023-09426-1

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

Maxwell, S. E., Kelley, K., & Rausch, J. R. (2008). Sample size planning for statistical power and accuracy in parameter estimation. *Annual Review of Psychology, 59*, 537–563. https://doi.org/10.1146/annurev.psych.59.103006.093735

McElvany, N., Schroeder, S., Hachfeld, A., Baumert, J., Richter, T., Schnotz, W., … Ullrich, M. (2009). Diagnostische Fähigkeiten von Lehrkräften bei der Einschätzung von Schülerleistungen und Aufgabenschwierigkeiten bei Lernmedien mit instruktionalen Bildern [Teachers' Diagnostic Skills to Judge Student Performance and Task Difficulty When Learning Materials Include Instructional Pictures]. *Zeitschrift für Padagogische Psychologie, 23*(3–4), 223–235. https://doi.org/10.1024/1010-0652.23.34.223

McNeish, D. (2017a). Missing data methods for arbitrary missingness with small samples. *Journal of Applied Statistics, 44*(1), 24–39. https://doi.org/10.1080/02664763.2016.1158246

McNeish, D. (2017b). Multilevel mediation with small samples: A cautionary note on the multilevel structural equation modeling framework. *Structural Equation Modeling: A Multidisciplinary Journal, 24*(4), 609–625. https://doi.org/10.1080/10705511.2017.1280797

Meissel, K., Yao, E. S., & Meyer, F. (2022). Teacher judgment (in) accuracy: Differential relations with student progress in writing. *Contemporary Educational Psychology, 69*, Article 102067. https://doi.org/10.1016/j.cedpsych.2022.102067

Monahan, P. O., McHorney, C. A., Stump, T. E., & Perkins, A. J. (2007). Odds ratio, delta, ETS classification, and standardization measures of DIF magnitude for binary logistic regression. *Journal of Educational and Behavioral Statistics, 32*(1), 92–109. https://doi.org/10.3102/1076998606298035

Mu, J., Bayrak, A., & Ufer, S. (2022). Conceptualizing and measuring instructional quality in mathematics education: A systematic literature review. *Frontiers in Education, 7*, Article 994739. https://doi.org/10.3389/feduc.2022.994739

Muraki, E. (1992). A generalized partial credit model: Application of an EM algorithm. *ETS Research Report Series, 1992*(1), 1–30. https://doi.org/10.1002/j.2333-8504.1992.tb01436.x

Muthén, B. O., & Satorra, A. (1995). Complex sample data in structural equation modeling. *Sociological Methodology, 25*, 267–316. https://doi.org/10.2307/271070

OECD. (2023). *PISA 2022 technical report, PISA*. OECD Publishing. https://www.oecd.org/pisa/data/pisa2022technicalreport/.

O'Rourke, H. P., & MacKinnon, D. P. (2018). Reasons for testing mediation in the absence of an intervention effect: A research imperative in prevention and intervention research. *Journal of Studies on Alcohol and Drugs, 79*(2), 171–181. https://doi.org/10.15288/jsad.2018.79.171

Peng, P., Lin, X., Ünal, Z. E., Lee, K., Namkung, J., Chow, J., et al. (2020). Examining the mutual relations between language and mathematics: A meta-analysis. *Psychological Bulletin, 146*(7), 595–634. https://doi.org/10.1037/bul0000231

Pielmeier, M., Huber, S., & Seidel, T. (2018). Is teacher judgment accuracy of students' characteristics beneficial for verbal teacher-student interactions in classroom? *Teaching and Teacher Education, 76*, 255–266. https://doi.org/10.1016/j.tate.2018.01.002

Pituch, K. A., & Stapleton, L. M. (2012). Distinguishing between cross- and cluster-level mediation processes in the cluster randomized trial. *Sociological Methods & Research, 41*(4), 630–670. https://doi.org/10.1177/0049124112460380

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

Praetorius, A.-K., Charalambous, C., Wemmer-Rogh, W., Gossner, L., Herrmann, C., Ufer, S., et al. (2023). *MAIN-Teach-Modell*. Zenodo. https://doi.org/10.5281/zenodo.8280389

Praetorius, A.-K., & Gräsel, C. (2021). Noch immer auf der Suche nach dem heiligen Gral: Wie generisch oder fachspezifisch sind Dimensionen der Unterrichtsqualität [Still searching for the holy grail: How generic or subject-specific are dimensions of teaching quality]? *Unterrichtswissenschaft, 49*(2), 167–188. https://doi.org/10.1007/s42010-021-00119-6

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

Praetorius, A.-K., Koch, T., Scheunpflug, A., Zeinz, H., & Dresel, M. (2017). Identifying determinants of teachers' judgment (in)accuracy regarding students' school-related motivations using a Bayesian cross-classified multi-level model. *Learning and Instruction, 52*, 148–160. https://doi.org/10.1016/j.learninstruc.2017.06.003

Praetorius, A.-K., Lauermann, F., Klassen, R. M., Dickhäuser, O., Janke, S., & Dresel, M. (2017). Longitudinal relations between teaching-related motivations and student-reported teaching quality. *Teaching and Teacher Education, 65*, 241–254. https://doi.org/10.1016/j.tate.2017.03.023

Praetorius, A.-K., Pauli, C., Reusser, K., Rakoczy, K., & Klieme, E. (2014). One lesson is all you need? Stability of instructional quality across lessons. *Learning and Instruction, 31*, 2–12. https://doi.org/10.1016/j.learninstruc.2013.12.002

Preacher, K. J., Zyphur, M. J., & Zhang, Z. (2010). A general multilevel SEM framework for assessing multilevel mediation. *Psychological Methods, 15*(3), 209–233. https://doi.org/10.1037/a0020141

R Core Team. (2019). *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing. URL https://www.R-project.org/.

Ready, D. D., & Wright, D. L. (2011). Accuracy and inaccuracy in teachers' perceptions of young children's cognitive abilities: The role of child background and classroom context. *American Educational Research Journal, 48*(2), 335–360. https://doi.org/10.3102/0002831210374874

Robitzsch, A.. sirt: Supplementary item response theory models. R package version, 3.9-4. https://CRAN.R-project.org/package=sirt.

Robitzsch, A., Kiefer, T., & Wu, M. (2019). *Tam: Test analysis modules. R package version, 3.3-10*. https://CRAN.R-project.org/package=TAM.

Rosseel, Y. (2012). lavaan: An R package for structural equation modeling. *Journal of Statistical Software, 48*(2), 1–36. URL http://www.jstatsoft.org/v48/i02/.

Rucker, D. D., Preacher, K. J., Tormala, Z. L., & Petty, R. E. (2011). Mediation analysis in social psychology: Current practices and new recommendations. *Social and Personality Psychology Compass, 5*(6), 359–371. https://doi.org/10.1111/j.1751-9004.2011.00355.x

Ruelmann, M., Charalambous, C. Y., & Praetorius, A.-K. (2023). The representation of feedback literature in classroom observation frameworks: An exploratory study. *Educational Assessment, Evaluation and Accountability, 35*, 67–104. https://doi.org/10.1007/s11092-022-09403-0

Schönbrodt, F. D., Humberg, S., & Nestler, S. (2018). Testing similarity effects with dyadic response surface analysis. *European Journal of Personality, 32*(6), 627–641. https://doi.org/10.1002/per.2169

Schrader, F.-W. (1989). *Diagnostische Kompetenzen von Lehrern und ihre Bedeutung Fur die Gestaltung und Effektivität des Unterrichts* [Diagnostic competencies of teachers and their meaning for the design and effectivity of instruction]. *Frankfurt/Main: Peter Lang*.

Senden, B., Nilsen, T., & Teig, N. (2023). The validity of student ratings of teaching quality: Factorial structure, comparability, and the relation to achievement. *Studies In Educational Evaluation, 78*, Article 101274. https://doi.org/10.1016/j.stueduc.2023.101274

Stang, J., & Urhahne, D. (2016). Stabilität, Bezugsnormorientierung und Auswirkungen der Urteilsgenauigkeit [Stability, reference norm orientation, and effects of judgment accuracy]. *Zeitschrift für Padagogische Psychologie, 30*(4), 251–262. https://doi.org/10.1024/1010-0652/a000190

Südkamp, A., Kaiser, J., & Möller, J. (2012). Accuracy of teachers' judgments of students' academic achievement: A meta-analysis. *Journal of Educational Psychology, 104*(3), 743–762. https://doi.org/10.1037/a0027627

Thiede, K. W., Brendefur, J. L., Carney, M. B., Champion, J., Turner, L., Stewart, R., et al. (2018). Improving the accuracy of teachers' judgments of student learning. *Teaching and Teacher Education, 76*, 106–115. https://doi.org/10.1016/j.tate.2018.08.004

Thiede, K., Oswalt, S., Brendefur, J. L., Carney, M. B., & Osguthorpe, R. D. (2019). Teachers' judgments of student learning of mathematics. In J. Dunlosky, & K. A. Rawson (Eds.), *Handbook of education. Cambridge handbook of cognition and education* (pp. 678–695). Cambridge University Press. https://www.cambridge.org/core/books/cambridge-handbook-of-cognition-and-education/3983FDC96F4E72A7F57445406E10F4F4.

Tofighi, D., & MacKinnon, D. P. (2011). RMediation: An R package for mediation analysis confidence intervals. *Behavior Research Methods, 43*(3), 692–700. https://doi.org/10.3758/s13428-011-0076-x

Trautwein, U., & Baeriswyl, F. (2007). Wenn leistungsstarke Klassenkameraden ein Nachteil sind [When high-performing classmates are a disadvantage]. *Zeitschrift für Padagogische Psychologie, 21*(2), 119–133. https://doi.org/10.1024/1010-0652.21.2.119

Urhahne, D. (2015). Teacher behavior as a mediator of the relationship between teacher judgment and students' motivation and emotion. *Teaching and Teacher Education, 45*, 73–82. https://doi.org/10.1016/j.tate.2014.09.006

Urhahne, D., & Wijnia, L. (2021). A review on the accuracy of teacher judgments. *Educational Research Review, 32*, Article 100374. https://doi.org/10.1016/j.edurev.2020.100374

Vieluf, S., Praetorius, A.-K., Rakoczy, K., Kleinknecht, M., & Pietsch, M. (2020). Angebots-nutzungs-modelle der Wirkweise des unterrichts: Ein kritischer vergleich verschiedener modellvarianten [Opportunity-and-use-models for teaching effectiveness: A critical comparison of different model variants]. *Zeitschrift für Padagogik, 66 Beiheft 1/20*, 63–80.

Wadmare, P., Nanda, M., Sabates, R., Sunder, N., & Wadhwa, W. (2022). Understanding the accuracy of teachers' perceptions about low achieving learners in primary schools in rural India: An empirical analysis of alignments and misalignments. *International Journal of Educational Research Open, 3*, Article 100198. https://doi.org/10.1016/j.ijedro.2022.100198

Wammes, D., Slof, B., Schot, W., & Kester, L. (2023). Teacher judgement accuracy of technical abilities in primary education. *International Journal of Technology and Design Education, 33*, 415–438. https://doi.org/10.1007/s10798-022-09734-5

Westphal, A., Gronostaj, A., Vock, M., Emmrich, R., & Harych, P. (2016). Differenzierung im gymnasialen Mathematik und Deutschunterricht – vor allem bei guten Diagnostiker/innen und in heterogenen Klassen [Differentiation in high school mathematics and German lessons – especially with good diagnosticians and in heterogeneous classes]? *Zeitschrift für Padagogik, 62*(1), 131–148.

Westphal, A., Zuber, J., & Vock, M. (2018). Welche Rolle spielen Selbstwirksamkeit, Motivation und Einstellungen zu Diagnostik für die Nutzung datenbasierter Rückmeldungen [The link between teachers' use of empirical feedback and their selfefficacy, motivation, and attitudes towards diagnostics]. *Zeitschrift für Bildungsforschung, 8*(3), 289–307. https://doi.org/10.1007/s35834-018-0223-x

Zhu, Y. (2022). Reading matters more than mathematics in science learning: An analysis of the relationship between student achievement in reading, mathematics, and science. *International Journal of Science Education, 44*(1), 1–17. https://doi.org/10.1080/09500693.2021.2007552

Zhu, M., Urhahne, D., & Rubie-Davies, C. M. (2018). The longitudinal effects of teacher judgement and different teacher treatment on students' academic outcomes. *Educational Psychology, 38*(5), 648–668. https://doi.org/10.1080/01443410.2017.1412399

Zitzmann, S., Wagner, W., Hecht, M., Helm, C., Fischer, C., Bardach, L., et al. (2022). How many classes and students should ideally be sampled when assessing the role of classroom climate via student ratings on a limited budget? An optimal design perspective. *Educational Psychology Review, 34*, 511–536. https://doi.org/10.1007/s10648-021-09635-4