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

Content-specificity of teachers' judgment accuracy regarding students' academic achievement

Teaching and teacher education 100 (2021), 13 S.

Teaching and Teacher Education 100 (2021) 103298

Contents lists available at ScienceDirect

Teaching and Teacher Education

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

Content-specificity of teachers' judgment accuracy regarding students' academic achievement

Dimitra Kolovou [a, *], Alexander Naumann [b, d], Jan Hochweber [a, d], Anna-Katharina Praetorius [c]

[a] St. Gallen University of Teacher Education (PHSG), Institute for Educational Assessment, Notkerstrasse 27, 9000, St. Gallen, Switzerland
[b] DIPF | Leibniz Institute for Research and Information in Education, Department of Educational Quality and Evaluation, Rostocker Straße 6, 60323, Frankfurt am Main, Germany
[c] University of Zurich, Institute of Education, Chair for Research on Learning, Instruction and Didactics, Freiestrasse 36, 8032, Zurich, Switzerland
[d] IDeA Research Center Frankfurt, Rostocker Straße 6, 60323, Frankfurt am Main, Germany

HIGHLIGHTS

- Explored content-specificity of teachers' judgment accuracy in mathematics and German language class.
- Modelled accuracy and the relationships between content domains simultaneously.
- Used Bayesian multivariate multilevel models with latent predictor variables.
- Found low to medium average judgment accuracy in all content domains.
- Content domains were substantially correlated, yet empirically distinguishable.

ARTICLE INFO

Article history:
Received 26 January 2020
Received in revised form
21 December 2020
Accepted 19 January 2021
Available online 8 February 2021

Keywords:
Judgment accuracy
Teacher judgments
Content domains
Academic achievement
Multilevel modelling
Multilevel IRT

ABSTRACT

Teachers' accuracy in judging students' achievement is often assumed to be a general ability of teachers. Based on this assumption, teachers should be at least consistent in their accuracy across different content domains within a school subject. Yet, this assumption has rarely been investigated empirically so far. Data from 54 mathematics teachers ($N = 1170$ students) and 55 language teachers ($N = 1255$ students) were analysed using a Bayesian multivariate multilevel modelling approach. Results indicate that latent accuracy measures across content domains indeed are substantially correlated within both investigated subjects, but may still be considered to represent different dimensions.

Judging students' academic abilities is an important task of teachers, which drives their daily decision-making. Teachers' judgments influence, for example, their lesson planning, the selection and difficulty level of learning activities and materials, and serve as basis for adaptive interactions with their students (Alvidrez & Weinstein, 1999; Herppich et al., 2018;; Loibl, Leuders, & Dörfler, 2020). The ability of teachers to make accurate judgments, also referred to as *teachers' judgment accuracy*, is therefore considered a necessary condition for meaningful teaching activities, especially in terms of optimal tailoring teaching to students' strengths and needs (Begeny, Eckert, Montarello, & Storie, 2008; Hoge & Coladarci, 1989; Pielmeier, Huber, & Seidel, 2018). Adaptive teaching behaviour is in turn related to positive student academic outcomes (Brühwiler & Blatchford, 2011; Corno, 2008; Parsons et al., 2018). Therefore, it is assumed that a high level of teachers' judgment accuracy has a positive impact on teaching effectiveness

* Corresponding author.
E-mail addresses: dimitra.kolovou@phsg.ch (D. Kolovou), Naumanna@dipf.de (A. Naumann), jan.hochweber@phsg.ch (J. Hochweber), anna.praetorius@ife.uzh.ch (A.-K. Praetorius).

https://doi.org/10.1016/j.tate.2021.103298

(e.g., Hoge & Coladarci, 1989; Thiede, Oswalt, Brendefur, Carney, & Osguthorpe, 2019; Urhahne & Wijnia, 2021). Research has provided empirical evidence supporting this assumption (Anders, Kunter, Brunner, Krauss, & Baumert, 2010; Behrmann & Souvignier, 2013; Helmke & Schrader, 1987; Thiede et al., 2018). Consequently, many studies have focused on investigating teachers' judgment accuracy regarding students' academic achievement (for overviews see Südkamp, Kaiser, & Möller, 2012; Kaufmann, 2020; Urhahne & Wijnia, 2021).

Previous studies on teachers' judgment accuracy of student achievement have typically focused on one single academic domain.[1] That is, studies have investigated either judgments of overall achievement in one subject (i.e., subject domain) or in a single content area within a subject (i.e., content domain). Researchers have commonly assumed that judgment accuracy is a general ability of the teacher (Artelt & Rausch, 2014; Hurwitz, Elliott, & Braden, 2007; Schrader, 2010) and can therefore be generalised across (content) domains. Based on this assumption, studies have often used single measures of accuracy in one specific content domain (e.g., arithmetics) to examine how accurate teachers are in judging students' ability in a subject as a whole (e.g., mathematics; Gabriele, Joram, & Park, 2016; Lorenz & Artelt, 2009). Yet, there is a distinct lack of studies that investigate whether it is possible to infer from teachers' judgment accuracy in one content domain (e.g., arithmetic) that teachers will also judge accurately in other content domains (e.g., geometry) of a specific subject (for an exception see Lorenz & Artelt, 2009).

The question of content-generality versus content-specificity is relevant for understanding the structure of judgment accuracy, that is, whether judgment accuracy can be mapped to different content domains and therefore is content-general or whether it consists of distinguishable content-specific facets (Herppich et al., 2018). Clarifying the structure of teachers' judgment accuracy is also important with respect to its measurement. If judgment accuracy is content-general, a single measure of accuracy suffices for gaining insight into teachers' judgment accuracy across an array of content domains. Otherwise, multiple content-specific measures are necessary. A further implication concerns the ways in which the development of judgment accuracy can be fostered, that is, whether content-specific rather than content-general trainings are more effective.

To pursue this issue in a systematic manner, the present study investigates whether teachers' judgment accuracy concerning students' academic achievement is specific to different content domains within the two different subjects, mathematics and German language class. To examine judgment accuracy in different content domains and the relations among them simultaneously, we applied an innovative multivariate multilevel latent modelling approach. This approach mitigates typical methodological limitations of previous studies of teachers' judgment accuracy (see Challenges in Measuring Teachers' Judgment Accuracy section).

In the following sections, we first elaborate on teachers' judgment accuracy regarding students' academic achievement and summarise previous findings. Afterwards, we consider the content-specificity of teachers' judgment accuracy from a theoretical perspective and report related empirical results in prior studies. Subsequently, our methodological approach is described. Finally, our research questions and hypotheses are presented.

1. Accuracy of teachers' judgments regarding students' academic achievement

Teacher judgments are defined as accurate when they are consistent with objective assessments of students' academic achievement (e.g., test scores; Hoge & Coladarci, 1989; Kaufmann, 2020; Ready & Wright, 2011). Student achievement is usually measured either by standardised tests or by curriculum-based measurement procedures (CBM; see Eckert, Dunn, Codding, Begeny, & Kleinmann, 2006; Feinberg & Shapiro, 2003, 2009). A commonly used measure of teachers' judgment accuracy is based on computing correlations between teachers' judgments of their individual students' achievement and the students' test performance for each single classroom or teacher, respectively, and averaging after applying Fisher's z-transformation (e.g., Helmke & Schrader, 1987; Südkamp et al., 2012; Urhahne & Wijnia, 2021).

Overall, meta-analyses on teachers' judgment accuracy concerning students' achievement indicate that teacher judgments are fairly accurate (Hoge & Coladarci, 1989; Kaufmann, 2020; Südkamp et al., 2012). Compared to the mean correlation $r = 0.63$ reported by Südkamp et al. (2012; for similar results see Hoge & Coladarci, 1989), Kaufmann (2020) found a higher correlation of $r = 0.80$ when re-analysing Hoge and Coladarci's (1989) data. In both studies, no statistically significant differences were found in teachers‘ average judgment accuracy between mathematics and language classes (Kaufmann, 2020; Südkamp et al., 2012). However, Südkamp et al. (2012) found a wide range of accuracy coefficients with respect to language classes ($r = -0.03$ to $r = 0.84$) compared to mathematics ($r = 0.35$ to $r = 0.80$).

2. Content-specificity of teachers' judgment accuracy

Teachers' judgment accuracy of students' achievement is often implicitly conceptualised as content-general (Herppich et al., 2018). At the same time, some research evidence suggests that accuracy is specific to the content domain being judged (Karst, 2012).

First, in schools, learning is in large part content-specific (Baumert, Lüdtke, Trautwein, & Brunner, 2009; Seidel & Shavelson, 2007). The teaching and learning contents are typically structured around subjects, which consist of highly associated content domains that are nevertheless psychometrically distinguishable, suggesting, therefore, the use of domain-specific assessments (Brunner, 2006; Harks, Klieme, Hartig, & Leiss, 2014; Lonigan & Milburn, 2017). Accordingly, teachers need to judge students' achievement not only in domains specified at the subject level but also, and more importantly, in content domains within subjects. This in turn enables teachers to gain deeper insight into students' understanding and provide appropriate learning opportunities (Artelt & Rausch, 2014; Brunner, Anders, Hachfeld, & Krauss, 2013; Seidel & Shavelson, 2007; Shulman, 1987). Furthermore, results from interview studies on teachers' decision-making in lesson planning show that teachers focus their lesson planning on the specific content to be taught. In doing this, they take into account their judgments of students' respective abilities (Morine-Dershimer, 1978–1979; Randi & Corno, 2005; Shavelson & Stern, 1981).

Second, for content-specific judgments to be accurate, content knowledge (CK) and pedagogical content knowledge (PCK) are regarded as a basic prerequisite (Herppich et al., 2018; Shulman, 1987; Thiede et al., 2015). However, teachers can at the same time show strengths and weaknesses with regard to the content domains within a subject. For example, it has been shown that teachers may have sound knowledge of geometry while

[1] Academic domains are often defined broadly as subject matter or discipline (e.g., mathematics). They can also be defined more specifically with respect to distinct content areas within a subject, such as algebra in mathematics (Harks et al., 2014). In the present study, we will use the term *content domain* to refer to distinct content areas within a subject and *subject domain* to refer to the subject as a whole. Domain will be used to refer to both subjects and content areas.

simultaneously having relative weaknesses in algebra (Blömeke, Kaiser, Döhrmann, & Lehmann, 2010). Accordingly, the judgment accuracy of individual teachers may vary between different content domains in which students' achievement is being judged (Herppich et al., 2018). The specific nature of such knowledge and its impact on the content-specificity of teacher judgments is also supported by a recent study by Hoppe, Renkl, and Rieß (2020).The The study aimed at fostering pre-service biology teachers' ability to make "on-the-fly" judgments of students' conceptions by conveying topic-specific pedagogical content knowledge. It was found that the acquisition of pedagogical content knowledge related to a specific topic (e.g., importance of plants in ecosystems) was only effective for judgments on that topic and did not lead to better judgments on other topics (e.g., decomposition).

Third, research on teacher expertise also speaks to the possibility that the ability to accurately judge students' achievement may not be generalizable across various content domains. This ability is considered to be one of expert teachers (Bromme, 2014; Leinhardt & Smith, 1985; Weinert, Schrader, & Helmke, 1990), and one that is acquired and fostered during teacher education (Dünnebier, Gräsel, & Krolak-Schwerdt, 2009; Hoppe et al., 2020; Van Ophuysen, 2006) as well as professional development (Thiede et al., 2015, 2018). However, considering that expertise has consistently been found to be domain-dependent, (expert) teachers cannot be assumed to be experts in judging students' achievement to the same extent across all content domains (Palmer, Stough, Burdenski, & Gonzales, 2005; see also Berliner, 1994, 2004).

To date, only very few empirical studies have addressed the specificity of teachers' judgment accuracy concerning student achievement across different domains. The study by Lorenz and Artelt (2009) primarily focused on cross-subject variation of primary school teachers' judgment accuracy. In their study, accuracy in each of two language content domains (vocabulary range and reading comprehension) and in arithmetic were weakly correlated ($r = 0.07$ to $r = 0.18$). Across two measurement points, teachers' accuracy within each of the two language domains was substantially correlated (ranging from $r = 0.42$ to $r = 0.44$). Overall, however, the correlations between and within the different content domains were at best moderate, suggesting that teachers' judgment accuracy may be specific to individual content domains. Using confirmatory factor analyses, Lintorf et al. (2011) investigated the dimensionality of teachers' accuracy in assessing the difficulty of two reading tasks with six items each. The results provided no evidence to support one-dimensionality within a task. Instead, teachers' judgment accuracy showed to be dependent on the item difficulty of each task. For example, teachers who were able to make an accurate judgment on difficult items were less accurate on easy items. Praetorius, Karst, Dickhäuser, and Lipowsky (2011) investigated the domain-specificity of primary teachers' judgment accuracy of students' academic self-concept. In their study, they examined the correlations of judgment accuracy across different domains ("reading comprehension", "writing competence", and "mathematics") based on three different accuracy measures. Across the different accuracy measures, significant correlations were found only between the content domains "reading comprehension" and "writing competence" (ranging from $r = 0.38$ to $r = 0.80$), when the same judgment accuracy measures were used. Accordingly, a substantial degree of overlap between the language domains was evident.

The aforementioned studies focused on the subject-specificity of teachers' judgment accuracy, while mainly investigating primary school teachers. The only study that took into account the specificity of judgment accuracy concerning student achievement across different content domains, focused on consistency across two language content domains (Lorenz & Artelt, 2009). Initial findings from these studies suggest that teachers' judgment accuracy is specific to different subjects (mathematics and language classes) or reading tasks but not so much so to specific language content domains (e.g., vocabulary range and reading comprehension). However, the current state of knowledge about the degree to which teachers' judgment accuracy is specific to different content domains is rather limited. This is also due to methodological limitations of previous studies with respect to the measurement of teachers' judgment accuracy.

3. Challenges in Measuring Teachers' judgment accuracy

In previous investigations of the specificity of teachers' judgment accuracy across domains, judgment accuracy has been operationalised as the correlation between teachers' judgments and test performance for each single classroom or teacher (see Lorenz & Artelt, 2009). While this measure is common in the research on teachers' judgment accuracy, it has some significant limitations (Südkamp et al., 2012). Both teacher judgments and student achievement scores are subject to sampling and measurement error that may attenuate the correlation between teacher judgments and students' achievement (Kaiser, Südkamp, & Möller, 2017). In research on teachers' judgment accuracy, single-item measures of teacher judgments are common (see Südkamp et al., 2012), while achievement measures are based on a rather limited number of test items per student, leading to unreliable point estimates for both measures. Small sample sizes ($n < 30$), as are typical regarding the number of students judged per teacher, lead to imprecise estimates of accuracy at the classroom/teacher level (Schönbrodt & Perugini, 2013; see also Praetorius, Koch, Scheunpflug, Zeinz, & Dresel, 2017). Furthermore, teachers differ in the number of students for whom judgments are made. When averaging the computed correlations across teachers or classrooms, these differences are not being weighted accordingly. Hence, the calculated mean value does not reflect the mean judgment accuracy among teachers (Dollinger, 2013). Finally, although hierarchical data structures (i.e., judgments of students nested in classrooms or teachers) are common in teachers' judgment accuracy research, they are not directly taken into account in the previous measurement of judgment accuracy. Students in the same classroom tend to be more similar in terms of their level of achievement, and disregarding such dependencies may result in too small standard error estimates and too liberal significance tests (Hox, Moerbeek, & van de Schoot, 2018; see also; Dollinger, 2013).

In order to address these methodological challenges, multilevel modelling techniques (for an overview, see Snijders & Bosker, 2012) are increasingly being used in research on teachers' judgment accuracy (e.g., Dollinger, 2013; Kaiser et al., 2017; Karst & Bonefeld, 2020; Kilday, Kinzie, Mashburn, & Whittaker, 2012; Meissel, Meyer, Yao, & Rubie-Davies, 2017; Ready & Wright, 2011). However, previous studies have been limited to to modelling accuracy for a single domain at a time (i.e. a single outcome variable). Using multilevel regression, the accuracy measure in one domain is based on the (random) slope of test performance (i.e., test scores) when predicting teachers' judgments (Dollinger, 2013; Karst & Bonefeld, 2020; Karst, Hartig, Kaiser, & Lipowsky, 2017; Meissel et al., 2017). Still, measurement error in test performance and teacher judgments, and sampling error in test performance (the predictor variable) is commonly neglected. The multilevel modelling approach with latent variables used in this study enables model-based estimations of teachers' judgment accuracy in multiple domains simultaneously, appropriate handling of hierarchical and imbalanced data structures, and the specification of latent variables to deal with measurement error comparable, for instance, to "doubly latent" analyses of contextual effects (Lüdtke, Marsh, Robitzsch, & Trautwein, 2011; Marsh et al., 2012).

4. The present study

The present study seeks to systematically examine the content-specificity of secondary school teachers' judgment accuracy within each of two subjects, mathematics and German language class. More precisely, we examine the content-specificity of teachers' judgment accuracy by investigating the relations of judgment accuracy across three corresponding content domains within each subject. In each content domain within the two subjects, we focus on so-called global judgments of students' achievement. This type of judgment concerns ratings of students' overall performance in a domain and is typically examined using Likert-type rating scales (Karing, Matthäi, & Artelt, 2011; see also; Südkamp et al., 2012). In particular, we asked teachers to make a global rating in each content domain for each of their students. The choice of two subjects makes it possible to examine the extent to which the results can be generalised across subjects. The content domains examined are "number and variable", "shape and space", and "measures, functions, data, and probabilities" for mathematics, and "reading comprehension", "listening comprehension", and "language(s) in focus" for German language class.

To investigate the relations of judgment accuracy across multiple domains we used an innovative multivariate multilevel latent modelling approach that deals with typical methodological limitations of previous studies. In previous studies on this topic, teachers' judgment accuracy was typically operationalised as the correlation between teacher judgments and students' scores on standardised tests, calculated separately for each classroom or teacher. The domain-specificity was investigated examining the manifest intercorrelations of these accuracy measures across various content domains (Lorenz & Artelt, 2009; Praetorius et al., 2011). However, the operationalisation of judgment accuracy as the correlation between teacher judgments and test performance is commonly criticised for being unreliable (see Challenges in Measuring Teachers' Judgment Accuracy section). Accordingly, the previously used measures of accuracy could lead to an underestimation of the true relationships of teachers' judgment accuracy across multiple domains. In our study, therefore, we implemented a multivariate multilevel latent modelling approach which accounts for multivariate correlated outcomes enabling us to appropriately model teachers' judgment accuracy within each content domain, and simultaneously to examine the latent relations across them. By doing so, we explicitly took into account the hierarchical data structure of students nested within classrooms/teachers and considered both sampling error (due to a limited number of students in each classroom) and measurement error in test scores (due to a limited number of test items per student). Our hypotheses ($H_1 - H_4$) were as follows:

H1. Mathematics teachers' judgments are positively associated with their students' test performance in each mathematical content domain.

H2. Language teachers' judgments are positively associated with their students' test performance in each language content domain.

Based on previous research on the accuracy of teacher judgments (Südkamp et al., 2012), we expected positive and moderate to strong average associations between teacher judgments and test performance in each content domain.

H3. Mathematics teachers' judgment accuracy measures in different mathematical content domains correlate positively.

H4. Language teachers' judgment accuracy measures in different language content domains correlate positively.

In line with our theoretical considerations and previous findings on the content-specificity of teachers' judgment accuracy (Lorenz & Artelt, 2009; Praetorius et al., 2011), we expected that the content-specific accuracy measures within each subject correlate at least moderately positively, but remain clearly distinguishable (i.e., correlations are not close to perfect).

5. Method

5.1. Design and sample

The research project on which the present study is based was conducted with a sample of 18 public lower secondary level schools (*Sekundarstufe I;* part of the compulsory education) that comprised grade levels 7 to 9 from the German-speaking Swiss Canton of Zurich (see Helbling, Tomasik, & Moser, 2019 for a more detailed description of the educational context of Switzerland and the Canton of Zurich). Participation of the schools was voluntary, but within the participating schools, teachers and students were (with few exceptions) obliged to take part in the project. The project encompassed four measurement points. All students who were admitted to the seventh grade of each school in the 2016/2017 school year were drawn to participate. The content-specific measures of teacher judgments and respective measures of seventh graders' academic achievement that were used for the present study were collected at the first measurement point. Students completed computerised curriculum-based tests in the two subjects of mathematics and German language right at the beginning of Grade 7 (September/November 2016). Online questionnaires were used to collect all remaining data, including students' and teachers' demographic data as well as teacher judgments regarding their students' performance. Teachers made their judgments in December/January 2017 over a period of four weeks. At this time point, teachers and students had known each other for approximately four months.

In the present study, we limited our analyses to mathematics and language teachers and their students for whom the following data were available: (a) teachers' judgments of their students' performance in all three mathematical or language content domains and (b) standardised tests completed by the students. Both types of data were necessary for calculating the measures of teachers' judgment accuracy. When a teacher's judgments and/or students' performance on standardised tests were not available, that teacher and her/his students were not included in the analyses. In our study, this was the case for teachers of one school that participated in the research program from which data were used in this study after the students' performance was measured for the first time. In addition, there were also some teachers for whom their judgments or standardised test data of all their students were not available due to lack of participation. The resulting data set for our analyses comprised 54 mathematics teachers (out of $n = 63$) and 55 language teachers (out of $n = 61$) from 17 schools.

Of the nine mathematics teachers who were not included in the analyses, three did not judge their students in any content domain, four judged them in less than three domains (two of them due to teaching the same students but in different content domains), and for the remaining two teachers standardised test data was not available. Of the six language teachers who were not included in the analyses, four did not judge their students in any content domain, one judged them in less than three domains, and one teacher's standardised test data was not available.

Of the 54 mathematics teachers included in the analyses, 41%

were women, and 30% had 1–5 years of teaching experience, 38% 6–10 years, 15% 16–25 years, and 17% up to 25 years (1.9% missing data). Of the 55 language teachers, 58% were women, and 28% had 1–5 years of teaching experience, 34% 6–15 years, 13% 16–25 years, and 25% up to 25 years (3.6% missing data). Twelve teachers, who were teaching both mathematics and German language in their classrooms, provided judgments for both subjects.[2]

Furthermore, as a result of the aforementioned inclusion criteria, out of the project's overall student sample of 1462 students (49% female) at an average age of 13 years ($SD = 0.51$), data of 1170 students from mathematics classrooms and 1255 students from German language classrooms were analysed. For these students, teacher judgments in all three content domains in mathematics and/or German language class were available. With regard to the standardised test results, missing data on the individual domains varied between 1.88% – 3.07% in mathematics and between 2.39% – 3.67% in German language class.

5.2. Variables

Achievement in mathematics and German language. Tests were administered in three content domains in mathematics and German language, respectively, according to the common curriculum for German-speaking Switzerland: (a) "number and variable" (i.e., arithmetic and algebra); (b) "shape and space" (i.e., geometry); (c) "measures, functions, data, and probabilities"; (e) "reading comprehension"; (f) "listening comprehension"; and (g) "language(s) in focus" (assessing knowledge in language awareness, lexis, pronunciation, grammar, orthography and language learning reflection).

In the Canton of Zurich, lower secondary school consists of two or three levels (A, B, and possibly C, depending on the respective school), with A being the most demanding level. Additionally, students are taught in mathematics and German language in separate performance-based classrooms – I, II or III – with I being the most challenging. To enable the administration of tests corresponding to students' performance in different performance-based classrooms, a multi-matrix design was used for each test with three different test booklets of varying average item difficulty: easy, medium, and hard. Common items (anchors) in all test booklets were placed within the same relative position in order to ensure comparability of test performance at both individual student-level and group-level (i.e., classroom/teacher-level). The tests for the content domains consisted of 20–25 dichotomously scored items, which showed satisfactory fit to the Rasch model (Rasch, 1960). The reliability for each dimension was generally satisfactory (WLE reliabilities: "shape and space": 0.73; "number and variable": 0.76; "measures, functions, data and probabilities": 0.74; "listening comprehension": 0.68; "reading comprehension": 0.66; "language(s) in focus": 0.73). Moreover, due to the inclusion of test performance as latent variables in our models, unreliability in the point estimates was less of a concern.

Teacher judgments in mathematics and German language. Following the commonly used approach in research on teachers' judgment accuracy, mathematics teachers and German language teachers were asked to predict the test performance of each student (Südkamp et al., 2012; see also; Hoge & Coladarci, 1989). These judgments had to be provided separately for each of the six content domains in mathematics and German language. Prior to rating students' performance in each content domain, teachers were provided with ten (in mathematics) or seven (in German) preselected test items that were included in all test booklets from the corresponding content domain. This allowed the teachers to become familiar with the specific test content. Judgments were collected via 10-point Likert scales to allow for higher sensitivity in capturing teacher judgments (see Zhu & Urhahne, 2020). The lowest and highest response category of the rating scale were labelled as follows: "0–10%, the 10% lowest-performing students" and "90–100%, the 10% highest-performing students", respectively. Teachers were encouraged to give their individual appraisal of each student in comparison to all other students of the same grade level (i.e., seventh grade) in the Canton of Zurich. This is in line with current suggestions for comparisons beyond the classroom context, as teacher ratings are likely to be influenced by how well each student performs in relation to average performance of their classroom (Baudson, Fischbach, & Preckel, 2016; see also; Lazarides, Viljaranta, Aunola, & Nurmi, 2018; Wright & Wiese, 1988). For example, the following instruction was used for the content domain "number and variable": "For each student, please tick the box indicating how in your estimation he or she has performed on the test for the content domain NUMBER AND VARIABLE in comparison with all other students (roughly at the beginning of seventh grade) in the Canton of Zurich".

Control variables at classroom/teacher-level. The project from which this study draws data used a quasi-experimental design with assignment of schools to the treatment and control conditions. Teachers from schools in the treatment condition attended training programs in mathematics and German language didactics. The programs were designed to sensitize teachers to the learning difficulties of low-achieving students and provide guidance for teachers to provide adequate support for these students. Thus, the programs were primarily expected to have an indirect positive influence (via enhanced teaching quality) on these students' mathematics and German language achievement. However, due to this focus on student achievement, it is possible that the participating teachers were also paying particular attention when judging the performance of their low-achieving students. Since our interest was not in these potential effects of the training programs, we decided to control for potential treatment effects in our analyses. We did this in line with previous studies on teacher judgment accuracy, which drew on data from research projects with comparable designs (e.g., Furnari, Whittaker, Kinzie, & DeCoster, 2017). Accordingly, we controlled for teaching in a treatment school using a dummy-coded grouping variable, "treatment" ($0 =$ control group, $1 =$ treatment group). As some teachers taught in more than one participating classroom, we controlled for the "assignment of teachers" to multiple classes with another dummy variable ($0 =$ one class, $1 =$ several classes). 19 mathematics teachers (out of $n = 54$) and 17 language teachers (out of $n = 55$) taught multiple classes in mathematics and German language class, respectively.

5.3. Analyses

Multilevel Modelling. As our study addresses multiple content domains simultaneously, we enhanced previous approaches to account for multivariate outcomes and regression parameter correlations on a latent level. For each subject, we specified one multivariate regression model with students i nested in classrooms/teachers j and teacher judgments Y_{dij} as outcomes (see Fig. 1). Here, let Y_{dij} be the judgment in the dth mathematical or language content domain for the ith student of the jth classroom/teacher, with $d = \{1, 2, 3\}$ within each subject (mathematics: 1: "shape and space", 2: "measures, functions, data, and probabilities", 3: "number and variable"; German language: 1: "listening comprehension", 2: "reading comprehension", 3: "language(s) in

[2] These teachers were comparable to all other teachers in terms of teaching experience, age and accuracy in the mathematical domains. Minor differences in favour of the majority of German teachers were found in the language domains.

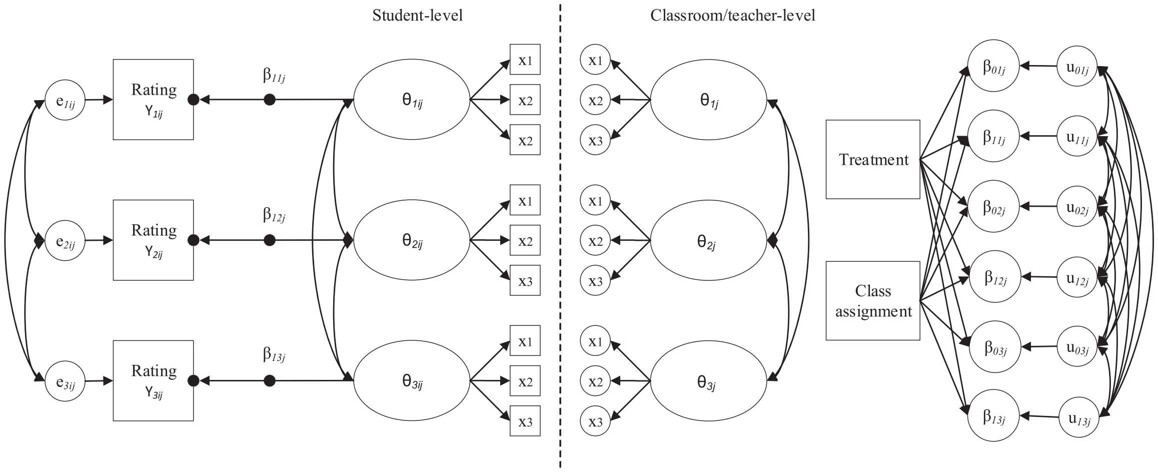

Fig. 1. Illustration of the multilevel random intercepts and random slopes regression model with multivariate outcomes (ratings of three different content domains $Y_{1ij} - Y_{3ij}$) predicted by latent ability on student-level in each content domain (θ_{1ij}, θ_{2ij}, θ_{3ij}). Parameters θ_{1j}, θ_{2j}, and θ_{3j} denote student ability on the classroom/teacher-level, as measured by test items x1, x2, etc. Random intercepts (β_{01j}, β_{02j}, β_{03j}) and random slopes (β_{11j}, β_{12j}, β_{13j}) may correlate across outcomes indicating the degree of content-specificity (right part of the illustration). We entered two teacher-level variables (dummy variables), "Treatment" and "Class assignment", as classroom/teacher-level control variables.

focus"). The outcome variables, that is, the teacher judgments Y_{dij}, were z-standardised before entering the model for each subject. In each multivariate model, we added the within-teacher-component of student test performance within the corresponding content domain, θ_{dij}, as latent predictor on the student-level, while we controlled for "treatment" and "class assignment" at the classroom/teacher-level. The resulting multilevel model for the multivariate outcomes in each subject is:

$$Y_{dij} \sim MN(\beta_{0dj} + \beta_{1dj}\theta_{dij}, \Sigma)$$

with

$$\beta_{kdj} \sim MN(\gamma_{0kd} + \gamma_{1kd} TREATMENT_j + \gamma_{2kd} CLASS\,ASSIGNMENT_j, \mathrm{T})$$

for each of the $k = \{0, 1\}$ random student-level regression coefficients β_{kdj} per classroom/teacher j and content domain d. That is, the regression slopes of student test performance θ_{dij} in each content domain d were allowed to vary across classrooms/teachers j, resulting in a random intercepts and random slopes regression. Accordingly, T is a 6 × 6 covariance matrix comprising a) three random intercepts variances, b) three random slopes variances, as well as c) information on the covariance of intercepts and slopes across and within the three content domains per subject.

The latent predictor variable θ_{dij}, that is, a student's within-classroom/teacher ability component, was estimated from a three-dimensional multilevel 1pl IRT (ML-MIRT) model with between-item multidimensionality (i.e., each item measures only one dimension; see Reckase, 2009). This student-level ability component is by definition group-mean centered. The ML-MIRT model for mathematics comprised the three content-specific dimensions "shape and space", "measures, functions, data, and probabilities", and "number and variable". Similarly, the model for German language class comprised the content-specific dimensions "listening comprehension", "reading comprehension", and "language(s) in focus".

When examining teachers' judgment accuracy ($H_1 - H_2$), we were interested in the pure student-level effect of test performance, $\theta_{dij,}$ on teacher judgments. Thus, we standardised the slope coefficients using standardisation on the student-level as part of the model fitting procedure (*within-group* standardisation; see Schuurman, Ferrer, de Boer-Sonnenschein, & Hamaker, 2016). In a first step, we estimated the student-level variances of the predictors (test performance in each content domain) and outcome variables (teacher judgment in each content domain) within each Markov chain Monte Carlo (MCMC) iteration (for details on estimation, see below). The classroom/teacher-specific standardised regression coefficients were then calculated as the product of the unstandardised coefficients and the ratio of the classroom/teacher-specific standard deviations of the predictor variable and the outcome variable. Subsequently, the average standardised regression coefficient (i.e., the average accuracy across classrooms/teachers) was estimated by calculating the average of the classroom/teacher-specific standardised coefficients in each MCMC iteration. Substantively, the estimated classroom/teacher-specific standardised regression coefficients reflect the amount of student-level standard deviations that teacher judgments will increase when the student achievement increases by one classroom/teacher-specific standard deviation (see Schuurman et al., 2016). Due to the standardisation on the student-level, the standardised coefficients can be interpreted in a similar way to the correlation coefficients in other studies on teachers' judgment accuracy (Kaiser et al., 2017; Kilday et al., 2012), while taking sampling and measurement error in the classroom/teacher-specific accuracy estimates into account.

To analyse the content-specificity of teachers' judgment accuracy ($H_3 - H_4$), we used the random-slope covariance parameters to compute latent correlations between random slopes across the mathematical and language content domains, respectively. The correlation coefficients provided information about the extent to which teachers' judgment accuracy is consistent across content domains. Lower correlations indicate a lower consistency, which in turn reflects a higher degree of content-specificity. Because we controlled for the effects of the two dummy variables, we evaluated the residual correlations rather than unconditional correlations (see also right part of Fig. 1). To gain further insight into the effects of the control variables, we also estimated the regression models without controlling for the effects of the dummy variables and compared the results with respect to the latent correlations.

To provide additional evidence for content-specificity, we compared the two subject-specific multivariate multilevel regression models to models in which the random slopes within each subject were set to be equal. That is, we checked whether teachers' judgment accuracy is multidimensional (i.e., content-specific) or unidimensional (i.e., content-general). We utilized the Deviance Information Criterion (DIC; Spiegelhalter, Best, Carlin, & Van Der Linde, 2002) for model comparison.

Estimation and inference. All analyses were carried out in the Bayesian framework (e.g., Fox, 2010) using MCMC estimation. As we had no reliable prior information available, we assumed vague prior distributions only. All models were estimated using four chains with 2500 samples each after a burn-in phase of 5000 samples and a thinning interval of ten (i.e., every 10th iteration was recorded).

We derived point estimates by computing the mean of the posterior distribution of each of the parameters. Additionally, we computed 95% Bayesian credibility intervals (BCI) for all parameters which indicate a statistically significant result when the BCI does not comprise zero. We tested the equality of parameters (e.g., between two standardised regression coefficients) by calculating the difference δ between each pair of parameters as well as its BCI. Following this approach, two parameters are equal if the BCI of the difference contains zero.

We assessed model convergence by visual inspection of the MCMC chains and by calculating the Gelman-Rubin R statistic (R-hat; Gelman et al., 2013) for each parameter. Convergence analyses for all model parameters showed that R-hat values are all less than 1.1, indicating that acceptable MCMC convergence was achieved. However, there was one mathematics classroom/teacher for which students' ability parameters did not converge well (i.e., R-hat > 1.1). The values of the model parameters with and without this classroom/teacher did not differ substantially, yet standard errors were larger when including this classroom/teacher. Hence, we report the results including this classroom/teacher, resulting in a more conservative way of hypothesis testing. All models were estimated using R 3.6.0 (R Core Team, 2019), JAGS 4.3.0 (Plummer, 2017), coda (Plummer, Best, Cowles, & Vines, 2006) and mcmcplots (Curtis, 2018).

6. Results

6.1. Descriptive analyses

Table 1 provides an overview of basic descriptive statistics for teachers' judgments. In addition to the means and standard deviations, percentile ranks are also given to provide more detailed information on the ranges and distributions of teacher judgments. All variables were, on average, close to the theoretical scale mean (5.5) but exhibited a large variation. In Table 2, intercorrelations on student-level for teacher judgments and test performance between the content domains for each subject are shown. Both teacher judgments and test performance (on student-level) between the content domains of mathematics and German were strongly interrelated. The intraclass correlation (ICC), which is the proportion of variance at the classroom/teacher-level, indicated for both achievement measures (i.e., teachers' judgments and test performance) that a considerable proportion of the variability existed between classrooms/teachers (see Table 2). However, in all content domains, the ICC was higher for test performance than for the teacher judgments.

Table 1
Descriptive statistics for teachers' judgments by content domain.

Variable	M	SD	Percentile				
			0	25	50	75	100
Mathematics[a]							
SHSP	5.34	2.27	1	4	5	7	10
MFDP	5.14	2.25	1	3	5	7	10
NV	5.46	2.24	1	4	6	7	10
German language[b]							
LC	5.91	2.14	1	4	6	8	10
RC	5.95	2.17	1	4	6	8	10
LiF	5.41	2.25	1	4	6	7	10

Note. SHSP = shape and space; MFDP = measures, functions, data, and probabilities; NV = number and variable; LC = listening comprehension; RC = reading comprehension; LiF = language(s) in focus.

[a] $n = 1170$.

[b] $n = 1255$.

6.2. Teachers' judgment accuracy

To evaluate the degree of correspondence between teacher judgments and test performance (i.e., teachers' judgment accuracy; $H_1 - H_2$), we examined the classroom/teacher-specific standardised coefficients of test performance when predicting teacher judgments in each mathematical or language content domain (see Analyses section). We first investigated the average effect of test performance on the corresponding teacher judgments, in other words, the mean judgment accuracy among teachers. We found—as hypothesized in H_1 and H_2—a positive and statistically significant relationship between teacher judgments and test performance in both the mathematical and language content domains with the corresponding BCIs not comprising zero. Across the mathematical content domains, the classroom/teacher-specific standardised regression coefficients showed a positive relationship to teacher judgments such that, on average, an increase of one SD in test performance was associated with a 0.25 [0.21, 0.30] SD increase in teacher judgments for „measures, functions, data, and probabilities“, a 0.28 [0.22, 0.33] SD increase for „shape and space”, and a 0.30 [0.24, 0.35] SD increase for „number and variable“. The mean of the standardised regression coefficients did not differ significantly among the three mathematical content domains ($\delta_{SHSP-MFDP} = 0.02$ [−0.02, 0.06]; $\delta_{SHSP-NV} = -0.02$ [−0.06, 0.02]; $\delta_{MFDP-NV} = -0.04$ [−0.08, 0.00]). This indicates that mathematics teachers' (in)accuracy in judging their test performance was similarly pronounced in all mathematical content domains.

Across the language content domains, the classroom/teacher-specific standardised regression coefficients showed a positive relationship to teacher judgments such that, on average, an increase of one SD in test performance was associated with a 0.30 [0.25, 0.34] SD increase in teacher judgments for „language(s) in focus“, a 0.32 [0.27, 0.36] SD increase for “listening comprehension”, and a 0.33 [0.28, 0.38] SD increase for “reading comprehension”. The mean of the standardised regression coefficients did not differ significantly among the three language content domains ($\delta_{RC-LiF} = 0.03$ [−0.01, 0.07]; $\delta_{LC-RC} = -0.01$ [−0.05, 0.03]; $\delta_{LC-LiF} = 0.02$ [−0.02, 0.06]. Accordingly, language teachers' (in) accuracy in judging their test performance was similarly pronounced in all content domains.

6.3. Content-specificity of teachers' judgment accuracy

To test the hypotheses that teachers' judgment accuracy measures in different content domains are positively correlated ($H_3 - H_4$), the latent correlations between the random slopes across the three content domains in each of the subjects of mathematics and German language were examined. More specifically, we report on the residual correlations between the random slopes (see Analyses section). As shown in Table 3 (for mathematics) and Table 4 (for German language), all effects captured by the two dummy variables—treatment and class assignment—were found to be nonsignificant. The results of the residual correlations for both subjects are presented in Table 5. With respect to mathematics teachers' judgment accuracy, the latent correlations across the

Table 2
Summary of intercorrelations on student-level and intraclass correlations of teacher judgments and students' test performance by content domain in mathematics (A) and German language (B).

A) Mathematics				
Variable	SHSP	MFDP	NV	ICC
SHSP		0.87 [0.80, 0.93]	0.85 [0.77, 0.92]	0.45 [0.36, 0.56]
MFDP	0.89 [0.86, 0.91]		0.79 [0.69, 0.88]	0.47 [0.37, 0.57]
NV	0.87 [0.84, 0.90]	0.92 [0.90, 0.94]		0.43 [0.33, 0.54]
ICC	0.63 [0.53, 0.73]	0.56 [0.46, 0.67]	0.62 [0.52, 0.72]	
B) German language				
Variable	LC	RC	LiF	ICC
LC		0.87 [0.80, 0.93]	0.86 [0.77, 0.92]	0.36 [0.27, 0.47]
RC	0.93 [0.91, 0.94]		0.88 [0.82, 0.93]	0.29 [0.20, 0.38]
LiF	0.87 [0.84, 0.90]	0.90 [0.88, 0.93]		0.47 [0.37, 0.57]
ICC	0.41 [0.31, 0.52]	0.42 [0.31, 0.52]	0.54 [0.43, 0.64]	

Note. Intercorrelations for teacher judgments on the student-level between the content domains are presented above the diagonal, and latent variable intercorrelations for students' test performance on the student-level between the content domains are presented below the diagonal. Intraclass correlation coefficients (ICC) for teacher judgments by content domain are presented in the vertical columns, and intraclass correlation coefficients for students' test performance by content domain are presented in the horizontal rows. Values in square brackets indicate the 95% Bayesian credible interval (BCI). SHSP = shape and space; MFDP = measures, functions, data, and probabilities; NV = number and variable; LC = listening comprehension; RC = reading comprehension; LiF = language(s) in focus.

Table 3
Predicting mathematics teacher judgments in each content domain.

Regression parameter	Outcomes		
	SHSP	MFDP	NV
	Fixed effects		
Intercept	−0.12 [−0.47, 0.24]	−0.15 [−0.51, 0.19]	−0.14 [−0.48, 0.21]
Student test performance	0.43 [0.22, 0.65]	0.32 [0.14, 0.51]	0.45 [0.24, 0.65]
Effect of treatment on intercept	0.06 [−0.37, 0.45]	0.10 [−0.31, 0.51]	0.06 [−0.35, 0.45]
Effect of treatment on slope	−0.08 [−0.33, 0.16]	−0.05 [−0.26, 0.17]	−0.05 [−0.30, 0.20]
Effect of class assignment on intercept	0.08 [−0.33, 0.53]	0.12 [−0.31, 0.55]	0.14 [−0.30, 0.54]
Effect of class assignment on slope	−0.01 [−0.25, 0.25]	0.05 [−0.18, 0.28]	−0.07 [−0.34, 0.18]
	Random effects		
Variance Intercept	0.51 [0.32, 0.74]	0.52 [0.32, 0.73]	0.48 [0.30, 0.70]
Variance Slope	0.15 [0.08, 0.23]	0.12 [0.06, 0.18]	0.15 [0.08, 0.22]
SD Intercept	0.71 [058, 0.87]	0.71 [0.58, 0.87]	0.69 [0.56, 0.84]
SD Slope	0.38 [0.29, 0.48]	0.34 [0.25, 0.43]	0.38 [0.29, 0.48]

Note. The predicted outcome is the z-standardised teacher judgment in the respective domain. Student test performance represents latent ability (group-mean-centered) on student-level in each content domain. Variances of intercepts and slopes are adjusted for the effects of treatment and class assignment. Values in square brackets indicate the 95% Bayesian credible interval (BCI). Note that the variances were rounded to two decimals. SHSP = shape and space; MFDP = measures, functions, data, and probabilities; NV = number and variable.

Table 4
Predicting German language teacher judgments in each content domain.

Regression parameter	Outcomes		
	LC	RC	LiF
	Fixed effects		
Intercept	0.03 [−0.28, 0.33]	0.00 [−0.28, 0.28]	0.00 [−0.34, 0.35]
Student test performance	0.58 [0.38, 0.78]	0.65 [0.46, 0.85]	0.53 [0.35, 0.70]
Effect of treatment on intercept	−0.04 [−0.40, 0.32]	0.00 [−0.32, 0.34]	−0.06 [−0.45, 0.37]
Effect of treatment on slope	−0.11 [−0.34, 0.13]	−0.17 [−0.39, 0.06]	−0.14 [−0.35, 0.07]
Effect of class assignment on intercept	−0.10 [−0.49, 0.29]	−0.06 [−0.42, 0.29]	−0.02 [−0.47, 0.40]
Effect of class assignment on slope	−0.09 [−0.35, 0.18]	−0.20 [−0.45, 0.04]	−0.13 [−0.37, 0.09]
	Random effects		
Variance Intercept	0.41 [0.25, 0.58]	0.33 [0.21, 0.48]	0.53 [0.33, 0.74]
Variance Slope	0.15 [0.08, 0.22]	0.12 [0.06, 0.18]	0.11 [0.06, 0.16]
SD Intercept	0.63 [0.51, 0.77]	0.57 [0.46, 0.69]	0.72 [0.59, 0.87]
SD Slope	0.38 [0.29, 0.47]	0.35 [0.26, 0.43]	0.33 [0.25, 0.41]

Note. The predicted outcome is the z-standardised teacher judgment in the respective domain. Student test performance represents latent ability (group-mean-centered) on student-level in each content domain. Variances of intercepts and slopes are adjusted for the effects of treatment and class assignment. Values in square brackets indicate the 95% Bayesian credible interval (BCI). Note that the variances were rounded to two decimals. LC = listening comprehension; RC = reading comprehension; LiF = Language(s) in Focus.

Table 5
Latent correlations of random slopes across content domains in mathematics (A) and German language (B).

A) Mathematics

Variable	SHSP	MFDP
MFDP	0.68 [0.49, 0.84]	
NV	0.67 [0.48, 0.84]	0.59 [0.34, 0.80]

B) German language

Variable	LC	RC
RC	0.63 [0.42, 0.81]	
LiF	0.58 [0.35, 0.78]	0.57 [0.34, 0.77]

Note. Residual correlations are depicted; the effects of the dummy variables (treatment, class assignment) were partialed out. Values in square brackets indicate the 95% Bayesian credible interval (BCI) for each correlation. SHSP = shape and space; MFDP = measures, functions, data, and probabilities; NV = number and variable; LC = listening comprehension; RC = reading comprehension; LiF = language(s) in focus.

mathematical content domains were between $r = 0.59$ [0.34, 0.80] and $r = 0.68$ [0.49, 0.84], revealing a substantial degree of overlap between them. H_3 is therefore strongly supported by the data. Similarly, German teachers' judgment accuracy—as hypothesized in H_4—strongly positively correlates across the language content domains: the correlation coefficients vary between $r = 0.57$ [0.34, 0.77] and $r = 0.63$ [0.42, 0.81]. Thus, at a latent level, the proportion of shared variance varies between 35% and 46% among the mathematical content domains, and between 32% and 40% among the language content domains.

For the sake of comparability, we briefly report below on the correlations between the slopes derived from the multivariate regression models uncontrolled for the effects of the dummy variables (not shown in Table 5; see Analyses section). The slopes within each subject correlated significantly, and the resulting correlation coefficients were all statistically significant and varied between $r = 0.74$ [0.59, 0.87] to $r = 0.82$ [0.72, 0.92] for the mathematical content domains and $r = 0.69$ [0.51, 0.84] and $r = 0.78$ [0.65, 0.89] for the language content domains. Although the correlation coefficients were generally higher, they were not significantly different from the coefficients when adding the control variables (reported in Table 5). The difference values ranged from $\delta_r = -0.12$ [-0.35, 0.09] to $\delta_r = -0.16$ [-0.39, 0.05] with respect to the mathematical content domains and from $\delta_r = -0.12$ [-0.41, 0.15] and $\delta_r = -0.16$ [-0.41, 0.07] regarding the language content domains.

Additional model comparisons indicated that the multidimensional models differentiating accuracy between content domains (DIC mathematics = 80,823; DIC German language: 110,940) showed better fit than the unidimensional models in each subject (DIC mathematics: 80,847; DIC German language: 111,046). The DIC differences (mathematics: $\delta_{DIC} = 23.78$, $SE = 21.39$; German language: $\delta_{DIC} = 105.65$, $SE = 21.47$) were greater than 10, suggesting that the unidimensional models can clearly be ruled out (Lunn, Jackson, Best, Thomas, & Spiegelhalter, 2013).

7. Discussion

The present study examined the extent to which teachers' judgment accuracy with respect to students' achievement is content-specific within each of the subjects of mathematics and German language class. To that end, the relationships between judgment accuracy in three mathematical content domains—"number and variable", "shape and space", "measures, functions, data, and probabilities"—and three language content domains—"reading comprehension", "listening comprehension", "language(s) in focus"—were examined using a Bayesian multivariate multilevel latent modelling approach. In doing so, we explicitly took into account the hierarchical data structure as well as sampling and measurement error (Lüdtke et al., 2008, 2011).

7.1. Teachers' judgment accuracy

In line with our expectations, both mathematics and language teacher judgments were positively associated with test performance in each content domain. To classify the level of judgment accuracy, our results can be compared to studies that operationalise judgment accuracy as the correlation between teacher judgments and test performance calculated separately for each classroom/teacher. This is possible because the classroom/teacher-specific standardised coefficients used to measure teachers' judgment accuracy in this study can be interpreted as correlation coefficients (see Analyses section; see also Kaiser et al., 2017; Kilday et al., 2012), although, importantly, they clearly refer to the relationships within classrooms/teachers. Compared to the average correlation of 0.63 in the meta-analysis of Südkamp et al. (2012), our analyses indicate low to medium average judgment accuracy (standardised coefficients ranging from 0.25 to 0.33) of mathematics and language teachers in all content domains. In particular, the results lie in the lower part of the correlation range (Südkamp et al., 2012; see also; Kaufmann, 2020), while no differences in average accuracy were found between the content domains of each subject.

7.2. Content-specificity of teachers' judgment accuracy

As was expected, the accuracy measures within each subject correlate strongly positively on a latent level across the content domains ($r = 0.57$ to $r = 0.68$). Hence, a noticeable amount of shared variance undoubtedly exists between accuracy in different content domains, indicating a relative similarity of judgment accuracy across them. Accordingly, mathematics or language teachers who make an accurate judgment in one content domain tend to form a judgment in other content domains that is not equivalent but comparable to a considerable extent in terms of its accuracy. However, in order to determine whether content domains of judgment accuracy can be empirically separated as distinguishable dimensions, it is necessary apart from model comparisons (i.e., multidimensional models versus unidimensional models) to interpret the magnitude of the latent correlations appropriately. To accomplish this, we refer to results on the separability of content domains in large-scale educational assessments like PISA (Programme for International Student Assessment; Organisation for Economic Co-operation and Development [OECD], 2019) and TIMSS (Trends in International Mathematics and Science Study; Mullis, Martin, Ruddock, O'Sullivan, & Preuschoff, 2009). In particular, we refer to results from latent correlations between content-specific dimensions based on multidimensional item response (MIRT) models. In these studies, latent correlations between content-specific dimensions in mathematics ranged from 0.62 to 0.91 (Blum et al., 2004; Brunner, 2006; Harks et al., 2014; Klieme, 2000; Liu, Wilson, & Paek, 2008) and were interpreted as indicating empirical separability. Accordingly, the latent correlations of the accuracy measures found in our study ($r = 0.59$ to $r = 0.68$ for mathematics and $r = 0.57$ to $r = 0.63$ for German language class) provide evidence that content-specific facets of judgment accuracy can be empirically distinguished both for mathematics and German language class. Based on our results, accuracy in different content domains of a subject cannot be understood to simply reflect a common “accuracy dimension”, and accuracy measures collected from different content domains cannot be used interchangeably without any reservations.

Previous results on the content-specificity only exist for language class. In the study of Lorenz and Artelt (2009) similar results were found (manifest correlations of $r = 0.42/0.44$ between vocabulary range and reading comprehension). Yet, the comparatively high latent correlations in our study could be due to methodological differences in operationalising teachers' judgment accuracy, the separability of the content domains, and the different grade levels being examined (see The Present Study section). In addition, one must keep in sight that in contrast to previous results based on manifest intercorrelations of accuracy measures, we examined the latent correlations between the accuracy measures across the content domains. It must also be noted that our study used global judgments with respect to *broad* curriculum-based content domains. While such content domains are psychometrically distinguishable in students' test data, they remain highly associated (the latent variable intercorrelations for students' test performance on the student-level in our study ranged between $0.87 \leq r \leq 0.92$ and $0.87 \leq r \leq 0.93$ in mathematical and language content domains, respectively), as they share skills and abilities which concurrently contribute and confound the learning development of each other (see Harks et al., 2014; Leinhardt, Zaslavsky, & Stein, 1990; Lonigan & Milburn, 2017). Furthermore, in the curricula, content domains are often confounded due to wording in the description of the related abilities and skills. As a result, teachers may tend to think about their students' ability in multiple content domains, although they are about to estimate students' ability only in a single content domain (Llosa, 2007). This could explain the high intercorrelations of teacher judgments within each subject that we found, which in turn may explain the high correlations between the content-specific accuracy measures.

Reflecting on our results, it is possible that teachers' judgment accuracy regarding students' achievement in both mathematics and German language class is organised in a multidimensional structure (Gabriele et al., 2016; Karst, Dotzel, & Dickhäuser, 2018; Lintorf et al., 2011; Spinath, 2005), which differentiates into content-specific facets. That is, content-specific facets of judgment accuracy may be nested in broad subject-specific factors reflecting judgment accuracy in mathematics and German language class, respectively. This is also supported by studies which indicate that teachers' judgment accuracy can be better described as a subject-related construct comprising more differentiated content-specific facets (Lorenz & Artelt, 2009; Praetorius et al., 2011). Additional support comes from the study of Hoppe et al. (2020), who showed that teachers' ability to make judgments of students' conceptions is acquired in a content-specific (i.e., topic-specific) way.

Given the multi-faceted structure that our results suggest, it is probably best to opt for content-specific measures of judgment accuracy, although the use of such specific measures depends on the research question under investigation, for example, whether the focus of a study is on a content domain or the subject as a whole. Even in the latter case, one could argue for the use of content-specific measures, since such measures are likely to provide more information than a single measure at the subject level. However, whether accuracy measures of aggregated content-specific judgments represent teachers' accuracy at the subject level cannot be answered in the present study and requires future research. Finally, designing trainings with content-specific foci may be fruitful in fostering judgment accuracy in different content domains within a subject (see Hoppe et al., 2020; Thiede et al., 2018). In particular, focusing on improving the utilisation of cues in the process of making judgments that are more predictive to students' achievement within a domain seems to be a promising approach (Oudman, van de Pol, Bakker, Moerbeek, & van Gog, 2018; Thiede et al., 2015, 2018). In such trainings, it might be of great importance to enhance teachers' content-specific knowledge so that the teachers are more likely to be able to use the appropriate cues (Artelt & Rausch, 2014; see Thiede et al., 2015, 2018, for an example). In this regard, Thiede et al. (2018), who examined the effects of different professional development programs on teachers' judgment accuracy, suggested that increasing *pedagogical content knowledge* may contribute to improved judgment accuracy and student achievement.

7.3. Limitations and directions for further research

The current study contributes to research on teachers' judgment accuracy as we extended previous studies on content-specificity by investigating whether secondary school teachers' judgment accuracy is specific to different mathematical or language content domains. Furthermore, we extended previously used multilevel modelling approaches to simultaneously model teachers' judgment accuracy in multiple content domains as well as the relationships among them.

Our study has, however, several limitations. First, these derive from the sample and the instruments. In the present study, we used a data set from a research project with a quasi-experimental design with an assignment of schools to treatment and control conditions. Although we controlled for potential treatment effects using a dummy-coded grouping variable at the level of the classrooms/teachers in line with previous studies (see Furnari et al., 2017) in all our models, an impact of the treatment (a teacher training program) on our results cannot be ruled out completely. However, it may be noted that we used data from the first of four measurement occasions in the project, that is, from an early stage of the training program, when any effects of the program can be expected to be relatively small.

With respect to the instruments used, teachers were asked to rate each students' test performance on a series of 10-point scales in comparison to other students of the same grade level and region. Standardised tests, on the other hand, measure student performance based on a series of tasks. Accordingly, it cannot be ruled out that teachers, taking into account their daily interaction with their students, judge students' overall competence rather than students' test performance, which could lead to over- or underestimation in judgments (Karing, 2009). Furthermore, the judgment task used in this study (rating of achievement in a domain) can be characterized less specific than other tasks such as rankings (i.e., ranking of students in their class with respect to their achievement) or estimating the number of correctly solved items in a test (Südkamp et al., 2012). However, in the meta-analysis of Südkamp et al. (2012), no effects of the specificity of the judgment task on teachers' average judgment accuracy were found.

In addition, the teachers in this study made their judgments at least two months after the performance test, which was administered right at the beginning of seventh grade (see Design and

Sample section). Accordingly, it cannot be ruled out that some students improved considerably – or, to the contrary, made less progress than should be expected – in the time span between the performance test and the teachers' judgments. Similarly, it is possible that teachers took into account the daily performance of students within this time interval when judging their students. This may have resulted in teachers rating their students higher or lower than they would have done if both measures had been collected simultaneously. Südkamp et al. (2012) considered the time gap between the collection of teacher judgments and measures of students' academic achievement in their meta-analysis. They first classified studies according to when performance tests were administrated: (a) at the same time as the teacher judgments (within a 1-month period; 73.3%), (b) at least one month after the teacher judgments (8.3%) and (c) at least one month before the teacher judgments (18.3%). Then, they investigated the moderating effect of the time gap for the 61 effect sizes included in their analysis. None of the effects of the time interval were statistically significant, that is, temporal proximity was not associated with higher judgment accuracy. Nevertheless, future studies should carefully consider the potential impact of time gaps when planning their studies.

Furthermore, although we have followed the most common approach in judgment accuracy research for measuring teacher judgments (see Südkamp et al., 2012; see also; Feinberg & Shapiro, 2009; Hoge & Coladarci, 1989), the proximity of this measure to assessment situations in daily teaching is limited (Kaiser, Praetorius, Südkamp, & Ufer, 2017). Accordingly, future research should be devoted to the development of measures with higher ecological validity and to the investigation of their content-specificity.

Moreover, we examined judgment accuracy in broadly defined mathematical and language content domains and their relations (see Harks et al., 2014). We investigated these relations separately for each subject. Cross-subject relations between the content domains could not be studied because only a very small subsample of teachers taught and judged (the same) students in both subjects. Future studies should, however, specifically aim to investigate cross-subject relations and plan the sampling of teachers accordingly. Besides, since teacher judgments depend on the nature of a domain, it is possible that our results cannot be generalised to other types of judgments. This could be the case for judgments that relate to domains defined at more fine-grained levels such as topic- and task-specific judgments (see Hoppe et al., 2020; Lintorf et al., 2011), or to other subjects. In this respect it is also not clear to what extent the results can be generalised to other grade levels and educational systems. For instance, the extent to which judgment accuracy is content-specific or -general might depend on the structure and contents of teacher education in the respective content domain (see Blömeke, Kaiser, Döhrmann, & Lehmann, 2010). However, as can be deduced from research results on teacher knowledge, content-specificity in relation to teachers should be defined more broadly than in expert research (Blömeke, Busse, Kaiser, König, & Suhl, 2016). Finally, the question remains open to which extent content knowledge and pedagogical content knowledge influence the accuracy of teacher judgments in different content domains and consequently the relations across them (Herppich et al., 2018; Thiede et al., 2018).

7.4. Conclusions

We investigated the content-specificity of teachers' judgment accuracy, a rather neglected topic in research on judgment accuracy. To that end, we used a multivariate multilevel modelling approach with latent predictor variables, which represents also a methodological extension of the multilevel modelling techniques previously used in this research area. We provided empirical evidence for strongly associated, but psychometrically separable content-specific facets of judgment accuracy for both mathematics and language teachers. Therefore, depending on the focus of the study, future studies should consider this aspect when deciding how to measure and promote teacher judgment accuracy. In order to gain differentiated insights into teachers' accuracy in assessing students' performance within a subject or in a particular content area, content-specific measures should undoubtedly be preferred. More generally, researchers should carefully consider when generalizing teachers' ability to accurately gauge students' performance across domains.

Author Statement

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

Author Note

This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors. Special thanks go to Lena Hollenstein and Olivia Rütti-Joy for their feedback and support.

References

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

Artelt, C., & Rausch, T. (2014). Accuracy of teacher judgments. When and for what reasons? In S. Krolak-Schwerdt, S. Glock, & M. Böhmer (Eds.), *Teachers' professional development: Assessment, training, and learning* (pp. 27–43). Rotterdam: Sense Publishers.

Baudson, T. G., Fischbach, A., & Preckel, F. (2016). Teacher judgments as measures of children's cognitive ability: A multilevel analysis. *Learning and Individual Differences, 52*, 148–156. https://doi.org/10.1016/j.lindif.2014.06.001

Baumert, J., Lüdtke, O., Trautwein, U., & Brunner, M. (2009). Large-scale student assessment studies measure the results of processes of knowledge acquisition: Evidence in support of the distinction between intelligence and student achievement. *Educational Research Review, 4*(3), 165–176. https://doi.org/10.1016/j.edurev.2009.04.002

Begeny, J. C., Eckert, T. L., Montarello, S. A., & Storie, M. S. (2008). Teachers' perceptions of students' reading abilities: An examination of the relationship between teachers' judgments and students' performance across a continuum of rating methods. *School Psychology Quarterly, 23*(1), 43–55. https://doi.org/10.1037/1045-3830.23.1.43

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

Berliner, D. C. (1994). Expertise: The wonder of exemplary performances. In J. N. Mangieri, & C. C. Block (Eds.), *Creating powerful thinking in teachers and students* (pp. 161–186). Forth Worth, TX: Holt, Richart & Wiston.

Berliner, D. C. (2004). Describing the behaviour and documenting the accomplishments of expert teachers. *Bulletin of Science, Technology & Society, 24*(3), 200–212. https://doi.org/10.1177/0270467604265535

Blömeke, S., Busse, A., Kaiser, G., König, J., & Suhl, U. (2016). The relation between content-specific and general teacher knowledge and skills. *Teaching and Teacher Education, 56*, 35–46. https://doi.org/10.1016/j.tate.2016.02.003

Blömeke, S., Kaiser, G., Döhrmann, M., & Lehmann, R. (2010). Mathematisches und mathematikdidaktisches Wissen angehender Sekundarstufen-I-Lehrkräfte im internationalen Vergleich [Mathematical content and mathematics pedagogical content knowledge of prospective secondary school I teachers in international comparison]. In S. Blömeke, G. Kaiser, & R. Lehmann (Eds.), *TEDS-M 2008. Professionelle Kompetenz und Lerngelegenheiten angehender Mathematiklehrkräfte für die Sekundarstufe I im internationalen Vergleich* (pp. 197–238). Münster: Waxmann.

Blum, W., Neubrand, M., Ehmke, T., Senkbeil, M., Jordan, A., Ulfig, F., & Carstensen, C. (2004). Mathematische Kompetenz [Mathematical competence]. In PISA-Konsortium Deutschland (Ed.), *PISA 2003. Der Bildungsstand der Jugendlichen in Deutschland–Ergebnisse des zweiten internationalen Vergleichs* (pp. 47–92). Münster, Germany: Waxmann.

Bromme, R. (2014). *Der Lehrer als Experte [The teacher as expert]*. Münster: 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. (2006). *Mathematische Schülerleistung: Struktur, Schulformunterschiede und Validität [Student achievement in mathematics: Structure, school type differences and validity] (Unpublished doctoral dissertation)*. Germany: Humboldt-Universität zu Berlin. Retrieved from http://library.mpib-berlin.mpg.de/diss/Brunner_Dissertation.pdf.

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* (pp. 229–248). New York, NY: Springer.

Corno, L. (2008). On teaching adaptively. *Educational Psychologist, 43*(3), 161–173. https://doi.org/10.1080/00461520802178466

Curtis, S. (2018). *mcmcplots: Create plots from MCMC output. R package version 0.4, 3* https://cran.r-project.org/web/packages/mcmcplots/mcmcplots.pdf.

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

Dünnebier, K., Gräsel, C., & Krolak-Schwerdt, S. (2009). Urteilsverzerrungen in der schulischen Leistungsbeurteilung: Eine experimentelle Studie zu Ankereffekten [Biases in teachers' assessments of student performance: An experimental study of anchoring effects]. *Zeitschrift für Padagogische Psychologie, 23*(3–4), 187–195. https://doi.org/10.1024/1010-0652.23.34.187

Eckert, T. L., Dunn, E. K., Codding, R. S., Begeny, J. C., & Kleinmann, A. E. (2006). Assessment of mathematics and reading performance: An examination of the correspondence between direct assessment of student performance and teacher report. *Psychology in the Schools, 43*(3), 247–265. https://doi.org/10.1002/pits.20147

Feinberg, A., & Shapiro, E. S. (2003). Accuracy in teacher judgments of student achievement in reading: A comparison of judgment measures. *School Psychology Quarterly, 18*, 62–65.

Feinberg, A. B., & Shapiro, E. S. (2009). Teacher accuracy: An examination of teacher-cased judgments of students' reading with differing achievement levels. *The Journal of Educational Research, 102*(6), 453–462. https://doi.org/10.3200/JOER.102.6.453-462

Fox, J.-P. (2010). *Bayesian item response modelling: Theory and applications*. New York, NY: Springer.

Furnari, E. C., Whittaker, J., Kinzie, M., & DeCoster, J. (2017). Factors associated with accuracy in prekindergarten teacher ratings of students' mathematics skills. *Journal of Psychoeducational Assessment, 35*(4), 410–423. https://doi.org/10.1177/0734282916639195

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

Gelman, A., Carlin, J. B., Stern, H. S., Dunson, D. B., Vehtari, A., & Rubin, D. B. (2013). *Bayesian data analysis* (3rd ed.). New York, NY: Chapman and Hall/CRC.

Harks, B., Klieme, E., Hartig, J., & Leiss, D. (2014). Separating cognitive and content domains in mathematical competence. *Educational Assessment, 19*(4), 243–266. https://doi.org/10.1080/10627197.2014.964114

Helbling, L. A., Tomasik, M. J., & Moser, U. (2019). Long-term trajectories of academic performance in the context of social disparities: Longitudinal findings from Switzerland. *Journal of Educational Psychology, 111*(7), 1284–1299. https://doi.org/10.1037/edu0000341

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

Herppich, S., Praetorius, A.-K., Förster, N., Glogger-Frey, I., Karst, K., Leutner, D., … Südkamp, A. (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

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

Hoppe, T., Renkl, A., & Rieß, W. (2020). Förderung von unterrichtsbegleitendem Diagnostizieren von Schülervorstellungen durch Video und Textvignetten [Fostering on-the-fly judgements of students' conceptions using video and text vignettes]. *Unterrichtswissenschaft, 48*, 573–597. https://doi.org/10.1007/s42010-020-00075-7

Hox, J. J., Moerbeek, M., & van de Schoot, R. (2018). *Multilevel analysis: Techniques and applications* (3rd ed.). New York, NY: Routledge.

Hurwitz, J. T., Elliott, S. N., & Braden, J. P. (2007). The influence of test familiarity and student disability status upon teachers' judgments of students' test performance. *School Psychology Quarterly, 22*(2), 115–144. https://doi.org/10.1037/1045-3830.22.2.115

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). Münster: Waxmann.

Kaiser, J., Südkamp, A., & Möller, J. (2017). The effects of student characteristics on teachers' judgment accuracy: Disentangling ethnicity, minority status, and achievement. *Journal of Educational Psychology, 109*(6), 871–888. https://doi.org/10.1037/edu0000156

Karing, C. (2009). Diagnostische Kompetenz von Grundschul- und Gymnasiallehrkräften im Leistungsbereich und im Bereich Interessen [Diagnostic competence of elementary and secondary school teachers in the domains of competence and interests]. *Zeitschrift für Padagogische Psychologie, 23*(3–4), 197–209. https://doi.org/10.1024/1010-0652.23.34.197

Karing, C., Matthäi, J., & Artelt, C. (2011). Genauigkeit von Lehrerurteilen über die Lesekompetenz ihrer Schülerinnen und Schüler in der Sekundarstufe I – Eine Frage der Spezifität? [Lower Secondary School Teacher Judgment Accuracy of Students' Reading Competence – A Matter of Specificity?]. *Zeitschrift für Padagogische Psychologie, 25*(3), 159–172. https://doi.org/10.1024/1010-0652/a000041

Karst, K. (2012). *Kompetenzmodellierung des diagnostischen Urteils von Grundschullehrern [Competency modelling of the diagnostic judgment of primary school teachers]*. Münster: Waxmann.

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. Article.

Karst, K., Dotzel, S., & Dickhäuser, O. (2018). Comparing global judgments and specific judgments of teachers about students' knowledge: Is the whole the sum of its parts? *Teaching and Teacher Education, 76*, 194–203. https://doi.org/10.1016/j.tate.2018.01.013

Karst, K., Hartig, J., Kaiser, J., & Lipowsky, F. (2017). Mehrebenenmodelle als Werkzeuge zur Analyse diagnostischer Kompetenz von Lehrkräften – ein lineares Mischmodell (LMM) und seine Anwendung in R [Multilevel models as tools for analysing diagnostic competence of teachers – a linear mixed model]. In A.-K. Praetorius, & A. Südkamp (Eds.), *Diagnostische kompetenz von Lehrkräften* (pp. 153–177). Münster: Waxmann.

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. Article.

Kilday, C. R., Kinzie, M. B., Mashburn, A. J., & Whittaker, J. V. (2012). Accuracy of teacher judgments of preschoolers' math skills. *Journal of Psychoeducational Assessment, 30*(2), 148–159. https://doi.org/10.1177/0734282911412722

Klieme, E. (2000). Fachleistungen im voruniversitären Mathematik- und Physikunterricht: Theoretische Grundlagen, Kompetenzstufen und Unterrichtsschwerpunkte [Subject specific achievement in preuniversity mathematics and physics instruction: Theoretical basis, competence levels and instructional focuses]. In J. Baumert, W. Bos, & R. Lehmann (Eds.), *TIMSS/III. Dritte Internationale Mathematik- und Naturwissenschaftsstudie–Mathematische und naturwissenschaftliche Bildung am Ende der Schullaufbahn: Bd. 2. Mathematische und physikalische Kompetenzen am Ende der gymnasialen Oberstufe* (pp. 57–128). Opladen, Germany: Leske & Budrich.

Lazarides, R., Viljaranta, J., Aunola, K., & Nurmi, J. E. (2018). Teacher ability evaluation and changes in elementary student profiles of motivation and performance in mathematics. *Learning and Individual Differences, 67*, 245–258. https://doi.org/10.1016/j.lindif.2018.08.010

Leinhardt, G., & Smith, D. A. (1985). Expertise in mathematics instruction: Subject matter knowledge. *Journal of Educational Psychology, 77*(3), 247–271. https://doi.org/10.1037/0022-0663.77.3.247

Leinhardt, G., Zaslavsky, O., & Stein, M. K. (1990). Functions, graphs, and graphing: Tasks, learning, and teaching. *Review of Educational Research, 60*(1), 1–64. https://doi.org/10.3102/00346543060001001

Liu, O. L., Wilson, M., & Paek, I. (2008). A multidimensional Rasch analysis of gender differences in PISA mathematics. *Journal of Applied Measurement, 9*(1), 18–35.

Lintorf, K., McElvany, N., Rjosk, C., Schroeder, S., Baumert, J., Schnotz, W., … Ullrich, M. (2011). Zuverlässigkeit von diagnostischen Lehrerurteilen – reliabilität verschiedener Urteilsmaße bei der Einschätzung von Aufgabenschwierigkeiten [Reliability of diagnostic teacher judgments – reliability of different judgment measures in the assessment of task difficulties]. *Unterrichtswissenschaft, 39*(2), 102–120. https://doi.org/10.3262/UW1102102

Llosa, L. (2007). Validating a standards-based classroom assessment of English proficiency: A multitrait-multimethod approach. *Language Testing, 24*(4), 489–515. https://doi.org/10.1177/0265532207080770

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. Article.

Lonigan, C. J., & Milburn, T. F. (2017). Identifying the dimensionality of oral language skills of children with typical development in preschool through fifth grade. *Journal of Speech, Language, and Hearing Research, 60*(8), 2185–2198. https://doi.org/10.1044/2017_JSLHR-L-15-0402

Lorenz, C., & Artelt, C. (2009). Fachspezifität und Stabilität diagnostischer Kompetenz von Grundschullehrkräften in den Fächern Deutsch und Mathematik [Domain specifity and stability of diagnostic competence among primary school teachers in the school subjects of German and mathematics]. *Zeitschrift für Padagogische Psychologie, 23*(3–4), 211–222. https://doi.org/10.1024/1010-0652.23.34.211

Lüdtke, O., Marsh, H. W., Robitzsch, A., & Trautwein, U. (2011). A 2× 2 taxonomy of multilevel latent contextual models: Accuracy—bias trade-offs in full and partial error correction models. *Psychological Methods, 16*(4), 444–467. https://doi.org/10.1037/a0024376

Lüdtke, O., Marsh, H. W., Robitzsch, A., Trautwein, U., Asparouhov, T., & Muthén, B. (2008). The multilevel latent covariate model: A new, more reliable approach to group level effects in contextual studies. *Psychological Methods, 13*(3), 203–229. https://doi.org/10.1037/a0012869

Lunn, D., Jackson, C., Best, N., Thomas, A., & Spiegelhalter, D. (2013). *The bugs Book: A practical Introduction to Bayesian Analysis*. Boca Raton, Florida: CRC Press.

Marsh, H. W., Lüdtke, O., Nagengast, B., Trautwein, U., Morin, A. J., Abduljabbar, A. S., & Köller, O. (2012). Classroom climate and contextual effects: Conceptual and methodological issues in the evaluation of group-level effects. *Educational Psychologist, 47*(2), 106–124. https://doi.org/10.1080/00461520.2012.670488

Meissel, K., Meyer, F., Yao, E. S., & Rubie-Davies, C. M. (2017). Subjectivity of teacher judgments: Exploring student characteristics that influence teacher judgments of student ability. *Teaching and Teacher Education, 65*, 48–60. https://doi.org/10.1016/j.tate.2017.02.021

Morine-Dershimer, G. (1978-79). Planning in classroom reality: An in-depth look. *Educational Research Quarterly, 3*(4), 83–99.

Mullis, I. V. S., Martin, M. O., Ruddock, G. J., O'Sullivan, C. Y., & Preuschoff, C. (2009). *TIMSS 2011 assessment frameworks*. Chestnut Hill, MA: Boston College.

OECD. (2019). *PISA 2018 Assessment and Analytical Framework*. Paris: OECD Publishing.

Oudman, S., van de Pol, J., Bakker, A., Moerbeek, M., & van Gog, T. (2018). Effects of different cue types on the accuracy of primary school teachers' judgments of students' mathematical understanding. *Teaching and Teacher Education, 76*, 214–226. https://doi.org/10.1016/j.tate.2018.02.007

Randi, J., & Corno, L. (2005). Teaching and learner variation. In P. D. Tomlinson, J. Dockrell, & P. H. Winne (Eds.), *Pedagogy – teaching for learning, (British Journal of Educational Psychology Monograph Series II, No. 3)*, (pp. 47–69). Leicester: British Psychological Society. doi:10.1348/000709905X62110.

Palmer, D. J., Stough, L. M., Burdenski, J., Thomas, K., & Gonzales, M. (2005). Identifying teacher expertise: An examination of researchers' decision making. *Educational Psychologist, 40*(1), 13–25. https://doi.org/10.1207/s15326985ep4001_2

Parsons, S. A., Vaughn, M., Scales, R. Q., Gallagher, M. A., Parsons, A. W., Davis, S. G., … Allen, M. (2018). Teachers' instructional adaptations: A research synthesis. *Review of Educational Research, 88*(2), 205–242. https://doi.org/10.3102/0034654317743198

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

Plummer, M. (2017). *JAGS version 4.3. 0 user manual [Computer software manual]*. Retrieved from https://sourceforge.net/projects/mcmc-jags/files/Manuals/4.x/.

Plummer, M., Best, N., Cowles, K., & Vines, K. (2006). CODA: Convergence diagnosis and output analysis for MCMC. *R News, 6*(1), 7–11.

R Development Core Team. (2019). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing. Retrieved from http://www.r-project.org/index.html.

Praetorius, A.-K., Karst, K., Dickhäuser, O., & Lipowsky, F. (2011). Wie gut schätzen Lehrer die Fähigkeitsselbstkonzepte ihrer Schüler ein? Zur diagnostischen Kompetenz von Lehrkräften [How teachers rate their students: On teachers' diagnostic competence regarding the academic self-concept]. *Psychologie in Erziehung und Unterricht, 58*, 81–91. https://doi.org/10.2378/peu2010.art30d

Rasch, G. (1960). *Studies in mathematical psychology: I. Probabilistic models for some intelligence and attainment tests*. Copenhagen: Nielsen & Lydiche.

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

Reckase, M. D. (2009). *Multidimensional item response theory*. New York, NY: Springer.

Schönbrodt, F. D., & Perugini, M. (2013). At what sample size do correlations stabilize? *Journal of Research in Personality, 47*(5), 609–612. https://doi.org/10.1016/j.jrp.2013.05.009

Schrader, F. W. (2010). Diagnostische Kompetenz von Eltern und Lehrern [Diagnostic competence of teachers and parents]. In D. H. Rost (Ed.), *Handwörterbuch pädagogische psychologie* (pp. 102–108). Weinheim: Beltz Verlag.

Schuurman, N. K., Ferrer, E., de Boer-Sonnenschein, M., & Hamaker, E. L. (2016). How to compare cross-lagged associations in a multilevel autoregressive model. *Psychological Methods, 21*(2), 206–221. https://doi.org/10.1037/met0000062

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

Shavelson, R. J., & Stern, P. (1981). Research on teachers' pedagogical thoughts, judgments, decisions, and behaviour. *Review of Educational Research, 51*(4), 455–498. https://doi.org/10.3102/00346543051004455

Shulman, L. (1987). Knowledge and teaching: Foundations of the new reform. *Harvard Educational Review, 57*(1), 1–22. https://doi.org/10.17763/haer.57.1.j463w79r56455411

Snijders, T., & Bosker, R. (2012). *Multilevel analysis: An introduction to basic and advanced multilevel modelling* (2nd ed.). London, UK: Sage.

Spiegelhalter, D. J., Best, N. G., Carlin, B. P., & Van Der Linde, A. (2002). Bayesian measures of model complexity and fit. *Journal of the Royal Statistical Society: Series B (Statistical Methodology), 64*, 583–639. https://doi.org/10.1111/1467-9868.00353

Spinath, B. (2005). Akkuratheit der Einschätzung von Schülermerkmalen durch Lehrer und das Konstrukt der diagnostischen Kompetenz [Accuracy of teacher judgments on student characteristics and the construct of diagnostic competence]. *Zeitschrift für Padagogische Psychologie, 19*(1/2), 85–95. https://doi.org/10.1024/1010-0652.19.1.85

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., & Osguthorpe, R. D. (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

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

Thiede, K. W., Brendefur, J. L., Osguthorpe, R. D., Carney, M. B., Bremner, A., Strother, S., … Jesse, D. (2015). Can teachers accurately predict student performance? *Teaching and Teacher Education, 49*, 36–44. https://doi.org/10.1016/j.tate.2015.01.012

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*. NY, NY: Cambridge University Press. https://www.cambridge.org/core/books/cambridge-handbook-of-cognition-and-education/3983FDC96F4E72A7F57445406E10F4F4.

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. Article.

Van Ophuysen, S. (2006). Vergleich diagnostischer Entscheidungen von Novizen und Experten am Beispiel der Schullaufbahnempfehlung [Comparison of diagnostic decisions between novices and experts: The example of school career recommendation]. *Zeitschrift für Entwicklungspsychologie und Pädagogische Psychologie, 38*(4), 154–161. https://doi.org/10.1026/0049-8637.38.4.154

Weinert, F. E., Schrader, F.-W., & Helmke, A. (1990). Educational expertise: Closing the gap between educational research and classroom practice. *School Psychology International, 11*(3), 163–180. https://doi.org/10.1177/0143034390113002

Wright, D., & Wiese, M. J. (1988). Teacher judgment in student evaluation: A comparison of grading methods. *The Journal of Educational Research, 82*(1), 10–14. https://doi.org/10.1080/00220671.1988.10885858

Zhu, C., & Urhahne, D. (2020). Temporal stability of teachers' judgment accuracy of students' motivation, emotion, and achievement. *European Journal of Psychology of Education*. https://doi.org/10.1007/s10212-020-00480-7. Advanced online publication.