Karing, Constance; Rausch, Tobias; Artelt, Cordula

Teacher Judgement Accuracy : Measurements, Causes and Effects

Date of secondary publication: 10.09.2025

Version of Record (Published Version), Bookpart

Persistent identifier: urn:nbn:de:bvb:473-irb-110334x

Primary publication

Karing, Constance; Rausch, Tobias; Artelt, Cordula (2024): Teacher Judgement Accuracy : Measurements, Causes and Effects, in: S. Weinert, H.-G.Roßbach, J. von Maurice, u. a. (Ed.), Educational Processes, Decisions, and the Development of Competencies from Early Preschool Age to Adolescence : Findings from the BiKS Cohort Panel Studies, Wiesbaden: Springer Fachmedien Wiesbaden, pp. 263–280, doi: 10.1007/978-3-658-43414-4_10.

Teacher Judgement Accuracy—Measurements, Causes and Effects

Constance Karing, Tobias Rausch and Cordula Artelt

Abstract

The formation of accurate judgements on students' performance is often considered as part of teachers' professional competence. Moreover, inaccurate judgements are seen as determinants of social inequality. Using data of BiKS-8-18 on teacher ratings and student performance, the paper gives an overview on different theoretical approaches and operationalization of judgement accuracy as well as their results in terms of homogeneity, stability over time, inter-individual differences and the effects of judgement accuracy on students' further achievement. Primary school teachers outperform secondary school teachers in accurately assessing student performance. Furthermore, judgement accuracy did not show to be a general ability. Applying to different student characteristics, however, related to subject areas/domains it proofed to be a relatively time persistent teacher ability. Teacher judgements are somewhat sensitive to characteristics at the class and student level, although bias

C. Karing (✉)
University of Jena, Jena, Germany
e-mail: constance.karing@uni-jena.de

T. Rausch
University of Bamberg, Bamberg, Germany

C. Artelt
Leibniz Institute for Educational Trajectories, Bamberg, Germany
e-mail: cordula.artelt@lifbi.de

S. Weinert et al. (eds.), *Educational Processes, Decisions, and the Development of Competencies from Early Preschool Age to Adolescence*, Edition ZfE 16,
https://doi.org/10.1007/978-3-658-43414-4_10

related to students' gender and social status was not found for teachers at secondary level. We found positive effects of teacher judgements on students' achievement gains, particularly in the domain of reading. Finally, by taking into account an add-on study of teachers' content related knowledge related to judgements on reading performance, we discuss the findings and further highlight the need to take into account judgement purposes and demands in future research.

Keywords

Judgement accuracy · Diagnostic competence · Learning outcomes · Summative assessment · Secondary school

Introduction

Teacher judgements are key to a variety of professional behaviors, including standard settings for lessons, educational feedback on students' performance and the awarding of grades and certificates. Some of these judgements are primarily related to educational goals (e.g., initiating assessment aimed at optimizing students' learning) and thus are labeled formative assessment or assessment ***for*** learning. Others refer to the formal evaluation and summative assessment of students' performance, like grades, certificates and decisions on placement and tracking (Looney 2011; OECD 2012) and are forms of assessment ***of*** learning. Given the range and relevance of teacher judgements, its potential bias and its effects on students' (future) performance, the topic is subject to research in educational and sociological research. Educational research on the topic is often motivated by the fact that judgement formation is an integral component of teacher professional competence (see Baumert and Kunter 2006; Herppich et al. 2018), and the notion of an underlying assessment competence which is regarded as a learnable cognitive disposition. Teachers' professional decisions (e.g., related to adaptive selection of tasks, content and didactic approaches) rely on more or less explicit assessments of students' learning. Furthermore, teachers' placement and tracking decisions are even more dependent on accurate summative assessments.

Thus, a number of studies focus on judgement accuracy and its different operationalizations (see Südkamp et al. 2012 for a review) and study the effects of judgement accuracy (teachers' diagnostic competence) on students' learning gains, partly by taking adaptive instruction as a mediator into account (e.g., Anders et al. 2010; Helmke and Schrader 1987). Based on Esser and Kroneberg's (2015) Model of Frame Selection, sociological approaches often focus on social disparities and bias in teacher judgements (e.g., due to students' social background) and corresponding consequences on students. In a similar vein, Krolak-Schwerdt et al. (2013, see also Artelt 2016) studied teacher judgements according to the social cognition paradigm and argued, that modes of teachers' information processing and the accuracy of their judgements are dependent on the relevance of the according judgement.

Within BiKS-8-18 we were interested in the nature as well as in prerequisites and effects of teacher judgements and their accuracy. Research findings were based either on longitudinal data of the BiKS-8-18 project, or on a specific add-on study focusing on teachers' knowledge prerequisites. BiKS-8-18 panel data allowed us to study the phenomenon of judgement accuracy by focusing on different theoretically sound indices of accuracy as trait- or state-measures and also to analyze their effect on students' progress over the course of different school years in the respective domain. The add-on study focused (pre-service) teachers' content and pedagogical knowledge in the area of text comprehension. By combining these approaches, a significant contribution to the ongoing debate about the necessity of accurate teacher judgements and the impact of teacher judgements on students' further learning could be gained.

In the first part of the article, we will provide an overview of different indicators for measuring judgement accuracy. The second part will shed light on factors explaining inter-individual differences in teacher judgement accuracy. In the third part, we will illustrate major empirical findings from the BiKS-8-18 study on the effects of teachers' judgement accuracy on students' further achievements. And finally, we will shed lights on teachers' professional content-related knowledge for judgements on students' competencies in the domain of reading.

1 Measuring Teachers' Judgement Accuracy: Indicators, Structural Relations, and Descriptive Results

1.1 Indicators of Judgement Accuracy

For evaluating teachers' judgement accuracy, an appropriate criterion is needed. A common approach in the literature is to evaluate judgement accuracy by comparing teacher judgements related to student achievement with students' achievement scores in a standardized test in the respective domain (test as criterion). Research using students' results on a standardized test as a criterion thereby makes use of different indicators of judgement accuracy. The most common indicator measures rank-order accuracy. Here, teacher ratings on students' individual abilities are correlated with students' individual performance on a standardized test; a higher correlation implies higher judgement accuracy, indicating the degree to which the correct rank-order of students is reproduced. Rank-order correlations can be derived using different question types and stimuli. A common approach is a judgement type in which teachers are asked to judge students' academic achievement or general ability on a rating scale (e.g., ranging from poor to excellent). Other studies use teacher judgements of individual students' performance on particular items and tasks and compare these to the actual performance of the student on the specific tasks. Thus, judgement types differ with respect to their specificity: they can be task-specific or global. The corresponding tasks administered to the teachers differ in their demands. For task-specific judgements, more information is available to teachers. For global judgements, however, teachers have to infer which of the students' specific behaviors they consider relevant for their judgement. They also have to decide which detections of specific behavior they will integrate into their overall judgement.

Using data from BiKS-8-18, Karing et al. (2011a) investigated the accuracy of global and task-specific teacher judgements in the domain of reading. The sample consisted of 64 German language teachers and their fifth grade students. For global judgements, teachers were asked to make judgements about each of their students' general reading competence. For task-specific judgements, teachers were asked to predict whether each of the randomly selected seven students would pass or fail each single item on the reading competence test with seven multiple-choice items. A comparison of the accuracy of the global vs. the task-specific judgements was possible by computing class level correlations between the global teacher judgement and students' performance in the BiKS reading competence assessment for global ratings on the one hand and the correlation between task-specific judgements and students' actual performance on the individual items on the other hand. The class level correlations were computed as correlations between teacher judgement and student performance at the class level, and averaged across classes through Fisher-z-transformation. The task-specific judgement was formed by summing up the number of items that the teacher judged the student would pass (each coded as 1) and each student's performance was computed by summing up each student's correct answers (each coded as 1). The comparison revealed that teachers' judgement accuracy was rather low in general (global: $\overline{r} = 0.34$, task-specific: $\overline{r} = 0.20$). However, the rank-order component is higher for global judgements than for task-specific ones. A similar finding for the rank-order component in the domain of mathematics was reported by Karst et al. (2018; global judgements: $\overline{r} = 0.64$, specific judgements: $\overline{r} = 0.36$).

However, for the differentiation and level components the values were more accurate for the specific judgements than for the general judgements (Karst et al. 2018). Karst et al. (2018) argued that these findings might be due to differences in teachers' information processing (e.g., heuristic vs. controlled information processing) depending on the kind of judgement.

Based on the task-specific ratings, three other indicators of judgement accuracy could be derived for a detailed description of the computation of the indicators (see Karing et al. 2011a): The task-specific hit rate, the level component, and the differentiation component. The task-specific hit rate is the share of accurate judgements and was built by summing up the number of items for which a teacher's judgement and a student's actual performance were in agreement, divided by the number of items. In our study the task-specific hit rate of German language teachers was moderately high ($M = 0.66$), indicating that although the corresponding rank-order was rather low on the class level, teachers' individual ratings where higher. One explanation for the low correlation could be that the seven items of the reading test did not differentiate well between the students. Such a restriction of variance can depress correlations and therefore lead to an underestimation of the association between teacher judgements and students' actual test performance. Concerning teachers' estimation of the level of students' achievement—computed by subtracting the mean of the percentage of correct students' answers from the mean of the corresponding teacher judgements at the class level—our results ($M = 0.08$) are in line with the findings of the recent overview on judgement accuracy by Urhahne and Wijnia (2021): secondary school teachers tend to overestimate the level of students' achievement. The third task-specific indicator reflects the degree to which teacher ratings map the variation of students' achievement in their class. The corresponding differentiation component was calculated as the mean within-class quotient between the standard deviation of teacher judgement and the standard deviation of student actual performance, where a value of 1 shows a perfect judgement accuracy. T-tests for one sample were used to determine whether the calculated differentiation component differed significantly from the value 1. The results for our sample indicate that deviation of students' reading competence was accurately judged ($M = 0.94$, $t = 1.04$, $p > 0.05$). As could be shown previously for global teacher ratings (e.g., Spinath 2005), also for the task-specific ratings only low correlations could be found between the different components of judgement accuracy ($r = -0.06$ to 0.31). Judgement accuracy related to rank-order, level and deviation within a class is not a homogeneous construct. Teachers might score high when reproducing the rank order of the achievement level within the class but fail to estimate the level of performance or the variation within the class—or vice versa. Dependent on the task and the relevance of the according judgements, failure in one or the other judgement component is more or less problematic.

1.2 Dimensionality

Complementary to relations between judgement components (rank-order, level and differentiation component) and types of judgements (global vs. task-specific), we further investigated the structure of teachers' diagnostic competence by focusing on the dimensionality of the ability construct with respect to different areas of evaluations (e.g., students' achievement vs. motivation). Karing (2009) could show that for both primary and secondary school teachers, the judgement accuracy is higher for competence assessment than for the assessment of subject-specific interest. Moderate correlations were found in the cognitive domain (rank-order components: arithmetic: $\overline{r} = 0.52$ to 0.65; vocabulary: $\overline{r} = 0.44$ to 0.55; text comprehension: $\overline{r} = 0.40$ to 0.61), whereas low correlations were found in the non-cognitive domain (rank-order components: interest in mathematics: $\overline{r} = 0.32$ to 0.37; interest in German language arts: $\overline{r} = 0.21$ to 0.30). The BiKS-8-18 data further revealed only low correlations between secondary school teachers' judgements of children's test anxiety (worry, emotionality) and children's self-reported test anxiety in the school subjects of German (worry: $\overline{r} = 0.28$, emotionality: $\overline{r} = 0.27$) and mathematics (worry: $\overline{r} = 0.44$, emotionality: $\overline{r} = 0.20$; Karing et al. 2015). Similar findings were reported by Zhu and Urhahne (2021) who investigated Chinese elementary school teachers' judgement accuracy of sixth graders' mathematical competence, motivation (e.g., self-efficacy, self-concept, effort, expectancy for success), test anxiety, and interest in the subject of mathematics. They found that teachers predict students' mathematical competence with high accuracy, motivation with moderate to high accuracy, and test anxiety as well as interest with low accuracy. These studies provide further evidence that diagnostic competence is not a unidimensional ability construct that encompasses both the cognitive and the non-cognitive domain. Thus, a general ability to assess student characteristics cannot be assumed (c.f. Kolovou et al. 2021).

1.3 Stability of Diagnostic Competence

As far as the stability of diagnostic competence over time is concerned, there has been almost no research so far. However, the notion of relative continuity and stability of judgement accuracy (related to a specific domain) is an important feature for the assumption of a trait-like concept like competence. Using the longitudinal BiKS-8-18 data, the stability of teachers' judgement accuracy could be examined (Lorenz 2011; Lorenz and Artelt 2009): For primary school teachers, moderately high correlations were found between two measurement points within a six-month interval for judgements in the areas of mathematics ($r_{t1t2}=0.38$/ $r_{t2t3}=0.44$)[,12], vocabulary ($r_{t1t2}=0.57$/ $r_{t2t3}=0.49$)[1], text comprehension ($r_{t1t2}=0.51$/ $r_{t2t3}=0.47$)[1], joy of learning ($r_{t1t2}=0.42$), school attitude ($r_{t2t3}=0.47$)[1] and subject interest in language arts ($r_{t2t3}=0.40$)[1]. At least in the language domain, the stabilities are approximately at the same level as the judgement accuracies at the individual measurement points. The results offer some reason to assume that teachers' domain-related judgement accuracy can be considered as diagnostic competence. Recently, also Zhu and Urhahne (2021) reported a high stability of Chinese elementary school teachers' judgement accuracy of students' mathematical competence ($M_{t1}=0.70$ / $M_{t2}=0.74$), motivation (for different indicators: $M_{t1}=0.42$ to 0.65 / $M_{t2}=0.0.42$ to 0.61), test anxiety ($M_{t1}=0.28$ / $M_{t2}=0.16$), and interest in the subject of mathematics ($M_{t1}=0.33$ / $M_{t2}=0.35$) over a 4-week interval.

2 Explaining Inter-Individual Differences: Context and Teacher Effects

Throughout the literature, differences between teachers in the accuracy of their judgements are frequently reported (for an overview see Kaufmann 2020; Südkamp et al. 2012). Theoretical considerations and empirical findings suggest that inter-individual differences in the accuracy of teacher judgements of students' academic achievement can be explained by multiple characteristics of teachers, students, judgements, and tests (Südkamp et al. 2012).

In BiKS-8-18, characteristics of teachers, students, classes (i.e., context for social comparisons), and domain of the judgement (mathematics and reading) were examined as moderators of judgement accuracy. Although it is often

[1] r_{t1t2}: t1 = second half of third grade, t2 = first half of fourth grade for all available teachers (n = 125 for mathematics and n = 128 for vocabulary and text comprehension), r $_{t2t3}$: t2 = first half of fourth grade, t3 = second half of fourth grade for all available teachers (n = 118 for mathematics and n = 130 for vocabulary and text comprehension, n = 129 for joy of learning and interest in language arts and n = 129 for school attitude).

questioned whether judgement accuracy related to students' academic achievement can be regarded as a general ability such as diagnostic competence (e.g., Kolovou et al. 2021; Spinath 2005), theoretically it can still be justified that the accuracy of diagnostic judgements relates to certain teacher characteristics. Specifically, experience on the job and other variables related to teacher expertise (for results on relevant professional knowledge, see below) are potential candidates for such characteristics. For primary school teachers, Lorenz (2011) investigated the correlation between several teacher characteristics and their judgement accuracy. However, reliable correlations could not be found for any of these variables (professional experience [years on the job], gender, teaching time in the respective class, self-reported ability to perspective taking, attitudes towards and assessment of one's own diagnostic competence, striving for perfection, extent of further training). For the indicators of judgement accuracy used in the BiKS study (rank-order component, differentiation component, level component; see above), none of the teacher variables that were considered to be potentially influential made a difference.

However, there are further factors that are likely to impact teacher judgements. Referring to social cognition research and frame selection models, theoretically important student characteristics in this regard are personality traits, gender, immigration background or socio-economic status. It is argued, e.g., in the continuum model of impression formation (Fiske and Neuberg 1990) that people rely on social categories for judgement formation whenever possible, potentially leading to biased judgements. In this respect, a judgement is considered to be biased, if teachers systematically evaluate "two groups as differing on some criterion more or less than they really do differ" (Jussim et al. 1996, p. 329), based on variables that are irrelevant to the criterion. As for personality traits, Westphal et al.'s (2021) findings indicate that students' conscientiousness positively influences teacher judgement accuracy in mathematics. Bonefeld et al. (2020) found preservice teachers' judgements to be less favorable for students with (vs. without) immigration background and for female (vs. male) students in a virtual classroom setting. In the BiKS-8-18 study, primary school teacher judgements on students' competence in the domains of mathematics and vocabulary were also found to be significantly different depending on the students' gender. Student test performance, however, did not systematically differ between boys and girls, indicating a gender bias in teacher judgements (Lorenz and Artelt 2009). Lorenz (2011) was also able to show social disparities in primary education: students from the upper half of the distribution of socio-economic status (HISEI) were overestimated or underestimated to a lesser extent than students from the lower half of the social distribution. However, these gender-specific and social status-related differences in teacher judgements in the domain of reading could not be found in the BIKS-data for the teachers on secondary school level (Karing et al. 2011a, b).

Social status-related differences can also be interpreted in the light of similarity of habitus between teacher and student, indicating that a higher distance in social status between teacher and student might lead to lower expectations and teacher disaffection (e.g., Alexander et al. 1987). With this in mind, it also seems likely that also other aspects of similarity between students and teachers might play a role for the accuracy of teacher judgements. Rausch et al. (2016) showed that the similarity of personality traits between a student and their teacher can have a small, but significant influence on the accuracy of teacher judgements. For 168 teacher-student dyads in German language classes and 241 dyads in mathematics classes (each at the end of Grade 8), a similarity index between student and teacher was calculated for every student. Multiple regression models were run to map the judgement bias that can be attributed to personality similarity. Results indicate, that while students being more similar to their respective teacher did not show higher performance in the domains of mathematics and reading, the more similar students were judged more positively than students who were less similar. While this holds true for global judgements in both domains, this could not be replicated for task-specific judgements, where personality similarity did not have a significant impact on the judgement. One explanation could be that teachers focus more on individuating information about the students or on the task when conducting a task-specific judgement.

Not only individual characteristics of teachers and students—and their interaction—can play a role in the accuracy of judgements, but also characteristics aggregated at class level. The average performance level and the heterogeneity of the performance in the class were also considered to be theoretically significant factors influencing the judgement. High achievement heterogeneity in the class was found positively related to accuracy of teacher judgement in both primary and secondary schools (Karing 2009). Both high performance heterogeneity and a wide dispersion of subject interest led to a better discriminability of students with regard to these characteristics. This can have a facilitating effect for teachers (Karing 2009). Although these results may be also a statistical artefact because extreme values influence the correlation coefficients (e.g., rank component).

Considerable differences with regard to the heterogeneity of the performance in the class, teacher training, the class teacher principle vs. subject teacher principle, cooperation between teachers and cooperation with parents suggest that there are also differences between primary schools and secondary schools (Gymnasium) with regard to judgement accuracy. As compared to secondary school (Gymnasium) teachers, primary school teachers show higher accuracy in assessing student performance in the areas of mathematics (d = 0.60), vocabulary (d = 0.55), and text comprehension (d = 0.58), as well as subject-specific interest in German language arts (d = 0.30), but not in mathematics (d = 0.18) (Karing 2009). However, it is not possible to distinguish whether the effect is rooted in differences in the homogeneity of student achievement in classes (lower homogeneity in primary school classes; cf. Tillmann and Wischer 2006) and/or differences in teacher training between the two school forms.

3 Does the Accuracy of Teacher Judgements Affect Students' Learning?

The longitudinal design of the BiKS-8-18 study also made it possible to investigate the effects of teachers' judgement accuracy on students' competence development. Despite the high plausibility of the assumption of a positive correlation between the accuracy of judgements and the development of students' performance, only a few empirical studies have been conducted in this area (Anders et al. 2010; Behrmann and Souvignier 2013; Helmke and Schrader 1987; Karst et al. 2014; Urhahne 2015). In the BiKS research group, this question was examined again by investigating the effects of the accuracy of global and task-specific judgements on performance development in the competence areas of mathematics and text comprehension (Karing et al. 2011a, b). Longitudinal data were obtained from a sample of 502 students and their 40 German language teachers and 29 mathematics teachers (measurement points: at the end of grades 5 and 6). Nearly 80% of the students attended higher academic track schools. The multilevel analyses conducted for this purpose showed a significant positive effect of the task-specific hit rate on the students' development between the fifth and sixth grade of reading competence, but not on the development of the mathematical competence. In addition, the correlation was moderated by instructional variables such as teachers' use of structural cues (e.g., teacher summarizes the lessons that students can remember the gist) and the degree to which lessons were individualized. A high task-specific hit rate in combination with a high degree of individualization of lessons is related to an increased development in students' reading competence, whereas a high task-specific hit rate in combination with a low degree of individualization of lessons had no effect on the development of students' reading competence. Moreover, a high task-specific hit rate in combination with a low frequency of structural cue use during lessons is also associated with an increase in the development of students' reading competence. However, no association was found when structural cues were used frequently. One explanation for this unexpected result could be that because of their learning strategies and prerequisites, high-ability students rely more on self-directed learning and individualized instructions instead of being dependent on teachers' use of structural cues during lessons. However, for low-ability students a highly structured learning environment makes it easier for them to focus their attention on relevant aspects of the lessons and combine prior knowledge with new knowledge (Blumberg et al. 2004; Lipowsky 2009). Similar findings were reported by Möller et al. (2002). The authors interpret their findings in the sense that high-ability students did not require highly structured learning environments in science and social studies for their learning gains, while low-ability students profited more from a highly structured lesson.

For the rank-order component of judgement accuracy, no significant positive association or interactions between this indicator and any instructional variable were found in the domains of reading and mathematics. A possible explanation might be the low value of the rank-order component (reading: $\overline{r} = 0.19$, mathematics: $\overline{r} = 0.44$). Thus, no significant associations with the development of reading and mathematical competence could be identified. This corresponds with the assumption that a minimal degree of judgement accuracy and instructional quality is required for a significant relationship or interactions (Schrader 1989). Further, few studies reported that the effect of accurate teacher judgements on student achievement is moderated (e.g., Behrmann and Souvignier 2013; Karing et al. 2011b) or mediated (e.g., Anders et al. 2010) by instructional activities (e.g., frequent feedback, individualization of lessons).

4 Knowledge Base

Teacher judgements on student achievement are usually embedded in school subjects or even more specific fields of study within the respective subject domains. Within the BiKS-8-18 study, teacher judgement accuracy was also assessed related to specific subject or competence domains respectively. This also reflects the notion of a content-specificity of judgement accuracy (Kolovou et al. 2021). However, the open question remains, if, how, and to what extent teachers' domain-specific knowledge base affects the accuracy of teacher judgements in that very domain. From a theoretical point of view, it can be argued, that knowledge about the nature and demands of a domain are positively associated with accurate judgements of student performance (cf. Artelt 2016). This also implies that domain-specific cues need to be recognized and considered in the judgement process, to judge student performance effectively and accurately (cf. National Institute of Child Health and Human Development 2000). Using an additional sample of pre-service-teachers, teachers, and expert teachers within the BiKS-8-18 study, a knowledge test was developed in order to assess teachers' knowledge in the domain of reading. Individual student performance on a specific task is dependent not only on students' characteristics, such as their prior knowledge, intelligence, and motivation, but also on the quality and composition of materials (such as texts), demands of the (text-specific) tasks and the activities required by the student to perform the task (such as the use of reading strategies appropriate to the text and task demands (Campione and Armbruster 1985, see also Artelt et al. 2005). Specific teacher judgements on student performance in a given situation need to take into account all these aspects in order to deliver an accurate judgement. Therefore, the test aims at covering teachers' knowledge base on text and item difficulties as well as on cognitive processes of text comprehension. Thus, the test comprised the three dimensions of text characteristics, item characteristics, and (necessary) reading strategies. It was administered to teachers with German language arts as their main teaching subject. From these teachers ($n = 44$), global and task-specific assessments of individual student performance ($n = 233$ students) were also available. No significant correlations between teacher knowledge in the domain of reading (total test score as well as all subscores on text characteristics, task characteristics, and reading strategies) and the accuracy of student assessments on reading tasks (global and task-specific judgements) were found (Rausch et al. 2015). Other research for the domain of biology shows that pre-service teachers' professional knowledge was related to diagnostic activities but could not generally be related to diagnostic accuracy (Kramer et al. 2021). The results seem to indicate that while domain-specific knowledge might play a role in the judgement process, other more distal cues or heuristics might have more impact on judgement accuracy (see research on influence of gender or immigration status on judgement accuracy, e.g., Bonefeld et al. 2020). The BiKS study was able to contribute to the research body on how teachers' domain-specific knowledge influences judgement processes and thus judgement accuracy. In the context of judgement processes in educational settings, it remains an open question, what elements of teacher knowledge are crucial in the judgement process, and how these domain-specific aspects are overlayed for example by (simple) heuristics in concrete judgement situations.

5 Summary and Outlook

Our research was dedicated to questions on judgement accuracy of teachers at primary and secondary level. Especially through the recording of diagnostic competence over several measurement points as well as through the broad recording of cognitive and emotional-motivational variables and the corresponding teacher assessments, contributions could be made to the further scientific clarification of conditions, structure and effects of diagnostic competence. New insights were also gained with the development of a test on teachers' knowledge bases in the area of text comprehension.

Primary school teachers are relatively good at assessing student performance in the performance areas, whereas they find it much more difficult to correctly assess students' subject interest. In contrast, teacher judgements in lower secondary school show rather low correlations with students' reading competence, both in global and in task-specific judgements. Global judgements are more accurate than task-specific judgements. For task-specific judgements, it can be seen that German language teachers accurately assess the dispersion of reading competence, while on average they overestimate the level of reading competence.

With regard to the structure of diagnostic competence, it could be shown in our studies that it is a subject-related ability that is relatively persistent over time. On the other hand, a general ability to assess students' characteristics across subject areas cannot be assumed. Likewise, it is not a homogeneous construct in the sense of a competence that is evident across the different judgement components.

Our results also show that the repeatedly reported large inter-individual differences in the accuracy of judgements between teachers can be attributed to a complex set of conditions consisting of characteristics of the teacher, the students, the class and the subject of judgement. Influencing factors in primary education are primarily to be found at class and student level. The dispersion of performance in the class has a considerable influence on the accuracy of judgements, as does gender and the students' social status. This correlation—an indication of bias—was not found for teachers at the secondary level.

Furthermore, there is a positive correlation between the task-specific hit rate and the development of reading competence in lower secondary school. This is particularly evident when the degree of individualization of instruction is high and when fewer structuring aids are offered in the classroom.

For the work related to teachers' knowledge base for judgements in the area of text comprehension, it can be summarized that we did find differences between more or less experienced teachers, but not for all task formats. If there were differences, they were mostly in favor of experts who scored more positively. In accordance with the question about the connection between knowledge and accurate assessments, it should be noted that we did not find any significant correlations, i.e., there is no connection on the basis of different components.

We also found that teacher judgement accuracy varies as a function of the respective achievement/competence domain under study (e.g., students' mathematical or reading competence, see Artelt and Rausch 2014 for a discussion), between competence and motivational characteristics of students (see also Spinath 2005) and—to some degree—as a function of features and demands of the judgement process. However, there is also reason to assume that there is a latent trait (or disposition) rooted in professional knowledge and expertise. As such, assessment competence is regarded as a learnable and situation-specific disposition (c.f. Herppich et al. 2018) and judgement accuracy is only a quantifiable product of this disposition.

To further elaborate on this, theories and research systematically need to consider the role of contextual factors like judgement purposes and demands. As demonstrated by Herppich et al. (2018), models need to integrate research on assessment processes, practices, and products.

Rooted in Rosenthal and Jacobson's (1968) much-cited work on expectancy effects, research not only focused the effects of non-accurate (but rather positive) teacher judgements, assuming that they impact teachers' learning fostering behavior and feedback (e.g., Gentrup et al. 2020). However, being accurate or overestimating students seems to follow different functions. Whereas accuracy seems to be important for formative assessments, since students profit most from receiving accurate feedback on their actual performance, including mistakes, (unrealistic) positive judgements operate on students' motivation and self- concept and might thereby foster learning. Further, Urhahne and Wijnia (2021) already criticized in their review on teachers' judgement accuracy that the majority of the studies investigated judgement accuracy only at one measurement point. Thus, the BiKS-8-18 panel data is unique in undertaking longitudinal measurements of teachers' judgements and students' performance. It allowed us to study the effects of teacher judgement accuracy on students' progress over the course of different school years in different cognitive and emotional-motivational domains. Hence, future research should use longitudinal designs to investigate teacher judgement accuracy in order to draw more robust conclusions, e.g., on the effects of judgement accuracy on student performance or on aspects of professional development of teachers' diagnostic competence from teacher education to the classroom. The findings can also inform teacher education, especially in subject-related didactics.

Acknowledgements Acknowledgment is given to the German Research Foundation (DFG) for the funding as a DFG research unit (FOR 543; funding numbers of the included projects: RO 820/12 (1-4), RO 820/11 (1-4), WE 1478/4 (1-4), AR 301/9 (1-2), BL 381/3 (1-4), FA 650/1 (1-4), BL 381/7 (1-3), AR 301/6 (1-3)), and for DFG funding of the subsequent BiKSplus projects (funding numbers RO 820/15-1, WE 1478/8-1, FA 650/3-1, and RE 3553/1-1, BU 2553/5-1, AR 301/10-1) as well as to the Federal Ministry for Education and Research for funding of a further subsequent project and assessment wave of BiKS-3-18 (funding number B8578).

We would like to sincerely thank all participants as well as all those who contributed to the survey and evaluations for their great commitment.

References

Alexander, K. L., Entwisle, D. R., & Thompson, M. S. (1987). School Performance, Status Relations, and the Structure of Sentiment: Bringing the Teacher Back. *American Sociological Review, 52*(5), 665–682. https://doi.org/10.2307/2095602.

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. *Psychologie in Erziehung und Unterricht, 57*, 175–193. https://doi.org/10.2378/peu2010.art13d.

Artelt, C., McElvany, N., Christmann, U., Richter, T., Groeben, N., Köster, J., Schneider, W., Stanat, P., Ostermeier, C., Schiefele, U. Valtin, R., & Ring, K. (2005). *Expertise – Förderung von Lesekompetenz*. BMBF.

Artelt, C. (2016). Teacher Judgments and their Role in the Educational Process. In R. A. Scott & S.M. Kosslyn (Eds.), *Emerging Trends in the Social and Behavioral Sciences. An Interdisciplinary, Searchable, and Linkable Resource* (p. 1–16). Wiley. https://doi.org/10.1002/9781118900772.etrds0402.

Artelt, C., & Rausch, T. (2014) Accuracy of teacher judgments. When and for what reasons? In S. Krolak-Schwerdt, S. Glock & M. Böhmer (Eds.), *The future of educational research: Vol. 3. Teachers' professional development: Assessment, training and learning*. (p. 27–43). Sense Publishers.

Baumert, J., & Kunter, M. (2006) Stichwort: Professionelle Kompetenz von Lehrkräften. *Zeitschrift für Erziehungswissenschaft, 9*(4), 469–520. https://doi.org/10.1007/s11618-006-0165-2.

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

Blumberg, E., Möller, K., & Hardy, I. (2004). Erreichen motivationaler und selbstbezogener Zielsetzungen in einem schülerorientierten naturwissenschaftsbezogenen Sachunterricht - Bestehen Unterschiede in Abhängigkeit von der Leistungsstärke? In W. Bos, E. Lankes, N. Plaßmeier & K. Schwippert (Eds.), *Heterogenität. Eine Herausforderung an die empirische Bildungsforschung* (p. 41–55). Waxmann.

Bonefeld, M., Dickhäuser, O., & Karst, K. (2020). Do preservice teachers' judgments and judgment accuracy depend on students' characteristics? The effect of gender and immigration background. *Social Psychology of Education, 23,* 189–216. https://doi.org/10.1007/s11218-019-09533-2.

Campione, J. C., & Armbruster, B. B. (1985). Acquiring information from texts: An analysis of four approaches. In S. Chipman, J. Segal & R. Glaser (Eds.), *Thinking and learning skills: Relating instruction to basic research* (Vol. 1, p. 317–359). Erlbaum.

Esser, H., & Kroneberg, C (2015). An Integrative Theory of Action: The Model of Frame Selection. In E. J. Lawler, S. R. Thye & J. Yoon (Eds.), *Order on the Edge of Chaos: Social Psychology and the Problem of Social Order* (p. 63–85). University Press.

Fiske, S. T., & Neuberg, S. L. (1990). A Continuum of Impression Formation, from Category-Based to Individuating Processes: Influences of Information and Motivation on Attention and Interpretation. *Advances in Experimental Social Psychology, 23,* 1–74. https://doi.org/10.1016/S0065-2601(08)60317-2

Gentrup, S., Lorenz, G., Kristen, C., & Kogan, I. (2020). Self-fulfilling prophecies in the classroom: Teacher expectations, teacher feedback and student achievement. *Learning and Instruction, 66*, 1–17. https://doi.org/10.1016/j.learninstruc.2019.101296

Helmke, A., & Schrader, F.-W. (1987). Interactional effects of instructional quality and teacher judgement accuracy on achievement. *Teaching and Teacher Education, 3*, 91–98.

Herppich, S., Praetorius A.-K., Förster, N., Glogger-Frey, I., Karst, K., Leutner, D., Behrmann, L., Böhmer, M., Ufer, S., Klug, J., Hetmanek, A., Ohle, A., Böhmer, I., Karing, C., Kaiser, J., & 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

Jussim, L., Eccles, J., & Maddon, S. (1996). Social perception, social stereotypes, and teacher expectations: Accuracy and the quest for the powerful self-fulfilling prophecy. *Advances in Experimental Social Psychology*, 29, 281–388.

Karing, C. (2009). Diagnostische Kompetenz von Grundschul- und Gymnasiallehrkräften im Leistungsbereich und im Bereich Interessen. *Zeitschrift für Pädagogische Psychologie, 23*, 197–209. https://doi.org/10.1024/1010-0652.23.34.197

Karing, C., Dörfler, T., & Artelt, A. (2015). How accurate are teacher and parent judgements of lower secondary school children's test anxiety. *Educational Psychology, 35*, 909–925. https://doi.org/10.1080/01443410.2013.814200

Karing, C., Matthäi, J., & Artelt, C. (2011a). Genauigkeit von Lehrerurteilen über die Lesekompetenz ihrer Schülerinnen und Schüler in der Sekundarstufe I - Eine Frage der Spezifität? *Zeitschrift für Pädagogische Psychologie, 25*, 159–172. https://doi.org/10.1024/1010-0652/a000041

Karing, C., Pfost, M., & Artelt, C. (2011b). Hängt die diagnostische Kompetenz von Sekundarstufenlehrkräften mit der Entwicklung der Lesekompetenz und der mathematischen Kompetenz ihrer Schülerinnen und Schüler zusammen? *Journal for Educational Research Online, 3*, 121–149. https://doi.org/10.25656/01:5626

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., Schoreit, E., & Lipowsky, F. (2014). Diagnostische Kompetenzen von Mathematiklehrern und ihr Vorhersagewert für die Lernentwicklung von Grundschulkindern. *Zeitschrift für Pädagogische Psychologie, 28*, 237–248. https://doi.org/10.1024/1010-0652/a000133.

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

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

Kramer, M., Förtsch, C., Boone, W. J., Seidel, T., & Neuhaus, B. J. (2021). Investigating pre-service biology teachers' diagnostic competences: relationships between professional knowledge, diagnostic activities, and diagnostic accuracy. *Education Sciences, 11*(3), 89. https://doi.org/10.3390/educsci11030089

Krolak-Schwerdt, S., Böhmer, M., & Gräsel, C. (2013). The impact of accountability on teachers' assessments of student performance: A social cognitive analysis. *Social Psychology of Education*, *16*, 215–239. https://doi.org/10.1007/s11218-013-9215-9

Lipowsky, F. (2009). Unterricht. In E. Wild & J. Möller (Eds.), *Pädagogische Psychologie* (p. 73–102). Springer.

Looney, J. (2011). Developing High-Quality Teachers: teacher evaluation for improvement. *European Journal of Education, (46),* 440–455. https://doi.org/10.1111/j.1465-3435.2011.01492.x

Lorenz, C. (2011). *Diagnostische Kompetenz von Grundschullehrkräften. Strukturelle Aspekte und Bedingungen.* University of Bamberg Press.

Lorenz, C., & Artelt, C. (2009). Fachspezifität und Stabilität Diagnostischer Kompetenz von Grundschullehrkräften in den Fächern Deutsch und Mathematik. *Zeitschrift für Pädagogische Psychologie, 23*(3-4), 211–222. https://doi.org/10.1024/1010-0652.23.34.211

Möller, K., Jonen, A., Hardy, I., & Stern, E. (2002). Die Förderung von naturwissenschaftlichem Verständnis bei Grundschulkindern durch Strukturierung der Lernumgebung. In M. Prenzel & J. Doll (Eds.), *Bildungsqualität von Schule: Schulische und außerschulische Bedingungen mathematischer, naturwissenschaftlicher und überfachlicher Kompetenzen* (p. 176–191). Beltz. https://doi.org/10.25656/01:3946

National Institute of Child Health and Human Development, NIH, DHHS. (2000). *Report of the National Reading Panel: Teaching Children to Read: Reports of the Subgroups* (00-4754). U.S. Government Printing Office.

OECD (2012). Education at a Glance 2012: Highlights. OECD Publishing. https://doi.org/10.1787/eag_highlights-2012-en

Rausch, T., Karing, C., Dörfler, T., & Artelt, C. (2016). Personality Similarity between Teachers and their Students Influences Teacher Judgment of Student Achievement. *Educational Psychology, 36*(5), 863–878. https://doi.org/10.1080/01443410.2014.998629

Rausch, T., Matthäi, J., & Artelt, C. (2015). Mit Wissen zu akkurateren Urteilen? Zum Zusammenhang von Wissensgrundlagen und Urteilsgüte im Bereich des Textverstehens. *Zeitschrift für Entwicklungspsychologie und Pädagogische Psychologie, 47,* 147–158. https://doi.org/10.1026/0049-8637/a000124

Rosenthal, R., & Jacobson, L., (1968). Pygmalion in the classroom. *The Urban Review, 3,* 16–20

Schrader, F.-W. (1989). *Diagnostische Kompetenzen von Lehrern und ihre Bedeutung für die Gestaltung und Effektivität des Unterrichts*. Peter Lang.

Spinath, B. (2005). Akkuratheit der Einschätzung von Schülermerkmalen durch Lehrer und das Konstrukt der diagnostischen Kompetenz. *Zeitschrift für Pädagogische Psychologie, 19*, 85–95. https://doi.org/10.1024/1010-0652.19.12.85

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

Tillmann, K.-J., & Wischer, B. (2006). Heterogenität in der Schule. Forschungsstand und Konsequenzen. *Pädagogik, 3,* 44–48.

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

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

Westphal, A., Lazarides, R., & Vock, M. (2021). Are some students graded more appropriately than others? Student characteristics as moderators of the relationships between teacher-assigned grades and test scores in mathematics. *British Journal of Educational Psychology, 91*, 865–881. https://doi.org/10.1111/bjep.12397

Zhu, C., & Urhahne, D (2021). Temporal stability of teachers' judgment accuracy of students' motivation, emotion, and achievement. *European Journal of Psychology of Education*, 36(2), 319–337. doi: https://doi.org/10.1007/s10212-020-00480-7