Hertel, Sophia; Urton, Karolina; Wilbert, Jürgen; Krull, Johanna; Bosch, Jannis; Hennemann, Thomas

Teachers' judgment accuracy of students' reading comprehension in inclusive primary schools

Empirische Sonderpädagogik 16 (2024) 4, S. 297-314

Quellenangabe/ Reference:
Hertel, Sophia; Urton, Karolina; Wilbert, Jürgen; Krull, Johanna; Bosch, Jannis; Hennemann, Thomas: Teachers' judgment accuracy of students' reading comprehension in inclusive primary schools - In: Empirische Sonderpädagogik 16 (2024) 4, S. 297-314 - URN: urn:nbn:de:0111-pedocs-340228 - DOI: 10.25656/01:34022; 10.2440/003-0034

https://nbn-resolving.org/urn:nbn:de:0111-pedocs-340228
https://doi.org/10.25656/01:34022

in Kooperation mit / in cooperation with:
Pabst Science Publishers*https://www.psychologie-aktuell.com/journale/empirische-sonderpaedagogik.html*

Empirische Sonderpädagogik, 2024.16:297-314
DOI https://doi.org/10.2440/003-0034
ISSN 1869-4845 (Print) · ISSN 1869-4934 (ebook)

Teachers' Judgment Accuracy of Students' Reading Comprehension in Inclusive Primary Schools

Sophia Hertel[1], Karolina Urton[2], Jürgen Wilbert[2], Johanna Krull[1], Jannis Bosch[2] & Thomas Hennemann[1]

[1] University of Cologne
[2] University of Münster

Abstract
Reading comprehension is crucial in primary education. Yet a quarter of German fourth-graders, especially those with special educational needs (SEN), struggle with it. Teachers need diagnostic abilities to provide tailored support, but previous studies have identified limitations in how teachers assess low and average achievers and their tendency to consider irrelevant characteristics. This study examines the accuracy with which teachers in inclusive classrooms assess students' reading comprehension at word, sentence and text levels and the extent to which student characteristics such as grade level, SEN and sex influence reading levels. The reading comprehension of 1,693 students with and without SEN was assessed using a standardized test and rated by 102 teachers. Using a multilevel analysis, we examined the degree to which the teachers' assessments corresponded to the students' standardized test performance and the extent to which this was related to the students' characteristics. Results showed moderate correlations between teachers' ratings and students' reading performance at sentence and text levels, and low correlations at word level. The accuracy of judgments varied greatly among teachers and judgment accuracy at sentence and text levels increased as grade level increased. There was no effect for sex. SEN in learning was associated with lower accuracy in teachers' assessments, as was SEN in emotion and behaviour but only on word level. Low-achieving students were assessed less accurately, although they are in particular need of tailored support in reading. As teachers must be able to identify struggling students and monitor their development, this highlights the need for further research into the characteristics of teachers with higher diagnostic skills.

Keywords: special educational needs, inclusive education, reading comprehension, accuracy of teachers' judgement

Die Urteilsgenauigkeit von Lehrkräften bei der Beurteilung des Leseverständnisses von Schüler*innen in inklusiven Grundschulen

Zusammenfassung
Die Entwicklung eines guten Leseverständnisses ist insbesondere in der Grundschulbildung von entscheidender Bedeutung. Dennoch hat ein Viertel der deutschen Viertklässler*innen, insbesondere diejenigen mit sonderpädagogischem Förderbedarf (SPF), Schwierigkeiten damit. Lehrkräfte benötigen diagnostische Kompetenzen, um eine bedarfsgerechte Förderung anzubieten. Bisherige Studien haben gezeigt, dass Lehrkräfte bei der Beurteilung von Schüler*innenleistungen im geringen und mittleren Bereich nur begrenzt in der Lage sind diese genau zu beurteilen und dazu neigen, irrelevante Merkmale zu berücksichtigen. Daher untersucht diese Studie, wie genau Lehrkräfte in inklusiven Klassen das Leseverständnis ihrer Schüler*innen auf Wort-, Satz- und Textebene beurteilen und inwieweit bestimmte Schüler*innenmerkmale wie die Klassenstufe, der SPF und das Geschlecht das Leselevel beeinflussen. Das Leseverständnis von 1693 Schüler*innen mit und ohne SPF wurde anhand eines standardisierten Tests bewertet und von 102 Lehrkräften beurteilt. Mittels einer Mehrebenenanalyse wurde untersucht, inwieweit die Bewertungen der Lehrkräfte mit den standardisierten Testergebnissen der Schüler*innen übereinstimmten und inwieweit dies mit den Schüler*innenmerkmalen zusammenhing. Die Ergebnisse zeigten moderate Korrelationen zwischen den Bewertungen der Lehrer*innen und den Leseleistungen der Schüler*innen auf Satz- und Textebene und geringe Korrelationen auf Wortebene. Die Urteilsgenauigkeit variierte stark zwischen den Lehrkräften und stieg auf Satz- und Textebene mit der Klassenstufe, während kein Einfluss des Geschlechts festgestellt wurde. Ein SPF Lernen war mit einer geringeren Urteilsgenauigkeit der Lehrkräfte verbunden, was, jedoch nur auf Wortebene, auch für den SPF im Bereich Emotionale und soziale Entwicklung festgestellt wurde. Schüler*innen mit geringen Leseverständnisleistungen wurden ungenauer beurteilt, obwohl sie besonders auf eine passgenaue Leseförderung angewiesen sind. Da Lehrkräfte in der Lage sein müssen, Schüler*innen mit Schwierigkeiten zu identifizieren und ihre Entwicklung zu beobachten, zeigt dies den Bedarf an weiterer Forschung zu den Merkmalen von Lehrkräften mit besseren diagnostischen Fähigkeiten.

Schlüsselwörter: Sonderpädagogischer Förderbedarf, inklusive Bildung, Leseverständnis, Urteilsgenauigkeit, Lehrkräfte

Students’ reading competence

Reading competence is a fundamental skill that students need to acquire during their school years as it is of high importance for success in later life (Rambuyon & Susada, 2022; Schmitterer & Brod, 2021) as competent readers are able to orient themselves more reliably amid the great amount of information now available (Golly Ledoux et al., 2023; Grigoryan, 2020; Rambuyon & Susada, 2022). Furthermore, good reading literacy paves the way for further education, enables the expansion of a wide variety of skills, and represents the most important means of transmitting information and communicating (Grigoryan, 2020). A lack of reading comprehension leads to a wide range of problems such as failing school subjects, behavioural issues, dropping out of school, and an increased risk of unemployment and poverty after school (K. J. Bennett et al., 2003; Bos et al., 2017).

Reading comprehension can generally be defined as “a cognitive activity in which readers construct a coherent and integrated mental representation of the text content" (Golly Ledoux et al., 2023, p. 20). Specifically, fourth-grade students with proficient reading comprehension skills "should be able to demonstrate an overall understanding of the text [… and] extend the ideas in the text by making inferences, drawing conclusions, and making connections to their own experiences" (National Assessment Governing Board, 2007, p. 24). Reading is a complex process comprised of many sub-skills that are intertwined with each other and multiple skills are required to develop proficient reading comprehension (Mendoza-Pinargote & Reyes-Meza, 2022; Rambuyon & Susada, 2022). In addition to prior knowledge and general cognitive abilities (such as working memory capacity), fluency decoding words (word level), fluency decoding syntactic structures (sentence level), and the ability to read across sentences coherently (text level) are important sub-skills that increase in complexity and contribute significantly to reading comprehension performance (Lenhard, 2013). Word comprehension relates to decoding and synthesis, while sentence comprehension relates to comprehension and syntactic skills, and text comprehension relates to information retrieval, cross-sentence reading and inferential thinking. Confidence in understanding a text arises when all these sub-processes work together smoothly, and primary education is particularly important for acquiring these reading comprehension skills at all (word, sentence and text) levels (Lenhard, 2013; Mendoza-Pinargote & Reyes-Meza, 2022).

In terms of students' reading skills, the Progress in International Reading Literacy Study Survey (PIRLS) showed that between 2001 and 2016, the share of fourth-graders in Germany with low comprehension skills increased from 17 to 19 per cent (Bos et al., 2017). This trend has continued, and a quarter of fourth graders now have low-level reading skills (McElvany et al., 2023). In addition, by ratifying the Convention on the Rights of Persons with Disabilities (article 24, The United Nations, 2006), Germany committed itself to implementing an inclusive education system at all levels. Therefore, more students with special educational needs (SEN) are being educated in general educational elementary schools. Unfortunately, it is particularly evident that when compared to students with learning, developmental, cognitive or physical disabilities (Erickson & Geist, 2016; Noll et al., 2018), many students with emotional behavioural disorders (EBD) have significant deficits in academic achievement (Reid et al., 2004) and often experience reading difficulties (Garwood et al., 2014; Lane et al., 2008).

Teachers' judgment accuracy

Teachers play a key role in developing their students' reading skills and "matter more to student achievement than any other school-related factor" (International Literacy Association, 2019, p. 1). Accordingly, Hudson (2022) demonstrated that there is a significant correlation between upper elementary teachers' knowledge and higher reading comprehension achievement in students. Therefore, accurate assessments of students' performance are an important prerequisite for a variety of instructionally relevant decisions affecting, for example, feedback design and the adaptation of instruction methods (Praetorius et al., 2013; Ready & Wright, 2011). Teachers' judgment accuracy, defined as "the correlation between teachers' judgments of students' academic achievement and students' actual test performance" (Südkamp et al., 2012, p. 755), is related to the development of students' reading comprehension (Behrmann & Souvignier, 2013; Urhahne & Wijnia, 2021). The judgments can be considered in terms of both the assessment outcome and the assessment process. One framework that addresses the cognitive processes underlying diagnostic judgments is called DiaCoM (Explaining Teachers' Diagnostic Judgements by Cognitive Modeling; Loibl et al., 2020). In addition to the diagnostic behaviour of teachers, including their assessment of reading comprehension, the external components of the model include situational characteristics. These provide the context for the diagnostic process and contain information that the teacher can use to draw conclusions about, among other things, the reading comprehension performance of their students (Leuders et al., 2020). This information can be determined using cues. Cues are the central element of the lens model, a general conceptual framework for explaining and analysing judgment accuracy and the processes involved in it (Back & Nestler, 2016; Brunswik, 1956). It is based on the assumption that people infer characteristics that are not directly perceptible or observable by focusing on perceptible cues or signals, such as behaviours and external features (Brunswik, 1956; Nestler & Back, 2013). While, for example, a learner's reading comprehension is not directly observable, learners' reading-related behaviour can be observed and, on this basis, reading comprehension can be inferred reasonably accurately (Förster & Böhmer, 2017). Thus, when Südkamp et al. (2012) conducted a meta-analysis on the accuracy of teachers' judgments with a focus on product indicators of diagnostic behaviour, their results indicated that teachers can assess students' academic performance quite accurately on average (r = .63). However, there is also evidence that assessment performance can vary widely from teacher to teacher (Gabriele et al., 2016; Wilbert et al., 2020).

Teachers' assessment of students' reading comprehension

Paleczek et al. (2017) demonstrated that the accuracy of teachers' assessments of reading comprehension varies based on numerous factors and increases as the students' year level increases (Paleczek et al., 2017; Schmidt & Schabmann, 2010). Furthermore, teachers also appear to be more accurate when assessing more complex reading skills, and certain student characteristics (e.g. disability status, behaviour) appear to influence the accuracy of teachers' assessments of students' reading comprehension (Südkamp et al., 2012; Urhahne & Wijnia, 2021). For example, Begeny et al. (2011) and Paleczek et al. (2017) found that teachers rated the reading achievement of low- and average-performing readers less accurately. However, there is evidence that in their evaluations, teachers not only consider relevant student characteristics but also characteristics that are irrelevant to what is being assessed (Praetorius et al., 2010; Südkamp et al., 2018). Behavioural problems (R. E. Bennett et al., 1993; Schmidt & Schabmann, 2009), low prior knowledge, having been identified as having special educational needs (Hurwitz et al., 2007; Wilbert et al., 2020), being a second language learner (L2) (Limbos & Geva, 2001; Paleczek et al., 2017) and the sex of the student (Klapp, 2015; Klapp & Cliffordson, 2009) have all been investigated as student characteristics that may influence diagnostic judgments.

Research Question and Hypothesis

Primary education is crucial for the development of reading comprehension. As teacher ratings of academic skills have been shown to vary widely, it is essential to examine the accuracy of teachers' assessments of reading comprehension in inclusive primary schools. This is particularly important in view of the shift towards more inclusive education and the resulting high level of academic heterogeneity in German primary schools. Given these previous findings, the present study investigates the extent to which teachers of inclusive classes are able to accurately assess reading comprehension at different levels of complexity when the student population is diverse in various ways including grade level, SEN and sex.

Based on the results mentioned above, we derived the following research questions:

1. How accurately are teachers of inclusive classrooms able to judge the reading comprehension of their students on word, sentence and text levels?

a. Teacher's judgment accuracy is higher at higher levels of complexity (i.e., text > sentence > word).
b. Judgment accuracy of word, sentence and text levels varies between different teachers.

2. In what way does teacher judgment accuracy of reading comprehension differ based on the student characteristics of grade level, SEN and sex?

a. Teachers judge the reading comprehension of students in higher grade levels more accurately than they judge the reading comprehension of students in lower grade levels.
b. Teachers judge the reading comprehension of students with SEN in learning less accurately than the reading comprehension of students without SEN in learning.
c. Teachers judge the reading comprehension of students with SEN in emotion and behaviour less accurately than the reading comprehension of students without SEN in emotion and behaviour.
d. Teachers' judgement accuracy of reading comprehension varies depending on the sex of the student. [exploratory]

Methods

Participants

This study is part of a 4-year research project in primary schools in one district area of North Rhine Westphalia (Hennemann et al., 2018; Urton et al., 2018). The initial sample was taken from grade levels two to four and included 1,108 students and 62 teachers in 2017, and 585 students and 35 teachers in 2018.

In order to ensure sample coherence, classes with less than 10 valid ratings were excluded. We considered ratings valid when a teacher has assessed the reading comprehension of a student and the corresponding ELFE 1-6 (Lenhard & Schneider, 2006) data were available. The final sample included 97 teachers (median age category: 41-50 years, 8.74% male, median time working as a teacher category: 17 years) and 1,693 students. The students, aged between 7 and 12 years, were enrolled in grade levels two to four. The sex distribution (female: Min = 48.25%; Max = 51.76%) and total number of students in 2017 (Min = 342; Max = 384) were roughly even across all grade levels whereas, in 2018, most students were in grade level two (see Table 1).

The percentage of students with each category of SEN (individual students may be represented in multiple categories) can be seen in Table 2. Emotion and behaviour (8.4%) and learning (8.7%) were the most

Table 1
Sample description by grade and year

Grade level	n	Mean age	SD age	% Female	% SEN	% Migration
			2017			
2	342	8.41	0.42	48.25	25.15	37.85
3	382	9.43	0.47	50.00	17.02	39.83
4	384	10.46	0.45	51.30	13.28	35.65
			2018			
2	236	8.08	0.53	49.58	24.15	38.58
3	170	9.17	0.57	51.76	25.29	30.33
4	179	10.02	0.45	49.16	16.20	29.94

common SEN, while 'other' SEN were less common (6.1%). This distribution aligns with the policy of federal states to focus on including students with learning and/ or emotional and behavioural difficulties in mainstream schools.

Materials and measures

Standardized Assessment of Students' Reading Comprehension

A standardized diagnostic reading comprehension assessment instrument was used to measure the students' reading comprehension (ELFE 1-6; Lenhard & Schneider, 2006). ELFE 1-6 is a timed test designed especially for elementary school students that measures reading comprehension at the word, sentence and text levels. At the word level (72 items), each item consists of a picture next to which there are four possible words to choose from (example item: Next to the picture of an owl (German Eule) there are the words "Keule, Eule, Ende, Erde"). At the sentence level (28 items), a sentence is presented that includes several options which the student must choose from to complete part of the sentence (example item: Tim got ice cream … the woman "Tim bekam das Eis durch/mit/auf/in/von der Frau."). At the text level (20 items), a text is offered with a corresponding question. There are different levels in that some of the information has to be found in isolation, some of it has to be read across sentences and some of it has to be inferred. The student must find the correct alternative from four possible answers (example item: The sun is shining all day today. Which sentence is true? The weather is nice today; The weather was nice yesterday; It will rain tomorrow; It is raining today. "Heute scheint den ganzen Tag die Sonne. Welcher Satz stimmt? Heute ist schönes Wetter; Gestern war schönes Wetter; Morgen wird es regnen; Heute regnet es.").

Teachers' Assessment of Students' Reading Comprehension

Teachers were asked to rate the reading comprehension of students using four items in total that matched the domains gathered in the standardized diagnostic reading comprehension tool (ELFE 1-6). Teachers conducted criteria-based assessments of each student's reading comprehension at word, sentence and text levels, based on both short and long texts, using an eleven-point Likert scale (0 = not proficient at all; 10 = very proficient).

Special Educational Needs (SEN)

To identify children with SEN, teachers were asked to indicate whether SEN had been diagnosed for each student in their class. They were also asked to indicate the area(s) in which the student had been diagnosed with SEN (multiple responses were possible). For SEN in learning the possible answers were emotion and behaviour, language, hearing and communication, intellectual development, vision development and physical development. Administrative formal diagnostic procedures are

Table 2
Percentage SEN (diagnosed and undiagnosed)

Variable	% (n)
No SEN	80.4 (1362)
Learning	8.7 (147)
Emotion and behaviour	8.4 (142)
Other	6.1 (103)

Note. Total n = 1693. Multiple choices are possible and values do not sum up to 100%.

currently either avoided or suspended in the North-Rhine-Westphalian primary education system in accordance with the inclusive approach. Teachers were, therefore, asked to indicate the areas in which each student needs increased support, regardless of whether they have a diagnosed SEN. The resulting responses regarding diagnosed and suspected SEN were each combined into one SEN category that included both diagnosed and suspected SEN in that category (see Table 2). Students with either diagnosed or additional SEN in learning were assigned to the category 'SEN in learning'. Students with either diagnosed or additional SEN in emotion and behaviour were assigned to the category 'SEN in emotion and behaviour'. Students with other SEN were categorized as 'other SEN'. The category 'other SEN' was created to ensure all students were included in the analyses while making sure students with 'other SEN' were not mistakenly included in the 'no SEN' reference group.

Procedure

The study received approval from the district education authority by meeting the following approval criteria: compliance with data protection regulations and educational relevance of the research. Moreover, all participating students obtained consent from their parents or legal guardians.

Data collection took place in the first half of the second school semester (February to April) in 2017 and 2018. Bachelor's and master's students collaborated in pairs to carry out the standardized reading comprehension test (ELFE 1-6). The project team provided a standardized data collection script and students received training in data collection procedures. To prevent children from copying from each other, classes were split in half for the data collection and seated at a distance from each other. The ELFE 1-6 was conducted in the classroom and took approximately 45 minutes. During this time, teachers compiled a list of personal information about the children including SEN and increased educational support requirements, sex, migration background and age.

For the teachers' assessment of their students' reading comprehension, teachers filled in a 3-minute questionnaire for each student in which they rated the students' reading comprehension.

Analyses

Data analyses were carried out using R Statistical Software (R Core Team, 2024) with packages nlme (Pinheiro, Bates, & R Core Team, 2023) and psych (Revelle, 2024).

To investigate how accurately the different teachers were able to assess reading comprehension (Hypothesis 1a), correlations between each teacher's ratings of students' reading proficiency at word, sentence and text levels and the corresponding scores (T-values) from the standardized reading comprehension measures were calculated. Furthermore, a multivariate regression model allowing for variances in the T-values at word, sentence and text levels, as well as different correlations of these levels within the double nested structure teacher/student/complexity level, was calculated according to Snijders and Bosker (2012). This model was used to check for inferential statistical differences between correlations at word, sentence and text levels.

To test for differences in individual rating accuracy between teachers (Hypothesis 1b), a random-intercept model with respective scores (T-values) as the criterion and teacher rating as a predictor, but not including individual variability in rating accuracy, was calculated. This was compared to a random-slope model factoring in individual variability in rating accuracy separately for the word, sentence and text levels. Chi-square tests were calculated to find out whether the inclusion of random slopes significantly increased the model fit (further detail is provided in Supplement C https://doi.org/10.17605/OSF.IO/MWPDN).

To investigate what influence the student's grade level, SEN status and sex might have on the accuracy of the teachers' assessment (Hypotheses 2a to 2d), we set up multilevel regression models (students nested within teachers) separately for word, sentence and text levels. In these models, the scores (T-values) for word, sentence and text level comprehension were predicted by teachers' ratings, grade level, student's sex and SEN (SEN in learning, SEN in emotion and behaviour, and other SEN). In order to test Hypotheses 2a to 2d, we examined the respective interaction between grade level (a), SEN learning (b), SEN emotion and behaviour (c), and sex (d) and the teacher rating (further details can be found in Supplement E https://doi.org/10.17605/OSF.IO/MWPDN).

Results

Research Question 1: How accurately are teachers of inclusive classrooms able to judge the reading comprehension of their students at word, sentence and text levels?

In order to provide a basis on which to evaluate teachers' judgements regarding reading comprehension, we first present the descriptive statistics and correlations between variables (see Table 3). Students' reading comprehension (T-values) is the average of their scores on the word, sentence and text levels (M_{word} = 51.27, *SD* = 10.01; $M_{sentence}$ = 50.29, *SD* = 9.89; M_{text} = 50.26, *SD* = 9.99).

Regarding the teachers' rating (11-point Likert scale), it is apparent that, overall, they rated the students' reading comprehension high, although the teachers' ratings decreased with increasing complexity (M_{word} = 8.87, *SD* = 1.75; $M_{sentence}$ = 8.29, *SD* = 2.08; M_{text} = 7.54, *SD* = 2.36).

Hypothesis 1a: Judgment Accuracy at Word, Sentence and Text Levels

Teachers' average judgment accuracy at word, sentence and text levels proved to be moderate (see Figure 1) and was higher at sentence and text levels than at word level ($r_{\text{word level}}$ = .44, *SD* = 0.25; $r_{sentence\ level}$ = .59, *SD* = 0.20; $r_{text\ level}$ = .58, *SD* = 0.20). On average, teachers rated the reading compe-

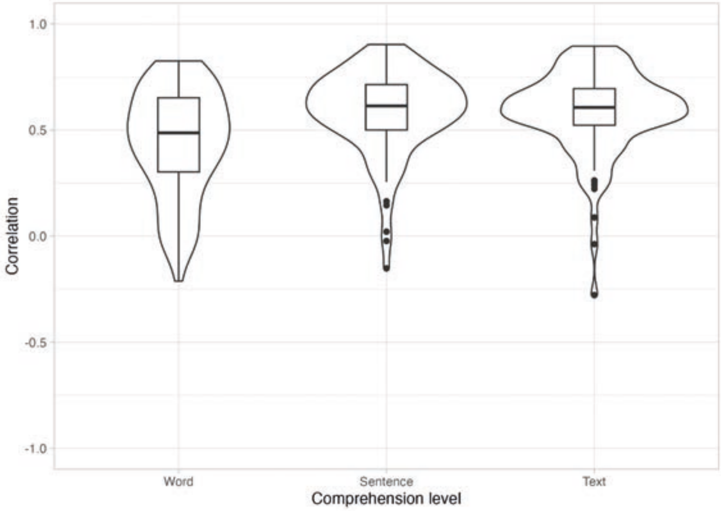

Figure 1

Correlation distributions per comprehension level

Table 3
Descriptives and correlations for all study variables

Variable	*M*	*SD*	1	2	3	4	5	6	7	8	9	10	11
1. ELFE word	51.27	10.01	–										
2. ELFE sentence	50.29	9.89	.76***	–									
3. ELFE text	50.26	9.99	.59***	.72***	–								
4. Rating word	8.87	1.75	.37***	.48***	.45***	–							
5. Rating sentence	8.29	2.08	.40***	.54***	.52***	.85***	–						
6. Rating text	7.54	2.36	.44***	.59***	.56***	.76***	.90***	–					
7. Sex	1.50	0.50	.00	-.04 †	-.04 †	-.04	-.05*	-.07**	–				
8. Age	9.31	0.98	-.11***	-.16***	-.10***	-.07**	.02	.01	.02	–			
9. Grade level	2.99	0.82	-.12***	-.12***	-.04	.01	.13***	.13***	-.02	.86***	–		
10. SEN in learning	0.09	0.28	-.21***	-.28***	-.24***	-.29***	-.33***	-.37***	.01	.06**	-.06*	–	
11. SEN in emotion and behaviour	0.08	0.28	-.05*	-.09***	-.09***	-.05*	-.06*	-.11***	.16***	-.02	-.06*	.10***	–
12. Other SEN	0.06	0.24	-.16***	-.18***	-.17***	-.22***	-.26***	-.25***	.05*	-.01	-.09***	.15***	.06*

Descriptives and correlations for all study variables

Note. *n* = 1693 for all variables.

Table 4
Multivariate regression model

	ELFE Word / Sentence / Text T-Value			
Predictors	*ß*	*se*	*t*	*p*
(Intercept)	0.02	0.05	0.48	.634
Sentence	0.00	0.03	-0.01	.989
Text	-0.01	0.04	-0.33	.741
Rating	0.32	0.02	15.37	<.001***
Rating * Sentence	0.15	0.02	7.91	<.001***
Rating * Text	0.15	0.02	6.18	<.001***
Random Effects				
Word Teacher	0.16			
Sentence Teacher	0.12			
Text Teacher	0.10			
Residual	0.71			
ICC	0.09			
N Teacher	97			
Observations	5079			
R² Marginal / Conditional	0.19 / 0.26			

tence of students as *M* = 17.45 (*SD* = 5.01). Table 4 shows the corresponding multivariate regression model. The predictor rating (*BetaRating* = 0.32, *pRating* < .001) represents the association with the ELFE scores for the reference category word level. As can be seen in the significant interactions between text and rating, and sentence and rating, ratings on both sentence and text levels show a stronger association with actual test scores compared to the word level (*Betarating*sentence* = 0.15, *prating*sentence* < .001; *Betarating**text = 0.15, *prating*text*, *p* < .001).

Hypothesis 1b: Variation in Judgment Accuracy at Word, Sentence and Text Levels

In order to investigate the extent to which the judgement accuracy of individual teachers differed, it is necessary to examine the variability of teachers' judgement accuracy. Our examination revealed a high degree of variance. While some teachers were able to assess their students' reading comprehension very accurately, there were negative correlations between the ELFE and teacher ratings made by other teachers (Min_{word} = -.21 to Max_{word} = .83; $Min_{sentence}$ = -.15 to $Max_{sentence}$ = .90; Min_{text} = -.28 to Max_{text} = .90). The results of inferential statistical analyses revealed a significantly better model-fit for the random-slope model on all levels (*Chiword* = 7.71, *pword* = .02; *Chisentence* = 6.45, *psentence* = .04; *Chitext* = 6.95, *ptext* = .03) suggesting substantial interindividual differences in judgment accuracy between teachers.

Research Question (2): How Do Student Characteristics Influence Judgment Accuracy?

Table 5 depicts three multilevel regression models in which ELFE word (Model 1), sentence (Model 2) and text (Model 3) level

Table 5
Regression model for each comprehension level

	Model 1 **ELFE Word T-Value**				**Model 2** **ELFE Sentence T-Value**				**Model 3** **ELFE Text T-Value**			
Predictors	*ß*	*se*	*t*	*p*	*ß*	*se*	*t*	*p*	*ß*	*se*	*t*	*p*
(Intercept)	0.07	0.05	1.47	0.143	0.06	0.04	1.47	0.143	0.01	0.04	0.15	0.883
SEN learning	-0.49	0.09	-5.43	**<0.001**	-0.56	0.09	-6.08	**<0.001**	-0.34	0.10	-3.31	**0.001**
SEN emotion and behaviour	-0.06	0.08	-0.80	0.426	-0.12	0.07	-1.67	0.096	-0.04	0.07	-0.48	0.633
Other SEN	-0.40	0.10	-3.78	**<0.001**	-0.25	0.10	-2.46	**0.014**	-0.34	0.10	-3.38	**0.001**
Sex	-0.04	0.02	-1.74	0.081	-0.01	0.02	-0.33	0.743	-0.01	0.02	-0.62	0.534
Grade level	-0.10	0.04	-2.71	**0.007**	-0.18	0.03	-5.48	**<0.001**	-0.12	0.03	-3.58	**<0.001**
Rating	0.44	0.03	15.74	**<0.001**	0.62	0.02	24.84	**<0.001**	0.64	0.02	26.01	**<0.001**
SEN learning*rating	-0.16	0.06	-2.61	**0.009**	-0.22	0.06	-3.57	**<0.001**	-0.24	0.07	-3.57	**<0.001**
SEN emotion and behaviour*rating	-0.21	0.08	-2.76	**0.006**	-0.03	0.07	-0.35	0.725	0.06	0.07	0.81	0.417
Other SEN*rating	-0.13	0.07	-1.99	**0.047**	-0.08	0.06	-1.37	0.172	-0.23	0.07	-3.41	**0.001**
Grade level*rating	-0.01	0.02	-0.53	0.594	0.05	0.02	2.31	**0.021**	0.09	0.02	4.49	**<0.001**
Sex*rating	-0.00	0.02	-0.12	0.904	0.02	0.02	1.06	0.289	0.03	0.02	1.38	0.166
Random Effects												
σ^2	0.67				0.55				0.55			
τ_{00}	0.16 $_{\text{id\_class\_teacher}}$				0.10 $_{\text{id\_class\_teacher}}$				0.10 $_{\text{id\_class\_teacher}}$			
ICC	0.19				0.15				0.16			
Marginal R^2 / Conditional R^2	0.208 / 0.359				0.382 / 0.477				0.369 / 0.467			

Note. n = 1693 for all variables, n = 97 teacher

T-values were used as criteria and student characteristics SEN (in learning, emotion and behaviour, and other), sex, grade level and teacher-rated competence for the corresponding complexity level were included as predictors. In addition, the interactions between student characteristics and teacher ratings were also included as predictors.

In general, the results indicate that students with SEN in learning (*Betaword* = -0.49, *pword* < .001; *Betasentence* = -0.56, *psentence* < .001; *Betatext* = -0.34, *ptext* = .001) and other SEN (*Betaword* = -0.40, *pword* < .001; *Betasentence* = -0.25, *psentence* = .014; *Betatext* = -0.34, *ptext* = .001) have lower ELFE scores compared to students with no SEN on all three complexity levels. Furthermore, students in higher grade levels had lower reading comprehension scores than students in lower grade levels (*Betaword* = -0.10, *pword* = .007; *Betasentence* = -0.18, *psentence* < .001; *Betatext* = -0.12, *ptext* < .001).

Hypothesis 2a: Differences in Judgment Accuracy Based on Grade Level

As can be seen in the interaction of Grade Level * Rating, the reading comprehension levels of children in higher grades were more accurately judged by teachers at the sentence (Model 2) and the text (Model 3) levels. The interaction effect in Model 1 is small and not significant suggesting that there are no grade-level differences in teachers' judgment accuracy of their students' word-level comprehension.

Hypothesis 2b: Differences in Judgment Accuracy Based on SEN in Learning

When considering SEN in Learning * Rating, students with SEN in learning were less accurately judged by teachers for word (Model 1), sentence (Model 2) and text (Model 3) levels than students without SEN.

Research Question 2c: Differences in Judgment Accuracy Based on SEN in Emotion and Behaviour

As reflected in the significant negative interaction of SEN in Emotion and Behaviour * Ratings, the reading comprehension of students with SEN in emotion and behaviour was, similarly, less accurately judged at the word (Model 1) level than children without SEN. However, in contrast to students with SEN in learning, no differences were detected in the teachers' judgment accuracy of students with SEN in emotion and behaviour and children without SEN at word (Model 2) and sentence (Model 3) levels.

Hypothesis 2d: Differences in Judgment Accuracy Based on Sex

Only non-significant interactions were detected for Sex * Rating at the word (Model 1), sentence (Model 2) and text (Model 3) levels suggesting that there is no difference in teachers' judgment accuracy of male and female students.

Discussion

The aim of this study is to examine teachers' ability to accurately assess their students' reading comprehension at word, sentence and text levels in inclusive classrooms. It also explores the extent to which teachers' judgment accuracy differs according to student characteristics.

According to *Hypothesis 1a*, teachers' assessment accuracy in inclusive classrooms varied when assessing their students' reading comprehension at word, sentence and text levels. There were low correlations at word level and moderate correlations at sentence and text levels. This corresponds to the results of Paleczek et al. (2017) and indicates that teachers have more issues assessing decoding ability than general reading comprehension. This also corroborates Schmidt and Schabmann (2009), who showed that teachers' diagnostic skills are inaccurate, particularly in the early stages of learning to read. Here, it is conceivable that teachers in higher grades focus more on the comprehension of whole sentences and texts and less on individual words and decoding (Clarke et al., 2014; Paleczek et al., 2015).

Hypothesis 1b, suggesting a high interindividual variability in teachers' assessment accuracy, was also confirmed. This is in line with previous findings (Gabriele et al., 2016; Wilbert et al., 2020).

Hypothesis 2a, which assumed that teachers assess their students' reading comprehension more accurately in higher grades than in lower grade levels, was confirmed for the sentence and text levels but not for word level. This is consistent with previous findings which show that teachers' assessments of reading comprehension improve with each additional year of primary school (Paleczek et al., 2017; Schmidt & Schabmann, 2010).

Hypothesis 2b was also confirmed in that teachers assessed the reading performance of students with SEN in learning less accurately than that of students without SEN. These findings correspond with those from previous studies (Begeny et al., 2011; Hurwitz et al., 2007; Paleczek et al., 2017). Although it can be assumed that both lower performance and the attribution of SEN influence assessment accuracy (Hurwitz et al., 2007; Wilbert et al., 2020), specific conclusions beyond this cannot be drawn from the available data.

Regarding *Hypothesis 2c*, the assumptions could only be confirmed to a limited extent, as SEN in emotion and behaviour only had an effect on the assessment of reading comprehension at word level. This corresponds to the results of Schmidt & Schabmann (2009).

Hypothesis 2d, which assumed differences in the rating accuracy of male and female students, could not be confirmed by the data. These results are in line with Urhahne and Wijnia (2021) whose review also showed that there is no, or at most a weak, correlation between student gender and teachers' assessment accuracy.

Limitations

The following limitations should be noted when interpreting the results. The participating teachers were not familiar with the standardized assessment of students' reading comprehension. This is relevant as previous research has shown that familiarity with the diagnostic instrument can increase the accuracy of the assessment (Begeny & Buchanan, 2010; Graney, 2008; Hurwitz et al., 2007). Additionally, the teachers' rating of each student's reading comprehension was based on a single item for the word and sentence levels, and two items for text-level comprehension. The use of a multi-item scale could provide a more reliable assessment (Südkamp et al., 2018).

Furthermore, SEN in learning and SEN in emotion and behaviour were not assessed by standardized instruments but based on unstandardized SEN assessment procedures, which is typical in Germany. If no official diagnosis was available, the suspected SEN was assessed by the teacher.

Outlook

According to PIRLS (McElvany et al., 2023) and PISA (OECD, 2023), a quarter of students in Germany are classified as low-achieving in reading and require tailored support. To combat this, teachers need to be able to identify students who are struggling and monitor their development (Paleczek et al., 2017). The results of our study suggest that, in particular, students with lower reading ability and students reading at lower levels of complexity are less accurately assessed, which may result in them not receiving the support they need. This may widen the gap in reading ability compared to their better reading classmates.

The Realistic Accuracy Model (Funder, 2012) suggests that a "good judge" is able to moderate for variables influencing their diagnostic accuracy. The present study suggests that there were some *good judges* who were able to accurately judge their students' reading comprehension. However, it was not possible to determine the criteria for a good judge of reading comprehension based on our data. Therefore, further research is needed to identify which characteristics teachers with higher diagnostic skills have and what interventions can be developed to increase teachers' diagnostic skills.

Therefore, further research is needed to identify which characteristics teachers with higher diagnostic skills have and what interventions can be developed to increase teachers' diagnostic skills.

References

Back, M. D., & Nestler, S. (2016). Accuracy of judging personality. In J. A. Hall, M. Schmid Mast, & T. V. West (Eds.), The social psychology of perceiving others accurately (pp. 98–124). Cambridge University Press. https://doi.org/10.1017/CBO9781316181959.005

Begeny, J. C., & Buchanan, H. (2010). Teachers' judgments of students' early literacy skills measured by the Early Literacy Skills Assessment: Comparisons of teachers with and without assessment administration experience. *Psychology in the Schools*, *47*(8), 859–868. https://doi.org/10.1002/pits.20509

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

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

Bennett, K. J., Brown, K. S., Boyle, M., Racine, Y., & Offord, D. (2003). Does low reading achievement at school entry cause conduct problems? *Social Science & Medicine*, *56*(12), 2443–2448. https://doi.org/10.1016/S0277-9536(02)00247-2

Bennett, R. E., Gottesman, R. L., Rock, D. A., & Cerullo, F. (1993). Influence of behavior perceptions and gender on teachers' judgments of students' academic skill. *Journal of Educational Psychology*, *85*(2), 347–356. https://doi.org/10.1037/0022-0663.85.2.347

Bos, W., Valtin, R., Hußmann, A., Wendt, H., & Goy, M. (2017). Wichtige Ergebnisse im Überblick. In W. Bos, R. Valtin, A. Hußmann, H. Wendt, & M. Goy (Eds.), Lesekompetenzen von Grundschulkindern in Deutschland im internationalen Vergleich (pp. 13–28). Waxmann.

Brunswik, E. (1956). *Perception and the representative design of psychological experiments* (2nd ed.). University of California Press.

Clarke, P. J., Truelove, E., Hulme, C., & Snowling, M. J. (2014). *Developing reading comprehension*. Wiley. https://doi.org/10.1002/9781118606711

Connor, C. M., Phillips, B. M., Kim, Y. G., Lonigan, C. J., Kaschak, M. P., Crowe, E., Dombek, J., & Al Otaiba, S. (2018). Examining the efficacy of targeted component interventions on language and literacy for third and fourth graders who are at risk of comprehension difficulties. *Scientific Studies of Reading*, *22*(6), 462–484. https://doi.org/10.1080/10888438.2018.1481409

Erickson, K. A., & Geist, L. A. (2016). The profiles of students with significant cognitive disabilities and complex communication needs. *Augmentative and Alternative Communication*, *32*(3), 187–197. https://doi.org/10.1080/07434618.2016.1213312

Förster, N., & Böhmer, I. (2017). Das Linsenmodell – Grundlagen und exemplarische Anwendungen in der pädagogisch-psychologischen Diagnostik. In A. Südkamp & A.-K. Praetorius (Eds.), *Diagnostische Kompetenz von Lehrkräften: Theoretische und methodische Weiterentwicklungen* (S. 46–50). Waxmann.

Funder, D. C. (2012). Accurate personality judgment. *Current Directions in Psychological Science*, *21*(3), 177-182.

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

Garwood, J. D., Brunsting, N. C., & Fox, L. C. (2014). Improving reading comprehension and fluency outcomes for adolescents with emotional-behavioral disorders: recent research synthesized. *Remedial and Special Education*, *35*(3), 181–194. https://doi.org/10.1177/0741932513514856

Glock, S., & Krolak-Schwerdt, S. (2014). Stereotype activation versus application: How teachers process and judge information about students from ethnic minorities and with low socioeconomic background. *Social Psychology of Education*, *17*(4), 589–607. https://doi.org/10.1007/s11218-014-9266-6

Golly Ledoux, V., Declercq, C., & Caillies, S. (2023). Psychological and nonpsychological inferences in reading comprehension in children: The role of initial level comprehension. *Canadian Journal of Experimental Psychology / Revue Canadienne de Psychologie Expérimentale*, *77*(1), 20–34. https://doi.org/10.1037/cep0000298

Graney, S. B. (2008). General education teacher judgments of their low-performing students' short-term reading progress. *Psychology in the Schools*, *45*(6), 537–549. https://doi.org/10.1002/pits.20322

Grigoryan, A. (2020). Major stages of reading skills development. *Armenian Journal of Special Education*, *4*(2), 42–51. https://doi.org/10.24234/se.2020.1.1.71

Hennemann, T., Hillenbrand, C., Fitting-Dahlmann, K., Wilbert, J., & Urton, K. (2018). Auf dem Weg zum inklusiven Schulsystem im Kreis Mettmann. Konzeption der wissenschaftlichen Begleitevaluation. *Zeitschrift für Heilpädagogik*, 4–16.

Hudson, A. K. (2022). Upper elementary teachers' knowledge of reading comprehension, classroom practice, and student's performance in reading comprehension. *Reading Research Quarterly*, rrq.491. https://doi.org/10.1002/rrq.491

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

International Literacy Association. (2019). *Children's rights to excellent literacy instruction [Position statement]*. https://www.literacyworldwide.org/docs/default-source/where-we-stand/ila-childrensrights-to-excellent-literacy-instruction.pdf

Kaiser, J., Retelsdorf, J., Südkamp, A., & Möller, J. (2013). Achievement and engagement: How student characteristics influence teacher judgments. *Learning and Instruction*, *28*, 73–84. https://doi.org/10.1016/j.learninstruc.2013.06.001

Klapp, A. (2015). Does grading affect educational attainment? A longitudinal study. *Assessment in Education: Principles, Policy & Practice*, *22*(3), 302–323. https://doi.org/10.1080/0969594X.2014.988121

Klapp, A., & Cliffordson, C. (2009). Effects of student characteristics on grades in compulsory school. *Educational Research and Evaluation*, *15*(1), 1–23. https://doi.org/10.1080/13803610802470425

Lane, K. L., Barton-Arwood, S. M., Nelson, J. R., & Wehby, J. (2008). Academic performance of students with emotional and behavioral disorders served in a self-contained setting. *Journal of Behavioral Education*, *17*(1), 43–62. https://doi.org/10.1007/s10864-007-9050-1

Lenhard, W. (2013). *Leseverständnis und Lesekompetenz: Grundlagen - Diagnostik - Förderung* (1. Auflage). Verlag W. Kohlhammer.

Lenhard, W., & Schneider, W. (2006). *ELFE 1-6. Ein Leseverständnistest für Erst- bis Sechstklässler*. Hogrefe.

Leuders, T., Loibl, K., & Dörfler, T. (2020). Diagnostische Urteile von Lehrkräften erklären – Ein Rahmenmodell für kognitive Modellierungen und deren experimentelle Prüfung. *Unterrichtswissenschaft*, *48*(4), 493–502. https://doi.org/10.1007/s42010-020-00085-5

Limbos, M. M., & Geva, E. (2001). Accuracy of teacher assessments of second-language students at risk for reading disability. *Journal of Learning Disabilities*, *34*(2), 136–151. https://doi.org/10.1177/002221940103400204

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

Macdonald, S. J., Deacon, L., & Merchant, J. (2016). "Too far gone": Dyslexia, homelessness, and pathways to drug use and dependency. *Insights into Learning Disabilities*, *13*(2), 117–134.

McElvany, N., Lorenz, R., Frey, A., Goldhammer, F., Schilcher, A., & Stubbe, T. C. (Eds.). (2023). IGLU 2021: Lesekompetenz von Grundschulkindern im internationalen Vergleich und im Trend über 20 Jahre. Waxmann Verlag GmbH. https://doi.org/10.31244/9783830997009

Mendoza-Pinargote, R. L., & Reyes-Meza, O. B. (2022). Language learning in the reading comprehension of elementary school students. *International Journal of Social Sciences*, *5*(2), 124–130. https://doi.org/10.21744/ijss.v5n2.1900

National Assessment Governing Board. (2007). *Reading framework for the 2007 National Assessment of Educational Progress*. U.S. Department of Education.

Nestler, S., & Back, M. D. (2013). Applications and extensions of the lens model to understand interpersonal judgments at zero acquaintance. *Current Directions in Psychological Science*, *22*(5), 374–379.

Noll, A., Roth, J., & Scholz, M. (2018). Fostering reading comprehension of learning tasks with pictorial symbols: A qualitative study of the subjective views and reading paths of children with and without special needs. *International Journal of Special Education*, *33*(3), 616–629.

Paleczek, L., Seifert, S., & Gasteiger-Klicpera, B. (2017). Influences on teachers' judgment accuracy of reading abilities on second and third grade students: A multilevel analysis. *Psychology in the Schools*, *54*(3), 228–245. https://doi.org/10.1002/pits.21993

Paleczek, L., Seifert, S., Schwab, S., & Gasteiger-Klicpera, B. (2015). Assessing reading and spelling abilities from three different angles – correlations between test scores, teachers' assessment and children's self-assessments in L1 and L2 children. *Procedia - Social and Behavioral Sciences*, *174*, 2200–2210. https://doi.org/10.1016/j.sbspro.2015.01.876

Paris, A. H., & Paris, S. G. (2003). Assessing narrative comprehension in young children. *Reading Research Quarterly*, *38*(1), 36–76. https://doi.org/10.1598/RRQ.38.1.3

Pinheiro, J., Bates, D., & R Core Team. (2023). nlme: linear and nonlinear mixed effects models (Version 3.1-164) [R package]. https://CRAN.R-project.org/package=nlme

Praetorius, A.-K., Berner, V.-D., Zeinz, H., Scheunpflug, A., & Dresel, M. (2013). Judgment confidence and judgment accuracy of teachers in judging self-concepts of students. *The Journal of Educational Research*, *106*(1), 64–76. https://doi.org/10.1080/00220671.2012.667010

Praetorius, A.-K., Greb, K., Lipowsky, F., & Gollwitzer, M. (2010). *Lehrkräfte als Diagnostiker. Welche Rolle spielt die Schülerleistung bei der Einschätzung von mathematischen Selbstkonzepten?* https://doi.org/10.25656/01:4570

Rambuyon, E. C., & Susada, B. L. (2022). Factors affecting reading comprehension in english of grade 4 pupils in owabangon elementary school. *International Journal Of Advance Research And Innovative Ideas In Education*, *8*(5), 1775–1786.

R Core Team. (2024). R: A language and environment for statistical computing (Version 2.4.6) [Software]. R Foundation for Statistical Computing. https://www.R-project.org/

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

Reid, R., Gonzalez, J. E., Nordness, P. D., Trout, A., & Epstein, M. H. (2004). A meta-analysis of the academic status of students with emotional/behavioral disturbance. *The Journal of Special Education*, *38*(3), 130–143. https://doi.org/10.1177/00224669040380030101

Revelle, W. (2024). psych: Procedures for psychological, psychometric, and personality research. Northwestern University, Evanston, Illinois. R package version 2.4.6, <https://CRAN.R-project.org/package=psych>.

Schmidt, B. M., & Schabmann, A. (2009). Sind Lehrer gute Lese-Rechtschreibdiagnostiker? Der Einfluss von problematischem Schülerverhalten auf die Einschätzungen der Lesekompetenz durch Lehrkräfte. *Heilpädagogische Forschung*, *35*(3), 133–145.

Schmidt, B. M., & Schabmann, A. (2010). „Es ist vorübergehend!" Lehrereinschätzungen über mögliche Lese- Rechtschreibprobleme. Eine klassifikatorische Analyse. *Heilpädagogische Forschung*, *3*, 106–115.

Schmitterer, A. M. A., & Brod, G. (2021). Which data do elementary school teachers use to determine reading difficulties in their students? *Journal of Learning Disabilities*, *54*(5), 349–364. https://doi.org/10.1177/0022219420981990

Schwab, S., Seifert, S., & Gasteiger-Klicpera, B. (2015). Leseunterricht in der Grundschule – Wer profitiert wirklich vom LARS- Leseförderprogramm? *Heilpädagogische Forschung*, *40*(4), 180–192.

Seifert, S., Schwab, S., & Gasteiger-Klicpera, B. (2015). Effects of a whole-class reading program designed for different reading levels and the learning needs of L1 and L2 children. *Reading & Writing Quarterly*, 1–28. https://doi.org/10.1080/10573569.2015.10291761

Snijders, T. A. B., & Bosker, R. J. (2012). *Multilevel analysis: An introduction to basic and advanced multilevel modeling* (2nd ed.). Sage.

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

Südkamp, A., Krawinkel, S., Lange, S., Wolf, S. M., & Tröster, H. (2018). Lehrkrafteinschätzungen sozialer Akzeptanz und sozialer Kompetenz: Akkuratheit und systematische Verzerrung in inklusiv geführten Schulklassen. *Zeitschrift für Pädagogische Psychologie*, *32*(1–2), 39–51. https://doi.org/10.1024/1010-0652/a000212

The United Nations. (2006). *United Nations Convention on the Rights of Persons with Disabilities*. https://www.un.org/disabilities/documents/convention/convention_accessible_pdf.pdf

Urhahne, D., & Wijnia, L. (2021). A review on the accuracy of teacher judgments. *Educational Research Review*, *32*, 1-. https://doi.org/10.1016/j.edurev.2020.100374

Urton, K., Börnert-Ringleb, M., Krull, J., Wilbert, J., & Hennemann, T. (2018). Inklusives Schulklima: Konzeptionelle Darstellung eines Rahmendmodells. *Zeitschrift für Heilpädagogik*, *69*, 40–52.

Wilbert, J., Urton, K., Krull, J., Kulawiak, P. R., Schwalbe, A., & Hennemann, T. (2020). Teachers' accuracy in estimating social inclusion of students with and without special educational needs. *Frontiers in Education*, *5*, 1–11. https://doi.org/10.3389/feduc.2020.598330

Author information

Sophia Hertel
https://orcid.org/0000-0003-3106-0981

Karolina Urton
https://orcid.org/0000-0002-5912-8143

Jürgen Wilbert
https://orcid.org/0000-0002-8392-2873

Johanna Krull
https://orcid.org/0000-0002-9067-6770

Jannis Bosch
https://orcid.org/0000-0002-1157-9914

Thomas Hennemann
https://orcid.org/0000-0003-4961-8680

Correspondence concerning this article should be addressed to
Sophia Hertel
Department of Special Education & Rehabilitation
Klosterstr. 79c, 50931 Cologne
sophia.hertel@uni-koeln.de

Erstmals eingereicht: 08.04.2024
Überarbeitung eingereicht: 02.09.2024
Angenommen: 12.10.2024

Offene Daten	Der anonymisierte Datensatz ist unter der folgender DOI Nummer zu finden: 10.5281/zenodo.15417832 Link: https://doi.org/10.5281/zenodo.15417832
Offener Code	Das für die Analyse verwendete R Script ist unter unter der folgenden DOI Nummer zu finden: 10.5281/zenodo.15417832 Link: https://doi.org/10.5281/zenodo.15417832
Offene Materialien	Materialien können auf Anfrage an johanna.krull@uni-koeln.de zur Verfügung gestellt werden.
Präregistrierung	Nein
Votum Ethikkommission	Es liegt kein Ethikvotum vor. Es fand eine Orientierung an der Deklaration von Helsinki statt. Die Schulbehörde, die Schulleitung, die Erziehungsberechtigten und die Schüler*innen wurden vorab über die Studie und den Umgang mit den erhobenen Daten informiert. Die Erziehungsberechtigten haben ihr Einverständnis zur Teilnahme ihrer Kinder an der Studie erteilt. Die Erhebung konnte jederzeit ohne Angabe von Gründen beendet und das Einverständnis zurückgenommen werden.
Finanzielle und weitere sachliche Unterstützung	Die Studie wurde vom Kreis Mettmann im Rahmen des Projektes ‚Mettmann 2.0 - Schulen auf dem Weg in die Inklusion' gefördert.
Autorenschaft	SH, KU, JW, JK conceived and designed the study; SH, KU, JK, JB and JW wrote the paper; JK and KU organized and supervised the data collection; TH, JK and KU administrated the project; TH raised the funding.