Research in Education and Rehabilitation 2026; 9(1): 179-199 **DOI**: **10.51558/2744-1555.2026.9.1.179**

OVERESTIMATING OR OVERLOOKING? TEACHERS' BIASED ASSESSMENT ON DIFFERENTIATED READING GROUPS

PRECJENJIVANJE ILI ZANEMARIVANJE? PRISTRASNE PROCJENE NASTAVNIKA PRI FORMIRANJU DIFERENCIRANIH GRUPA ZA ČITANJE

Judith Zellner [1*], Nikola Ebenbeck[1], Uta Hauck-Thum[1] and Markus Gebhardt[1]

[1]Department of Education and Psychology, University of Munich, Munich, Germany

Original Scientific Article

Received: 16/03/26
Revised: 24/04/26
Accepted: 30/05/26

ABSTRACT

In heterogeneous classrooms, teachers frequently make diagnostic decisions about how to differentiate instruction and allocate learning materials. However, research suggests that adaptive teaching decisions should be guided by data-based decision-making rather than social comparisons or personal traits. This study examines how N = 350 primary school students (grades 2 and 3) were assigned to differentiated reading materials – either intuitively by teachers or based on standardized reading assessments. Grouping patterns, decision factors, and biases were analyzed. Teachers matched the assessments in 52 % of cases, underestimated 34 %, and overestimated 14 %. They also tended to create comparable sized groups (24 % – 41 %), whereas standardized assessments identified 51 % of students as high-, 44 % moderate- and 5 % low-performing. Providing teachers with standardized results led to minor adjustments – 82 % of students remained in the intuitively assigned group. Multinomial logistic regressions indicated that intuitive group assignments and discrepancies between intuitive and standardized assignments were associated with selected student-related characteristics, including individual performance relative to the class mean, documented migration background, and documented behavioral or work-related issues. The findings highlight challenges in implementing data-based decision-making in everyday classroom practice and underline the importance of strengthening the systematic use of assessment data in teacher education.

Keywords: ability grouping; data-based decision-making; inclusive education

[*] **Correspondence author: Judith Zellner**, Special Education – Focus on Learning Disabilities and inclusive Education University of Munich ORCiD: 0009-0006-5113-9872
E-mail: Judith.zellner@edu.lmu.de

Research in Education and Rehabilitation 2026; 9(1): 179-199 **DOI**: **10.51558/2744-1555.2026.9.1.179**

SAŽETAK

U heterogenim razredima nastavnici često donose dijagnostičke odluke o tome kako diferencirati nastavu i rasporediti nastavne materijale. Međutim, istraživanja ukazuju na to da bi odluke o prilagođavanju nastave trebale biti zasnovane na podacima, a ne na socijalnim poređenjima ili ličnim karakteristikama učenika. U ovom istraživanju ispitano je kako je 350 učenika osnovne škole (drugog i trećeg razreda) raspoređeno prema diferenciranim materijalima za čitanje – bilo intuitivno, na osnovu procjena nastavnika, bilo na osnovu standardiziranih procjena čitanja. Analizirani su obrasci grupiranja, faktori koji utiču na donošenje odluka i pristrasnosti. Procjene nastavnika podudarale su se sa standardiziranim procjenama u 52 % slučajeva, dok su u 34 % slučajeva sposobnosti učenika bile potcijenjene, a u 14 % precijenjene. Nastavnici su također bili skloni formiranju grupa približno jednake veličine (24–41 %), dok su standardizirane procjene pokazale da 51 % učenika postiže visoke rezultate, 44 % umjerene, a 5 % niske rezultate. Upoznavanje nastavnika s rezultatima standardiziranih procjena dovelo je samo do manjih promjena – 82 % učenika ostalo je u grupi u koju su prvobitno intuitivno raspoređeni. Multinomijalne logističke regresijske analize pokazale su da su intuitivno raspoređivanje učenika u grupe i odstupanja između intuitivnog raspoređivanja i raspoređivanja zasnovanog na standardiziranim procjenama bili povezani s odabranim karakteristikama učenika, uključujući individualni uspjeh u odnosu na prosjek razreda, dokumentirano migracijsko porijeklo te dokumentirane probleme u ponašanju ili radu. Rezultati ukazuju na izazove u primjeni odlučivanja zasnovanog na podacima u svakodnevnoj nastavnoj praksi i naglašavaju važnost jačanja sistematske upotrebe podataka dobijenih procjenom u obrazovanju nastavnika.

Ključne riječi: grupiranje prema sposobnostima; odlučivanje zasnovano na podacima; inkluzivno obrazovanje

INTRODUCTION

Developing strong reading skills for all children is a central goal of primary education (KMK, 2024). In today's inclusive classrooms, teachers must accurately assess students' reading abilities to provide appropriate and effective support for diverse learners (Hußmann et al., 2017). Differentiated instruction is currently regarded as the state-of-the-art approach to addressing classroom heterogeneity and promoting effective teaching for diverse learners (Gheyssens et al., 2023). Nevertheless, it should not be seen as a universal or uncritically applicable model for inclusive education, as its effectiveness depends strongly on contextual and structural conditions (Letzel-Alt & Pozas, 2023; Lindner & Schwab, 2025).

At the core of differentiation lies the teacher's assessment of student performance. These assessments guide the formation of learning groups and the allocation of instructional materials but are often influenced by cognitive biases and reference group effects (Südkamp et al., 2012). Research shows that teachers may overestimate or underestimate students based on classroom context or student background characteristics rather than on actual performance data. Such biases threaten the accuracy of group assignments and, ultimately, the fairness of instructional differentiation.

Research in Education and Rehabilitation 2026; 9(1): 179-199 **DOI: 10.51558/2744-1555.2026.9.1.179**

Against this background, the present study examines how primary school teachers form differentiated reading groups, which indicators they rely on in this process, and how accurately these groupings reflect students' actual performance. We hypothesize that teachers' group assignments are systematically related to student performance indicators and contextual factors such as classroom reference effects. Furthermore, we expect that providing teachers with objective assessment data will lead to more consistent and data-based group assignments.

Assessment for differentiating students

Differentiation is an educational approach aimed at adapting instruction to the personal learning needs of students within a class. It is defined as the adjustment of curriculum, tasks, and teaching methods to meet individual learning requirements (Kyriakides & Creemers, 2008). While the principles of data-based and individualized decision-making for differentiation originate in special education research, they are increasingly relevant to all teachers in inclusive and primary education classrooms. Three main approaches to differentiation can be distinguished (Eikeland & Ohna, 2022). Differentiating teaching involves adapting instructional methods to address varying learning needs. Differentiating content entails providing diverse learning materials tailored to different performance levels. Differentiating students refers to assigning students to various groups based on prior material differentiation. Typically, content differentiation precedes the grouping of students, as they are assigned to groups only after materials have been tailored to different levels (Pozas et al., 2019). For this process to be effective, teachers need to accurately assess students' abilities in relation to the pre-differentiated materials to ensure each student receives fitting material and to ensure that the material is not fundamentally too easy or too difficult for the class. Since materials are often created in levels of differentiation, this process often results in the formation of rather homogeneous ability groups, where first all students in a class are grouped by their performance levels and then receive the same material within the assigned group. Differentiated instruction can be viewed as a classroom-based form of ability grouping, aiming to address diverse learning needs through flexible grouping and task adaptation. Although it has become a widely adopted pedagogical practice in response to classroom heterogeneity (Lindner, 2020; Gheyssens et al., 2023), empirical evidence on its effectiveness remains inconclusive. Meta-analytic research shows that within-class ability grouping yields, at best, small effects on student achievement and offers no consistent advantages for any performance group (Steenbergen-Hu et al., 2016). Consequently, while differentiated instruction represents the prevailing instructional approach, it should not be considered the ultimate pathway toward inclusive education (Letzel-Alt & Pozas, 2023).

The assessment of student performance relies on three reference norms (Jungjohann & Gebhardt, 2022). The individual reference norm evaluates a student's progress over time, using their own learning trajectory as the benchmark. The social reference norm compares students within the same class, identifying higher- and lower-performing groups (Becker & Rosen, 1992). The criterial reference norm measures performance against predefined learning objectives, e.g. anchored in the curriculum, which must be transparent and accessible to all students. Educational standards i.e. emphasize the social and critical reference norm, while in the context of inclusion, the individual reference norm is preferred (Holder & Kessels, 2018).

Research in Education and Rehabilitation 2026; 9(1): 179-199 **DOI**: **10.51558/2744-1555.2026.9.1.179**

Teachers favor the social reference norm when they assume stable personality traits in students or when instruction is minimally differentiated (Rheinberg, 2001), but it is also often applied unconsciously (Lintorf & Buch, 2020). In contrast, the individual reference norm is more commonly used when evaluating long-term development or when variables are perceived as complex constructs (Rheinberg, 2001). Also, special education teachers, who are expected to rely more heavily on the individual reference norm, often draw on social comparisons (Wilbert & Gerdes, 2009). Social reference norms are also assumed to be used to distinguish and assess performance levels both within-class and within-student (Ebenbeck et al., Review).

Differentiation decisions in the classroom are made by evaluating an individual student's performance in relation to both the other students in the class and the content being taught. Teacher judgments about student performance can be divided into three components (Schrader & Helmke, 1987):

1. The rank component measures the correlational agreement between teacher judgments and actual student performance. It describes how accurately teachers can rank students within the class. Studies show that the accuracy of these judgments varies significantly among teachers (Karing et al., 2011; Ebenbeck et al., Review).

2. The level component measures the absolute accuracy of teacher assessments. Teachers tend to systematically overestimate student performance, particularly for low-achieving students (Martin & Shapiro, 2011). Anders et al. (2010) found that teachers often inaccurately assess task difficulty, with a tendency to underestimate how challenging tasks are. The accuracy of these judgments varies widely among teachers, which may be attributed to individual judgment tendencies and educational content knowledge (Anders et al., 2010).

3. The differentiation component indicates how accurately teachers can assess the spread of student performance. This refers to their ability to correctly perceive the differences between high- and low-achieving students (Lorenz, 2011). While some studies have examined the accuracy of teacher judgments regarding this component, findings remain inconsistent (Hoffmann, 2017; Schrader, 2013; van Ophuysen, 2010). Similarly, although teacher judgments and expectations have been widely studied (e.g., Südkamp et al., 2012; Steenbergen-Hu et al., 2016), evidence on how these judgments inform teachers' group assignments for differentiated instruction is still limited and inconclusive, underscoring the need for further empirical investigation.

According to Feinberg and Shapiro (2009), teachers are accustomed to evaluating student performance in comparison to other students in the class. As a result, they find it easier to rank students than to assess the individual learning process of each child. Standardized assessments can therefore support teachers in inclusive classrooms by providing objective data to identify students at risk and to plan differentiated instruction accordingly. However, in Germany, teachers frequently rely on intuitive assessments to make educational decisions and assign students to differentiated learning groups, as inclusive structures are not yet widely implemented (Letzel-Alt & Pozas, 2023). Additionally, standardized tests are often perceived as time-consuming and requiring specific diagnostic competencies (Müller et al., 2017).

Research in Education and Rehabilitation 2026; 9(1): 179-199 **DOI: 10.51558/2744-1555.2026.9.1.179**

Weinert (1997) showed that teachers use a hypothetical "average student" as a benchmark in instructional planning, whose assumed performance level shapes teaching decisions. This practice resembles economic reference models, where median or mean values serve as decision-making benchmarks (Heller, 2009). This concept allows teachers to efficiently differentiate instruction and create differentiated instructional material but makes it difficult to individualize instruction at the student level (Bromme, 1992). The judgment of when a student is "too far" from the average depends heavily on the performance spread within the class (Labhart et al., 2018). In classes with high performance heterogeneity, teachers increasingly use within-class differentiation strategies, providing differentiated tasks for the entire class, while goal-differentiated instruction, where students are expected to achieve different learning outcomes, is less frequently deemed necessary. Instead, instruction again tends to be adapted to a hypothetical average student (Labhart et al., 2018; Pozas et al., 2020).

Bias in Assessment

Educational decisions in the classroom can be influenced by systematic biases, such as those stemming from the use of social reference norms and classroom composition effects, as well as student- and teacher-related characteristics that may introduce bias into assessments (e.g. Südkamp et al., 2012). This may especially be the case, when teachers rely on intuitive decisions, instead of standardized testing to gather information (Schellekens et al., 2021), as is often the case in the decision-making process in the classroom (Carter et al., 2020).

The social reference norm significantly impacts grading practices. For example, students with identical standardized test scores tend to receive better grades in low-achieving classes than in high-achieving classes (Schrader & Helmke, 2001). In classes with lower average performance, only the weakest students are identified as needing additional support, whereas in high-achieving classes, even average-performing students may be deemed to require intervention (Koßmann et al., 2024). Direct assessments unrelated to grading, however, appeared to be less influenced by the overall performance level of the class (Schrader & Helmke, 2001).

Closely related to this is the Big-Fish-Little-Pond Effect, which posits that students in high-achieving classes tended to develop a lower academic self-concept (Marsh, 1987). At the same time, teachers systematically rated students in high-achieving classes more harshly than equally capable students in lower-achieving classes (Meissel et al., 2017). Empirical path analyses confirmed this relationship (Marsh, 1987; Gerlach et al., 2007). Conversely, this likelihood decreases significantly if the at-risk student is in a class with peers who have special educational needs in the emotional and social domain (Koßmann et al., 2024). These mechanisms have long-term consequences for educational trajectories and transition recommendations (Rothenbusch, 2016).

Beyond social reference norms, Südkamp et al. (2012) provided evidence in their meta-analysis that factors at both the student and teacher level can distort or influence assessment decisions. A key factor is teacher experience. It is often assumed that greater teaching experience correlates with higher diagnostic accuracy, based on the idea that experienced teachers have more opportunities to develop diagnostic skills through teaching practice, collegial exchange, or professional development (McElvany et al., 2009). However, empirical

Research in Education and Rehabilitation 2026; 9(1): 179-199 **DOI: 10.51558/2744-1555.2026.9.1.179**

findings on this are mixed. While McElvany et al. (2009) found only a weak relationship between experience and judgment accuracy, other studies showed no clear effect (Praetorius et al., 2011). More recent evidence suggested that experience can contribute to higher accuracy and more elaborate diagnostic reasoning patterns, as experienced teachers tended to draw on a broader repertoire of observation and inference criteria when assessing student profiles (Kosel et al., 2024).

In addition to professional factors, there is also a link between intelligence and assessment literacy. According to Klauer et al. (2011), teachers' cognitive abilities were positively associated with the precision of their judgments. Demographic characteristics such as gender, however, had no measurable influence (Lorenz, 2011). There is consensus, however, that negative attitudes and biases toward certain student groups can distort assessments. Comparative research showed that teacher judgments were systematically influenced by students' social and migration backgrounds, often leading to over- or underestimations that affect long-term learning trajectories (Olczyk et al., 2022). Klug et al. (2016) further demonstrated that teachers' preconceptions can shape pedagogical decisions. Assumptions about students' backgrounds or learning behavior may subtly affect how teachers interpret performance data and form judgments (Terhart, 2022). At the same time, teachers with positive attitudes and higher self-efficacy tended to engage more systematically in diagnostic activities and to use assessment information more strategically for lesson planning (Ohle et al., 2015). Thus, variation in teacher beliefs represents one of several factors that can influence the quality of diagnostic decisions, without being the central focus of the present study.

Research question

To assign students to differentiated materials teachers need to make accurate educational decisions and assessments regarding student performance within their class. However, these decisions are often biased and influenced by reference group effects. We therefore examine how teachers group students for differentiated instruction in their classrooms, what factors they consider when making these decisions, and which elements contribute to the over- or underestimation of students within a class. We hypothesize that teachers' group assignment for differentiated instruction are systematically related to student performance indicators and contextual factors such as classroom reference effects. Furthermore, it is expected that providing teachers with objective assessment data will lead to more consistent and data-based group assignments.

METHODS

Sample

The sample included $N = 350$ students from second-grade ($N = 7$ classes) and third-grade ($N = 10$ classes), as well as their teachers ($N = 17$), across five primary schools in the German federal states of North Rhine-Westphalia and Bavaria. The participating teachers, all of whom were female, had an average age of $M = 39.35$ years ($SD = 10.3$, $Min = 27$, $Max = 60$). All teachers held degrees in primary education and had an average of $M = 11.18$ years of teaching

Research in Education and Rehabilitation 2026; 9(1): 179-199 **DOI: 10.51558/2744-1555.2026.9.1.179**

experience (*SD* = 10.35, *Min* = 1, *Max* = 33). The student sample had an average age of M = 9.15 years (*SD* = 0.76, *Min* = 8, *Max* = 11) and consisted of 128 females, 153 males, and 69 students for whom gender information was not available. Among the students, teachers reported for 15 special educational needs, including autism (*N* = 1), emotional and social development difficulties (*N* = 5), hearing impairment (*N* = 1), learning disabilities (*N* = 1), reading and spelling difficulties (*N* = 5), and speech and language impairments (*N* = 2).

Design and Instruments

The study design was based on triple-differentiated reading comprehension instruction materials for third-grade students (Diehl & Hanke, 2024). The materials included a text with the same content presented at three different levels of language complexity, along with accompanying comprehension questions.

The study followed a mixed-method design conducted across two cohorts. In the first cohort, teachers assigned students to groups based on their intuitive judgments and conducted a screening with their students independently of their initial group assignments. For the qualitative component, teachers were asked to briefly justify their decision for each student (free-text responses).

In the second step, students participated in a standardized reading screening (Jungjohann et al., 2023) administered through the digital testing platform Levumi.de. This assessment evaluated reading proficiency using two subtests: a Rapid Reading Test (RRT) and a Sentence Reading Test (SRT) (Jungjohann et al., 2023). In the second cohort, teachers (*N* = 8) were asked to reflect on their initial intuitive assessments in light of the raw test scores and, if necessary, adjust their assignments.

In the final step, we assigned all students criterion-based to one of the three differentiated groups based on the standardized test results. The criteria for group assignment were as follows: Students who scored 21 or more points on the SRT were assigned to *Group 3* (most challenging material). Students who scored 23 or fewer points on the RRT or 7 or fewer points on the SRT were assigned to *Group 1* (easiest material). All remaining students were assigned to *Group 2* (intermediate material) (Diehl & Hanke, 2024). This design enabled a comparison between intuitive and data-based decision-making.

Analysis

All statistical analyses were performed using RStudio (R core team, 2021) and the dplyr-package (Wickham et al., 2023). Given that the teachers' assessments are ordinally scaled, Pearson's Chi-squared test was used to determine whether the individual teacher has a significant impact on the intuitive teacher assessments and Spearman's Rho correlations were used to determine whether the intuitive teacher assessments correlated with the standardized test results. For teachers who revised their assessments based on raw test scores, this step was included in a further correlation analysis.

The differences in teachers' assessments were examined at both the student and class levels. At the student level, we analysed the extent to which differences between intuitive and standardized assignments could be attributed to individual student characteristics as predictors. We then calculated the mean of these differences for each teacher, with the average

Research in Education and Rehabilitation 2026; 9(1): 179-199 **DOI: 10.51558/2744-1555.2026.9.1.179**

intraindividual difference serving as an indicator of whether teachers tended to perceive uniform or differentiated reading profiles among their students at the class level.

Teacher justifications were analyzed using an inductive content analysis approach. Categories were developed directly from the material without predefined coding schemes. All justifications were coded dichotomously (1 = criterion mentioned, 0 = not mentioned). The final coding categories that emerged from the data were: (a) migration background, (b) special educational needs (SEN), (c) behavioral issues, (d) reading difficulties, (e) general positive performance characteristics, and (f) performance relative to the class average. To ensure coding reliability, two independent raters coded the data. Interrater agreement was calculated and reached a very high level (Cohen's $\kappa = .99$). Discrepancies were discussed and resolved through consensus. The coded qualitative data were then linked to the quantitative dataset to examine patterns in teachers' decision-making. To investigate factors associated with discrepancies between intuitive and standardised group assignments, we estimated two multinomial logistic regression models. Model 1 included the categories of assignment to *Group 1, 2* or *3*, with *Group 1* serving as the reference. Model 2 included the three nominal outcome categories: accurately estimated, underestimated, and overestimated. Accurate estimation was specified as the reference category. Both models included documented migration background and individual performance relative to the class mean. These predictors were selected because they showed sufficient variation and cell coverage for model estimation.

Several additional coded categories, including SEN, reading difficulties, general positive performance characteristics, and mathematics-related performance information, were rare and displayed empty cells across outcome categories. These variables were consequently reported descriptively but were not included in the regression model.

Given the presence of separation in some of the coded indicators, the multinomial model was estimated using mean bias-reduced adjusted scores with the brmultinom() function from the brglm2 package (Kosmidis, 2026; Kosmidis & Firth, 2011). This approach provides finite estimates under complete or quasi-complete separation and reduces small-sample estimation bias. We report regression coefficients (b), standard errors (SE), odds ratios (OR), 95% Wald confidence intervals, and two-sided Wald p values.

RESULTS

Comparison of intuitive and data-based group assignment

Based on the results of the standardized reading screening, 51 % of students demonstrated high reading performance and were assigned to *Group 3*. 44 % showed moderate reading ability (*Group 2*) and 5 % were assigned to *Group 1*.

Teachers intuitively formed comparably sized groups rather than strongly disproportionate ones. On average, 24 % of students were assigned to *Group 1* ($SD = 13.67$), 37 % to *Group 2* ($SD = 13.84$), and 41 % to *Group 3* ($SD = 15.48$). The correlation analysis showed a moderate correlation between the intuitive and the standardized group assignment with $r = .34$.

Most teachers (with the exception of one) assigned all three reading materials to their students. The proportion of students assigned to *Group 1* ranged from 5 % to 56 %, with an average deviation of 12.6 percentage points from the standardized placement ($SD = 10.25$, $Min = 0$, $Max = 31.58$). Assignments to *Group 2* ranged from 18 % to 67 %, deviating on average by 18 percentage points ($SD = 13$, $Min = 0$, $Max = 43$). For *Group 3*, placements ranged from 20 % to 68 %, with an average deviation of 24 percentage points from the

Research in Education and Rehabilitation 2026; 9(1): 179-199 **DOI: 10.51558/2744-1555.2026.9.1.179**

standardized grouping (*SD* = 15, *Min* = 0, *Max* = 47). In some cases, the group distribution followed a normal pattern (e.g., T6). As shown in Figure 1, four teachers (T10, T11, T12, T15) assigned students in a way that was closely aligned with the standardized grouping. These four teachers justified their intuitive decisions as follows: 28 % of students were placed based on their "migration background", 7 % due to "negative work behavior," 3 % due to "positive work behavior," and 1 % due to "reading difficulties" or "dyscalculia."

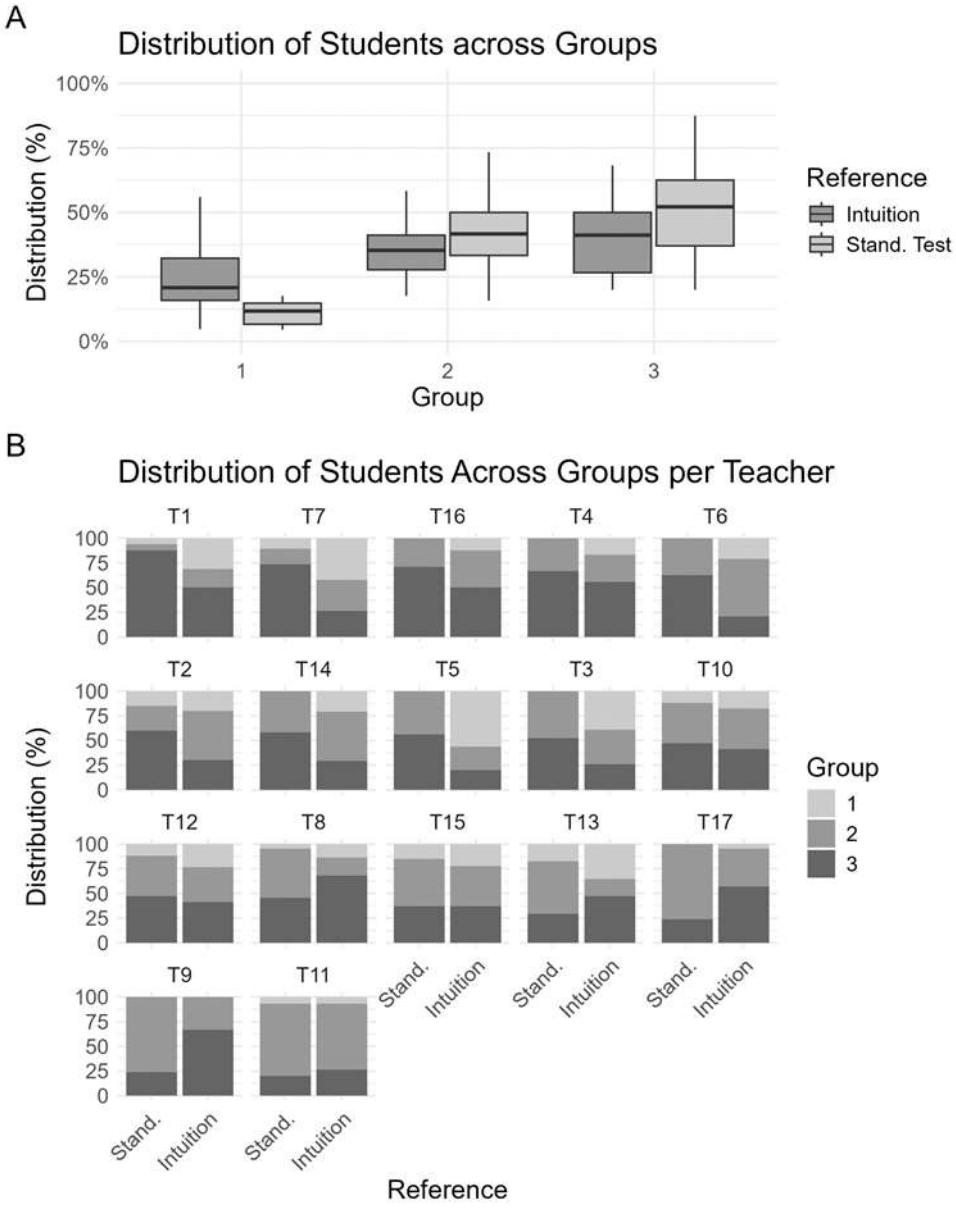

Figure 1. Intuitive and Standardized Group Assignment across all Students (A) and per Teacher (B)

The comparison between intuitive and standardized student grouping revealed clear discrepancies, as 52 % were placed accurately (*N* = 180), 34 % of students were underestimated (*N* = 119) and 14 % were overestimated (*N* = 50).

Students were therefore more than twice as likely to be underestimated as overestimated, suggesting a systematic tendency among teachers to assign students to lower rather than

Research in Education and Rehabilitation 2026; 9(1): 179-199 **DOI**: **10.51558/2744-1555.2026.9.1.179**

higher ability groups. A chi-square test confirmed the pattern of Figure 1 with $\chi^2(32) = 115.61$, $p < .001$ and $V = .41$, indicating that misjudgements were not randomly distributed, but instead showed a systematic relationship to individual teachers' decisions.

Toward data-based instructional decisions using test results

Furthermore, eight teachers ($N = 8$) had the opportunity to revise their initial intuitive group assignments of students ($N = 167$) and were explicitly instructed to base their revisions on the standardized test data. After reviewing the results, most teachers made only minor adjustments to their original decisions. Four teachers made no changes at all. The majority of students (83 %) remained in their original group. A total of 16 % were moved up or down by one level, and only 2 % ($N = 3$ students) experienced a shift of two group levels. The decision was therefore rarely revised, although the results of the standardized test suggested this for 48 % of the students.

It is noteworthy that two teachers reassigned a student from *Group 3* directly to *Group 1*, while only one teacher made a reverse change from *Group 1* to *Group 3*. Most of the adjustments were made by three teachers and involved 20 students who were moved one level downward. Only one teacher reassigned a single student one level upward. Adjustments with student results slightly increased the correlation between intuitive teacher and standardized assignment from $r = .34$ to .39. Overall, the review of raw standardized test results led to only limited changes in group assignments.

Prediction of bias in group assignment

Nine out of 17 teachers provided justifications for the group assignments of their $N = 172$ students. The distribution of assignment accuracy in this subsample differed slightly from that in the full sample with 74 students (43 %) were accurately estimated, 88 students (51 %) were underestimated, and 10 students (6 %) were overestimated.

The written justifications of those nine teachers contained several student-related criteria. The most frequently documented justification for the decision was a standardized large-scale test (IQB, 2024), mentioned for 132 students. Teachers' own observations also played an important role (83 mentions). In addition, teachers referred to migration background (61), observed reading difficulties (10), behavioral issues such as ADHD (6), , and generally positive performance characteristics (5). Interestingly, positive performance in mathematics as explanation for reading instruction decisions was documented for two students. Seven teachers based their entire class assignment on the large scale test, five of them uniformly for all students in the class without differentiation. Two teachers added their own observations in addition to the large-scale test. Generally positive performance traits were only noted for students placed in the highest group.

Several coded indicators were therefore too sparse for inclusion in the final regression model. Reading difficulties were documented for 10 students, of whom seven were accurately assigned according to the standardised grouping and three were underestimated. General positive performance characteristics were documented for five students; three were accurately assigned and two were overestimated. Mathematics-related performance information was

Research in Education and Rehabilitation 2026; 9(1): 179-199 **DOI: 10.51558/2744-1555.2026.9.1.179**

documented for two students, both of whom were underestimated. Behavioral or work-related issues were documented for six students, with one accurately assigned student and five underestimated students. Among the nine students with documented special educational needs, five were accurately assigned and four were underestimated; none was overestimated. These sparse distributions resulted in empty outcome cells and would have produced unstable estimates in a conventional multinomial regression. Accordingly, these variables are reported descriptively only and were not included in the final regression models.

To examine whether student- or class-related factors appear systematically in individual group assignments, a first bias-reduced multinomial logistic regression model was estimated with intuitive *Group 1* as the reference category. To investigate predictors of teachers' intuitive group assignments, a second bias-reduced multinomial logistic regression model analyzed predictors of over- or underestimation of student performance by teachers with accurate estimation as reference category. Both models included documented migration background and individual performance relative to the class mean. These variables were selected because they showed sufficient variation and cell coverage for model estimation.

The first bias-reduced multinomial logistic regression model examined predictors of intuitive group assignment, with *Group 1* as the reference category. Migration background was significantly associated with assignment to both *Group 2* and *Group 3* relative to *Group 1*. Students with a documented migration background had lower odds of being intuitively assigned to *Group 2* rather than *Group 1* than students without such documentation (*OR* = 0.33, 95% *CI* [0.13, 0.83], $p = .019$). The association was even stronger for assignment to *Group 3* rather than *Group 1* (*OR* = 0.02, 95% *CI* [0.004, 0.06], $p < .001$). Individual performance relative to the class mean was significantly associated with intuitive assignment to higher groups. For each one-unit increase in performance relative to the class mean, the odds of assignment to *Group 2* rather than *Group 1* increased by a factor of 5.68 (95% *CI* [2.37, 13.58], $p < .001$). Likewise, higher relative performance was associated with substantially higher odds of assignment to *Group 3* rather than *Group 1* (*OR* = 29.51, 95% *CI* [9.36, 92.76], $p < .001$).

The second bias-reduced multinomial logistic regression model examined predictors of discrepancies between intuitive and standardised group assignments, with accurate estimation as the reference category. Migration background was strongly associated with underestimation. Students with a documented migration background had substantially higher odds of being underestimated rather than accurately estimated than students without such documentation (*OR* = 25.82, 95% *CI* [8.99, 74.17], $p < .001$). Migration background was not significantly associated with overestimation relative to accurate estimation (*OR* = 2.68, 95% *CI* [0.36, 19.73], $p = .333$).

Performance relative to the class mean was not significantly associated with underestimation (*OR* = 1.46, 95% *CI* [0.71, 2.99], $p = .302$). However, higher relative performance was associated with lower odds of overestimation rather than accurate estimation (*OR* = 0.34, 95% *CI* [0.12, 0.94], $p = .038$). Thus, students performing relatively higher within their class were less likely to be assigned to a higher intuitive group than indicated by the standardised grouping.

Research in Education and Rehabilitation 2026; 9(1): 179-199 **DOI**: **10.51558/2744-1555.2026.9.1.179**

Table 1. Results of the Multinomial Logistic Regression on Over- and Underestimation

	Underestimation					Overestimation				
Predictors	b	SE	OR	95%-CI	p	b	SE	OR	95%-CI	p
Migration background	3.25	0.54	25.82	[8.99, 74.17]	< .001	0.99	1.02	2.68	[0.36, 19.73]	.333
Performance compared to class average	0.38	0.37	1.46	[0.71, 2.99]	.302	−1.08	0.52	0.34	[0.12, 0.94]	038

Note. b = regression coefficient, *SE* = standard error, *OR* = odds ratio, 95%-CI = 95% confidence interval, *p* = significance level. Significant results in bold.

DISCUSSION

Instructional decisions in the classroom are often based on the teacher's gut feeling instead of standardized data (Carter et al., 2022) and are therefore vulnerable to bias. Teachers' intuitive assignment to differentiated instructional groups diverged from grouping based on standardized test results. According to the screening data, half of all students were placed in the group with the hardest material (*Group 3*), slightly fewer in the intermediate group (*Group 2*), and only 5.43% in the l group with the easiest material (*Group 1*). This distribution aligned with our expectations based on the design of the instructional materials and the sample, which are second- and third-grade primary students, who, according to curriculum standards, are expected to read fluently (KMK, 2024). Nevertheless, it is precisely the accurate identification and targeted support of low-achieving students that is essential to ensure equitable learning opportunities in the classroom (Lindner & Schwab, 2020).

Teachers intuitively assigned significantly more students to the lowest-performing group than the standardized screening indicated. Their groupings were roughly evenly distributed across the three groups: 24 % of students were assigned to *Group 1*, 37 % to *Group 2*, and 41 % to *Group 3*. However, these proportions varied considerably across teachers, in line with prior research showing that the accuracy of differentiation judgments differs among teachers (Karing et al., 2011; Zellner & Gebhardt, 2025). While no specific instructions were given regarding the group size and how to allocate students, the near-equal distribution suggests that some teachers may have assumed all three groups needed to be filled, reflecting a norm-referenced rather than need-based approach. Such practices may inadvertently reinforce inequities by reproducing perceived "normal" distributions instead of responding to actual learner diversity. This pattern is consistent with the "rank component" described by Schrader and Helmke (1987), in which teachers rank students by performance and distribute them accordingly. Some teachers, (e.g., T15) appeared to have followed this trichotomous strategy. Others (e.g., T6) seemed to identify only the highest and lowest performers, placing all other students in the middle group. Another explanation might be that some teachers relied on a social norm based on a perceived average student, assigning most students to that group and only reallocating those who clearly deviated from it (e.g., T8, T11). Alternatively, some teachers may have applied a criterial norm, setting a performance standard for the class and differentiating only those who fell above or below it. Both approaches align with the

Research in Education and Rehabilitation 2026; 9(1): 179-199 **DOI: 10.51558/2744-1555.2026.9.1.179**

"differentiation component" (Schrader & Helmke, 1987), which reflects teachers' sensitivity to intra-class performance variation and their ability to identify both high and low achievers (Weinert, 1997; Labhardt et al., 2018; Lorenz, 2011).

Teachers intuitively placed about half of the students in the appropriate group, one third were underestimated, and one sixth were overestimated – meaning students were twice as likely to be underestimated as overestimated. However, prior research stated that teachers tend to overestimate student performance, particularly for low-achieving students (Martin & Shapiro, 2011). In addition, we also found in line with prior research that teacher judgments are systematically influenced by students' social and migration backgrounds, potentially leading to incorrect intuitive teacher judgements that can affect students' long-term learning trajectories and educational opportunities (Olczyk et al., 2022). Anders et al. (2010) found that teachers often misjudge task difficulty, with a tendency to underestimate how demanding learning materials are. This raises the possibility that inaccurate grouping reflects inaccurate perceptions of instructional material difficulty rather than inaccurate assessments of students. Alternatively, teachers may act out of a general pedagogical mindset aimed at avoiding negative learning experiences of students by assigning tasks that are more likely to be manageable. These interpretations remain speculative, as they are only based on quantitative data analysis and theoretical reasoning. To validate these assumptions, future research should include interviews or think-aloud studies to further understand teacher´s intuition.

Theoretical and empirical work emphasizes that standardized test data can support teachers in making informed, data-based instructional decisions (e.g., Fuchs & Fuchs, 2006). However, our findings showed that teachers made only minimal adjustments after receiving students' test scores. In 82 % of all cases, no changes were made to the initial group assignments. The remaining students were mostly adjusted by one group. This raises the question of why data are not used to inform decisions more consistently. One possible explanation is that teachers are not adequately trained to interpret standardized test results and may lack the data literacy required for data-based decisions. So, this points to a gap not in data availability but in data literacy and professional self-efficacy. Diverse learner profiles among preservice teachers require different instructional approaches (Zellner et al., 2024; Zellner & Gebhardt, 2025). Teacher education and professional development should therefore explicitly address barriers to evidence use, including uncertainty, mistrust, and habitual reliance on intuition. Teachers who lack confidence in using standardized test data for instructional decisions benefit most from a carefully scaffolded teacher education approach. This approach should begin with solid theoretical foundations, progress through guided case work, and gradually advance to complex application scenarios. Applied, classroom-oriented training formats like classroom-simulations (e. g. diagnostic conversations with an AI-generated teacher avatar; Zellner et al., 2025), foster data-based decision-making more systematically in teacher education programs, ideally allowing teachers not only to acquire knowledge but also to actively apply it in practical settings (Hauck-Thum, 2021).

Another explanation could be a lack of trust in standardized assessments. Teachers who have had consistently positive experiences with intuitive decision-making may not see the added value of relying on standardized test results. Similar hesitancy to evidence-based practices has been observed in other professional domains, such as medicine and psychology, where it is

Research in Education and Rehabilitation 2026; 9(1): 179-199 **DOI: 10.51558/2744-1555.2026.9.1.179**

discussed under terms like scepticism or even denialism (Prot & Anderson, 2020). The teachers in our study had an average of $M = 11.18$ years of teaching experience ($SD = 10.35$, $Min = 1$, $Max = 33$), indicating considerable heterogeneity in professional experience. This broad range may help explain the variation in teachers' trust in data-based assessment, as less experienced teachers might rely more on standardized information, whereas more experienced teachers may depend on established intuitive routines. Addressing such reservations requires professional learning opportunities that build diagnostic competence, foster critical reflection on bias, and strengthen self-efficacy in interpreting data. Promoting a culture of digitality – integrating both analogue and digital practices – can further support this transformation (Hauck-Thum, 2021), enabling teachers to view standardized assessments not as a threat but as a complementary tool for informed, equitable decision-making in inclusive classrooms.

The regressions analysis suggest that teachers' intuitive group assignments were associated with several student- and classroom-level characteristics. Therefore, developing prospective teachers' ability to critically reflect on their own expectations and biases, particularly in relation to perceived norm deviation, may represent a core element of diagnostic training in teacher education. Notably, stereotypical student characteristics such as migration background showed distinct patterns of association. Students with a documented migration background had lower odds of being intuitively assigned to *Group 2* and *Group 3* rather than *Group 1*. At the same time, documented migration background was associated with substantially higher odds of underestimation relative to accurate estimation. Migration background was not significantly associated with overestimation. However, these findings do not imply causal effects of the characteristics themselves. We do not assume that the characteristics themselves directly influenced grouping, but rather that teachers' perceptions of norm deviation shaped their decisions (Heller, 2009). The concept of normality may refer either to a teacher's personal idea of what is typical or to the class average as reference. This social comparison mechanism, often referred to as the big-fish-little-pond effect, was evident in our data: Performance relative to the class mean was strongly and positively associated with intuitive assignment to higher groups. This pattern indicates that teachers' intuitive assignments were related to students' relative standing within their classroom. However, the exact magnitude of these effects should be interpreted with caution. Previous research showed that teachers' expectations can unconsciously influence students' academic self-concept and self-efficacy, particularly among those receiving special education support (e.g., Klassen & Lynch, 2007). In this light, the observed pattern of associations in teachers' intuitive decisions warrants careful consideration, particularly because it coincided with a disproportionately high number of students being placed in the lowest group, even when their standardized scores did not fully align with these placements. These findings underscore the importance of making instructional decisions with sensitivity and awareness of their potential impact on student self-efficacy. In inclusive elementary classrooms, where student performance can vary widely, it is critical to raise teachers' awareness that norm deviation, such as atypical classroom behavior (Hoffmann, 2017), should not be uncritically interpreted as an indicator of low academic potential. A mindset shift that views standardized test results as a support rather than a threat to teachers professional judgment may help reduce potential biases and ensure that all students receive the instructional support they need.

Research in Education and Rehabilitation 2026; 9(1): 179-199 **DOI**: **10.51558/2744-1555.2026.9.1.179**

Although the present study provided insights into teachers' group assignments, it did not include further teacher-level variables such as attitudes toward students with migration backgrounds or prior diagnostic experience. This represents a limitation, as previous research has demonstrated that such factors can systematically influence teachers' assessment accuracy and decision-making processes (e.g., Kosel et al., 2024; Olczyk et al., 2022). Furthermore, the regression analysis does not include an evaluation based on each individual teacher, even though there were differences among them. In addition, the relatively small sample size limits the statistical power of the analyses and may have resulted in unstable parameter estimates. Future studies should therefore incorporate these variables by employing larger samples to provide a more comprehensive understanding of teacher-related sources of variance in diagnostic judgments.

Therefore, data-based decision-making in the classroom should be implemented collaboratively and be systematically embedded in teacher education. This is particularly crucial in open classroom models and in the context of individualized learning and a culture of digitality, where a transformation of pedagogical practice is both necessary and overdue. Yet in reality, many teachers continue to conceptualize their role primarily as classroom managers who allocate students into three equally sized groups – an approach shaped by systemic constraints and prevailing didactic traditions. This points to the pressing need for a shift in professional identity: away from static categorization toward a diagnostic, data-informed teaching culture that emphasizes reflective practice, individual learning trajectories, and instructional equity.

Ethics approvals

The Ethics Committee of the University of Regensburg (protocol code: 21-2592-101, date of approval: 10 December 2021) has positively approved the research methodology and data protection of the present study.

REFERENCES

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(3), 175-193. http://dx.doi.org/10.2378/peu2010.art13d

Becker, W. E. & Rosen, S. (1992). The learning effect of assessment and evaluation in high school. Economics of Education Review, 11(2), 107-118. https://doi.org/10.1016/0272-7757(92)90002-K

Bromme, R. (1992). Der Lehrer als Experte. Huber.

Carter, M., Webster, A., Stephenson, J., Waddy, N., Stevens, R., Clements, M., & Morris, T. (2022). Decision-making regarding adjustments for students with special educational needs in mainstream classrooms. Research Papers in Education, 37(5), 729-755.

Research in Education and Rehabilitation 2026; 9(1): 179-199 **DOI: 10.51558/2744-1555.2026.9.1.179**

Creemers, B. P. M., & Kyriakides, L. (2010). Using the Dynamic Model to develop an evidence-based and theory-driven approach to school improvement. Irish Educational Studies, 29(1), 5–23. https://doi.org/10.1080/03323310903522669

Diehl, K., & Hanke, J. (2024). LesePfad (Lesepakte mit formativem Assessment digital): LesePfad (Reading packages with formative Assessment, digital). QfI-Qualifizierung für Inklusion. Online-Zeitschrift zur Forschung über Aus-, Fort-und Weiterbildung pädagogischer Fachkräfte, 6(2). https://doi.org/10.25656/01:33041

Ebenbeck, Jungjohann, Zellner & Gebhardt. (Review). Which reference norms do teachers use to assess students' reading subskills in inclusive education?

Eikeland, I. & Ohna, S. E. (2022). Differentiation in education: a configurative review. Nordic Journal of Studies in Educational Policy, 8(3), 157-170. https://doi.org/10.1080/20020317.2022.2039351

Feinberg, A. B. & Shapiro, E. S. (2009). Teacher Accuracy: An Examination of Teacher-Based Judgments of Students' Reading With Differing Achievement Levels. The Journal of Educational Research, 102(6), 453–462. https://doi.org/10.3200/JOER.102.6.453-462

Fuchs, D., & Fuchs, L. S. (2006). Introduction to Response to Intervention: What, why, and how valid is it? Reading Research Quarterly, 41(1), 93–99. https://doi.org/10.1598/RRQ.41.1.4

Gerlach, E., Trautwein, U. & Lüdtke, O. (2007). Referenzgruppeneffekte im Sportunterricht. Zeitschrift für Sozialpsychologie, 38(2), 73-83. https://doi.org/10.1024/0044-3514.38.2.73

Gheyssens, E., Griful-Freixenet, J., & Struyven, K. (2023). Differentiated instruction as an approach to establish effective teaching in inclusive classrooms. In R. Maulana, M. Helms-Lorenz, R. M. Klassen (eds.). Effective teaching around the world: Theoretical, empirical, methodological and practical insights. 677-689. Cham: Springer International Publishing. https://doi.org/10.1007/978-3-031-31678-4_30

Hauck-Thum, U. (2021). Lehrer*innenbildung im Kontext kultureller Veränderungsprozesse. Ludwigsburger Beiträge zur Medienpädagogik, 21, 1-12. https://doi.org/10.21240/lbzm/21/23

Heller, U. (2009). Eine ökonomische Analyse von Referenzgruppeneffekten bei der Gestaltung von Managerbezügen. [Dissertation]. Universität Freiburg, Schweiz. https://folia.unifr.ch/documents/301494/preview/HellerU.pdf

Hoffmann, L. (2017). Lehrkräfte als Diagnostikerinnen und Diagnostiker. Untersuchung zu ausgewählten Tätigkeiten von Grundschullehrerinnen und Grundschullehrern im Bereich der pädagogisch-psychologischen Diagnostik. [Dissertation]. Universität Potsdam. https://doi.org/urn:nbn:de:kobv:517-opus4-404584

Holder, K., & Kessels, U. (2018). Between educational standards and inclusion: An experimental study on the effects of "standardization" and "individualization" on teachers' reference norm orientations. Unterrichtswissenschaft, 46, 87-104. https://doi.org/10.1007/s42010-018-0013-0

Hußmann, A., Wendt, H., Bos, W., Bremerich-Vos, A., Kasper, D., Lankes, E., McElvany, N., Stubbe, T. C. & Valtin, R. (2017). IGLU 2016. Lesekompetenzen von

Research in Education and Rehabilitation 2026; 9(1): 179-199 **DOI: 10.51558/2744-1555.2026.9.1.179**

Grundschulkindern in Deutschland im internationalen Vergleich. Waxmann. https://doi.org/10.25656/01:15476

Institut zur Qualitätsentwicklung im Bildungswesen e.V. an der Humboldt-Universität zu Berlin (IQB) (2024, September 9). VERA – Ein Überblick. Abgerufen am 26. April 2025, von https://www.iqb.hu-berlin.de/vera

Jungjohann, J., & Gebhardt, M. (2022). Reference norm orientation in the classroom. In M. Gebhardt, D. Scheer, & M. Schurig (eds.), Handbook of special educational diagnostics. Basics and concepts of status diagnostics, process diagnostics and support planning (S. 25–32). Regensburg: University Library. https://doi.org/10.5283/epub.53149

Jungjohann, J., Ebenbeck, N., Liebers, K., Diehl, K. & Gebhardt, M. (2023). Das Lesescreening LES-IN für inklusive Grundschulklassen. Entwicklung und psychometrische Prüfung einer Paper-Pencil-Version als Basis für computerbasiertes adaptives Testen (CAT). Empirische Sonderpädagogik 15(2), 141-156. https://doi.org/10.25656/01:2902

Karing, C., Matthäi, J. & Artelt, C. (2011). Genauigkeit von Lehrerurteilen über die Lesekompetenz ihrer Schülerinnen und Schüler in der Sekundarstufe I – Eine Frage der Spezifität? Zeitschrift für pädagogische Psychologie, 5(3), 159-172. https://doi.org/10.1024/1010-0652/a000041

Klassen, R. & Lynch, S. (2007). Self-Efficacy From the Perspective of Adolescents With LD and Their Specialist Teachers. Journal of learning disabilities, 40(6), 494-507. https://doi.org/10.1177/00222194070400060201.

Klauer, K. J. (2011). Lernverlaufsdiagnostik-Konzept, Schwierigkeiten und Möglichkeiten. Empirische Sonderpädagogik, 3(3), 207-224.

Klug, J., Gerich, M. & Schmitz, B. (2016). Can teachers' diagnostic competence be fostered through training and the use of a diary? Journal for educational research online, 8(3), 184-206. https://doi.org/10.25656/01:12825

Kosel, C., Bauer, J., & Seidel, T. (2024). Where experience makes a difference: Teachers' judgment accuracy and diagnostic reasoning regarding student learning characteristics. Frontiers in Psychology, 15. https://doi.org/10.3389/fpsyg.2024.1278472

Kosmidis, I. (2026). brglm2: Bias reduction in generalized linear models (R package version 1.1.0).

Kosmidis, I., & Firth, D. (2011). Multinomial logit bias reduction via the Poisson log-linear model. Biometrika, 98(3), 755–759. https://doi.org/10.1093/biomet/asr026

Koßmann, R., Kölm, J., & Gresch, C. (2024). Der „Fingerzeig "der Lehrkraft: Zusammenhänge zwischen Leistungs-und Verhaltensmerkmalen von Schulklassen und dem Vorliegen eines sonderpädagogischen Unterstützungsbedarfs im Lernen. Zeitschrift für Erziehungswissenschaft, 27(5), 1281-1303. https://doi.org/10.1007/s11618-024-01236-5

Kultusminister Konferenz (2024). Vereinbarung zur Arbeit an der Grundschule. Sekretariat der Kultusministerkonferenz.

Research in Education and Rehabilitation 2026; 9(1): 179-199 **DOI: 10.51558/2744-1555.2026.9.1.179**

https://www.kmk.org/fileadmin/veroeffentlichungen_beschluesse/2024/2024_03_15-Vereinbarung-Grundschule.pdf

Kutnick, P., Sebba, J., Blatchford, P., Galton, M., Thorp, J., MacIntyre, H. & Berdondini, L. (2005). The effects of pupil grouping: Literature review. DfES Publications, 688. https://dera.ioe.ac.uk/id/eprint/18143/1/RR688.pdf

Labhart, D., Pool Maag, S. & Moser Opitz, E. (2018). Differenzieren im selektiven Schulsystem. Der Widerspruch zwischen den gesellschaftlichen Funktionen der Schule und der Forderung nach individueller Förderung. Sonderpädagogische Förderung heute, 63(1), 71-87.

Letzel-Alt, V., & Pozas, M. (2023). Differentiated Instruction around the World: Exploring Differentiated Instructional Practice in General School Education. PEDOCS: Educational Research On-Line, 2023, 28077.

Lindner, K. T., & Schwab, S. (2020). Differentiation and individualisation in inclusive education: a systematic review and narrative synthesis. International journal of inclusive education, 1-21. https://doi.org/10.1080/13603116.2020.1813450

Lintorf, K., & Buch, S. R. (2020). Stable preference or flexibly oriented to the diagnostic goal - the choice of reference norms by prospective teachers. Journal of Educational Psychology. https://doi.org/10.1024/1010-0652/a000271

Lorenz, C. (2011). Diagnostische Kompetenz von Grundschullehrkräften: Strukturelle Aspekte und Bedingungen. University of Bamberg Press. https://doi.org/10.20378/irb-3956

Marks, R. (2025). Ability-grouping in Primary Schools: Case Studies and Critical Debate. Routledge. https://doi.org/10.4324/9781041053996

Marsh, H. W. (1987). The big-fish-little-pond effect on academic self-concept. Journal of Educational Psychology, 79(3), 280–295. https://doi.org/10.1037/0022-0663.79.3.280

Martin, S. D. & Shapiro, E. S. (2011). Examining the Accuracy of Teachers' Judgments of DIBELS Performance. Psychology in the Schools, 48(4), 343-356. https://doi.org/10.1002/pits.20558

McElvany, N., Schroeder, S., Hachfeld, A., Baumert, J., Richter, T., Schnotz, W., Horz, H., & Ullrich, M. (2009). Diagnostische Fähigkeiten von Lehrkräften. Zeitschrift für Pädagogische Psychologie, 23(34), 223–235. https://doi.org/10.1024/1010-0652.23.34.223.

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

Müller, X., Venetz, M. & Keiser, C. (2017). Nutzen von individuellen Förderplänen: Theoretischer Fachdiskurs und Wahrnehmung von Fachpersonen in der Schule. Vierteljahresschrift für Heilpädagogik und ihre Nachbargebiete, 86, 116-126.

Ohle, A., McElvany, N., Horz, H., & Ullrich, M. (2015). Text-picture integration—Teachers' attitudes, motivation and self-related cognitions in diagnostics. Journal for Educational Research Online, 7(2), 11–33. https://doi.org/10.25656/01:11488

Research in Education and Rehabilitation 2026; 9(1): 179-199 **DOI: 10.51558/2744-1555.2026.9.1.179**

Olczyk, M., Kwon, S. J., Lorenz, G., Perinetti Casoni, V., Schneider, T., Volodina, A., Waldfogel, J., & Washbrook, E. (2022). Teacher judgements, student social background, and student progress in primary school: A cross-country perspective. Zeitschrift für Erziehungswissenschaft, 26(2), 443-468. https://doi.org/10.1007/s11618-022-01119-7

Pozas, M., Letzel, V., & Schneider, C. (2020). Teachers and differentiated instruction: Exploring differentiation practices to address student diversity. Journal of Research in Special Educational Needs, 20(3), 217–230. https://doi.org/10.1111/1471-3802.12481

Praetorius, A.-K., Karst, K., Dickhäuser, O., & Lipowsky, F. (2011). Wie gut schätzen Lehrer die Fähigkeitsselbstkonzepte ihrer Schüler ein? Zur diagnostischen Kompetenz von Lehrkräften. Psychologie in Erziehung und Unterricht, 58(2), 81–91. https://doi.org/10.2378/peu2010.art30d

Prot, S. & Anderson, C. A. (2020). Science denial: Psychological processes underlying denial of science-based medical practices. In A. Lavorgna & A. Di Ronco (Hrsg.), Medical Misinformation and Social Harm in Non-science-based Health Practices (S. 24–37). Routledge.

R Core Team (2021). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. https://www.R-project.org/.

Rheinberg, F. (2001). Bezugsnormen und schulische Leistungsbeurteilung. In F. E. Weinert (eds.), Leistungsmessungen in Schulen (S. 59-71). Beltz Verlag.

Rothenbusch, S., Zettler, I., Voss, T., Lösch, T., & Trautwein, U. (2016). Exploring reference group effects on teachers' nominations of gifted students. Journal of Educational Psychology, 108(6), 883–897. https://doi.org/10.1037/edu0000085

Schellekens, L. H., Bok, H. G., De Jong, L. H., Van der Schaaf, M. F., Kremer, W. D., & Van der Vleuten, C. P. (2021). A scoping review on the notions of Assessment as Learning (AaL), Assessment for Learning (AfL), and Assessment of Learning (AoL). Studies in Educational Evaluation, 71. https://doi.org/10.1016/j.stueduc.2021.101094

Schrader, F. & Helmke, A. (2001). Alltägliche Leistungsbeurteilung durch Lehrer. In Weinert, F. E. (eds.), Leistungsmessungen in Schulen (S. 45-58). Beltz Verlag.

Steenbergen-Hu, S., Makel, M. C., & Olszewski-Kubilius, P. (2016). What one hundred years of research says about the effects of ability grouping and acceleration on K–12 students' academic achievement: Findings of two second-order meta-analyses. Review of Educational Research, 86(4), 849–899. https://doi.org/10.3102/0034654316675417

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.

Terhart, E. (2022, 9. Juni). Die Rolle von Lehrkräften bei der (Re-)Produktion von Bildungsgerechtigkeit. Bpb: Bundeszentrale für politische Bildung. https://www.bpb.de/themen/bildung/dossier-bildung/341050/die-rolle-von-lehrkraeften-bei-der-re-produktion-von-bildungsungerechtigkeit/

van Ophuysen, S. (2010). Professionelle pädagogisch-diagnostische Kompetenz. Eine theoretische und empirische Annäherung. In N. Berkemeyer, W. Bos, H. G.

Research in Education and Rehabilitation 2026; 9(1): 179-199 **DOI: 10.51558/2744-1555.2026.9.1.179**

Holtappels, N. McElvany & R. Schulz-Zander (eds.), Jahrbuch der Schulentwicklung (S. 203–234). Juventa.

Weinert, F. E. (1997). Notwendige Methodenvielfalt. In M. Meyer, U. Rampillon & G. Otto (Hrsg.), Lernmethoden – Lehrmethoden – Wege zur Selbstständigkeit (S. 50–52). Friedrich Jahresheft XV. Friedrich Verlag.

Wickham, H., François, R., Henry, L., Müller, K. & Vaughan, D. (2023). dplyr: A Grammar of Data Manipulation. R package version 1.1.4, https://github.com/tidyverse/dplyr, https://dplyr.tidyverse.org.

Wilbert, J., & Gerdes, H. (2009). The choice of reference norms for in the assessment of academic performance by prospective teachers specializing in learning disabilities. Curative Education Research, 35(3), 105-115.

Zellner, J., & Gebhardt, M. (2025). Scenario-Based Case Game to Promote Diagnostic Decision-Making in Inclusive Teacher Education. Empirische Sonderpädagogik, (2), 124-141. https://doi.org/10.2440/003-0044

Zellner, J., Ebenbeck, N., & Gebhardt, M. (2024). Entwicklung digitaler Simulationsspiele mit integrierten Entscheidungsbäumen zur Förderung der diagnostischen Entscheidungskompetenzen in der sonderpädagogischen Lehrkräfteausbildung. QfI-Qualifizierung für Inklusion, 6(2). https://doi.org/10.21248/qfi.162

Zellner, J., Koch, J., Fink, M., Ebenbeck, N., & Gebhardt, M. (2025). Fostering Diagnostic Decision-Making in Special Education Students through Simulation-Based Games and AI Pedagogical Agents: A Case-Based Learning Comparison in Higher Education. International Journal of Special Education, 40(2), 92-103. https://doi.org/10.52291/ijse.2025.40.24