



Article

Designing a Digital Flash Reading Test for Data-Based Decisions in Inclusive Classrooms: Duration and Word Length as Difficulty-Generating-Item Characteristics

Judith Zellner ^{*}, Nikola Ebenbeck  and Markus Gebhardt 

Special Educational Needs in Learning, Ludwig-Maximilians-University of Munich, Leopoldstraße 13, 80802 Munich, Germany

^{*} Correspondence: judith.zellner@edu.lmu.de

Abstract: Standardized assessment tools are essential for informed, data-driven decision-making. Reading speedily is a crucial early skill that all students should have the opportunity to develop in inclusive classrooms. To facilitate classroom-based reading diagnostics in this area of reading, we developed a flash reading test that reliably measures the performance of students with and without learning disabilities and intellectual disabilities. This test can be administered in the classroom and completed independently by students, taking only a few minutes, without requiring them to read aloud. The test is designed to provide an accurate assessment of the speed of lexical recall for all students. To evaluate the difficulty-generating-item characteristics of the new instrument, 400 primary and special school students participated in the test. The results indicate that students with low abilities and disabilities are particularly differentiated by the combination of a short display duration and short words. We provide information for test developers interested in designing similar assessments and teachers who can use this instrument to make informed decisions in the classroom.

Keywords: reading diagnostics; flash reading test; assessment; reading speed; inclusive teaching



Academic Editor: Garry Hornby

Received: 8 August 2024

Revised: 18 December 2024

Accepted: 20 December 2024

Published: 24 December 2024

Citation: Zellner, J., Ebenbeck, N., & Gebhardt, M. (2025). Designing a Digital Flash Reading Test for Data-Based Decisions in Inclusive Classrooms: Duration and Word Length as Difficulty-Generating-Item Characteristics. *Education Sciences*, 15(1), 5. <https://doi.org/10.3390/educsci15010005>

Copyright: © 2024 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

1. Introduction

An inclusive school recognizes the diverse performance levels of its students, necessitating differentiated and individualized teaching approaches. In reading, students are at various levels and require fitting instruction and support to achieve personal success (Al Otaiba et al., 2023). Screening instruments provide essential information for making informed support decisions (Grubb & Young, 2024). Most reading fluency assessments require one-on-one testing with a teacher (Morrison & Wilcox, 2020), which can be time-consuming. To address this, we developed a flash reading test that can be administered on a tablet and conducted within the classroom. This approach facilitates efficient diagnostics, providing teachers with data-driven insights into their students' speed of lexical recall across different word lengths.

1.1. Learning to Read in Heterogeneous Student Groups

Students enter school with varying levels of prior knowledge. While some can already read at the beginning of their schooling, others are just starting to learn the alphabet. This means that teachers are working with a heterogeneous group from the start. Especially concerning is that students with reading difficulties often go unidentified until the third

or fourth grade, which results in these different reading levels persisting throughout their school careers (PIRLS; Mullis & Martin, 2021). Particularly, students with reading difficulties need special support to catch up on learning gaps, avoid developing further difficulties in other subjects, and ultimately achieve good reading proficiency. This proficiency is foundational for their life outside school (Rosebrock & Nix, 2017; Galuschka & Schulte-Körne, 2015) and their future careers (Bennett et al., 2003). If they are not taught adequately at an early stage, unsuccessful experiences further risk their learning success. This preventive approach applies to all students with reading difficulties, including those with learning disabilities and intellectual disabilities.

Mastering basic reading skills is crucial for developing advanced reading abilities, as several authors show. According to Perfetti (2007), basic reading skills at the word and sentence levels free up cognitive capacity for constructing meaning at the text level. Lenhard (2019) states that phonological awareness, the lexical recall of words, and the formation and understanding of local coherence between sentences are fundamental to this process. Confidence and speed in the lexical retrieval of words from the mental lexicon form the basis for further reading processes, enabling secure reading comprehension at the word level. Mendoza-Pinargote and Reyes-Meza (2022) explained that reading comprehension develops first at the word level, then at the sentence level, and, finally, at the text level. If reading comprehension is not solid at the word level, there is a significantly increased risk of encountering difficulties at higher levels, both in acquiring written language and in other school subjects.

Reading comprehension is enabled or limited by the scope of one's vocabulary and the certainty and speed with which words from that vocabulary are recalled (Ennemoser et al., 2012). An insufficiently secure vocabulary hampers both reading acquisition and reading comprehension (Röthlisberger et al., 2021). For example, a study on third graders shows a moderate correlation between receptive vocabulary size and reading performance (Berendes et al., 2010). The importance of speed is highlighted in the dual-route model (Coltheart et al., 2001), where word reading occurs either directly via whole-word retrieval from the vocabulary or indirectly via recoding (combining phonological units into words). The difference between these two methods lies in their speed and efficiency. This effect is illustrated by Silverman et al. (2013), who report that fourth graders with good decoding skills but low reading fluency perform worse in reading comprehension than students with high reading fluency. High speed via the direct route is considered a prerequisite for adequate reading fluency in all grades, as it relieves working memory, thereby freeing up more capacity for reading comprehension (Perfetti, 2007).

1.2. Reading Abilities of Students with Disabilities

Individualized support and adaptive teaching in reading are needed for all students, but especially for students with disabilities (Schwab & Gasteiger-Klicpera, 2014), as their reading performances are usually lower than the reading performances of students without disabilities. Students with learning disabilities often show reading performances of one standard deviation below the average student (Gebhardt et al., 2015), but generally learn to read. However, they also often show weaker reading abilities than other students over the course of their schooling. On the other hand, about 30% of students with intellectual disabilities do not learn to read at all (Ratz & Lenhard, 2013). Among students with intellectual disabilities who do learn to read, there is great heterogeneity in reading abilities: 6.8% read at a logographic level, 31.9% at an alphabetic level, and 32% at an orthographic level. Di Blasi et al. (2019) compared the reading fluency, reading accuracy of words and pseudo-words, and reading and comprehension of texts of pupils with mild intellectual disability and students with borderline intelligence, which are comparable to students with

learning disabilities. Both groups of students showed weaker reading performance across many grades. Especially with reading fluency, students with intellectual disabilities showed more problems than students with borderline intelligence. For successful prevention, particular attention must be paid to those students who face considerable challenges in developing their reading skills (Di Blasi et al., 2019) and therefore require intensive reading support at school (Allor et al., 2010) to secure this basic skill.

1.3. Using Standardized Tests for Making Informed Support Decisions

Data-based Decision Making (DBDM) is a systemic approach that involves collecting, analyzing, and utilizing data to inform educational practices (Keuning et al., 2017; Schildkamp, 2019). The goal is to use economical, reliable, and valid diagnostics to create an optimal match between lessons and students' needs, with diagnostics and support directly linked through DBDM (Gebhardt et al., 2021; Blumenthal, 2017). In the context of inclusive education, DBDM helps teachers tailor their instructions and support decisions to meet the diverse needs of all students (Keuning et al., 2019). The aim is that eventually, the label of a disability will no longer be necessary for receiving appropriate support; instead, support will be selected solely based on a student's individual development and learning progress.

Data used for DBDM can vary but often include standardized test results. Formative procedures and ongoing support are considered more effective than extensive status tests (Voß, 2017). DBDM is particularly important for struggling students, like students with disabilities, because teachers' personal assessments are often influenced by students' behavior and social interaction (Schabmann & Schmidt, 2009), and these assessments are rarely differentiated, competency-oriented, comparable, and sufficiently preventive (Espin et al., 2021; Schmitterer & Brod, 2021).

Through DBDM, teachers receive test results that provide information about which students are still struggling and what types of support are best suited for each student. When DBDM is used as part of support planning, it can improve the quality of teaching and the development of students' reading skills (Schildkamp et al., 2014; Schildkamp et al., 2017). The collected data help teachers determine a student's current position in the learning process and identify the most appropriate intervention. For this approach to be effective, the assessment tools must closely align with support practices. Only then can testing and instruction integrate seamlessly, allowing teachers to derive further support measures more easily and effectively.

1.4. Requirements for the Test Design of an Inclusive Flash Reading Test to Assess the Speed of Lexical Recall in Inclusive Settings

There are numerous tests available to measure students' general reading fluency, but they vary significantly in quality and suitability for inclusive classrooms. In American schools, rating scales are commonly used to measure oral reading fluency, such as the DIBELS Oral Reading Fluency Test (Good & Kaminski, 2002), the NAEP Oral Reading Fluency Scale (Pinnel, 1995), and the Multi-level Academic Skill Inventory—Revised (MASI-R; Howell, 1982). Jungjohann et al. (2018) examined diagnostic instruments that reliably measure reading fluency based on classical test theory. However, these instruments have not been extensively tested for fairness and item difficulty, using methods like item response theory or measurement invariance analyses. Most of the instruments were not developed based on a theoretical model of reading acquisition. Instead, they rely on common models of reading development and target specific grade levels (e.g., MASI-R). Significant floor and ceiling effects can often be observed outside these grade levels. Typically, these test procedures involve oral reading tests where a skilled reader identifies the reading errors of a less skilled reader (Good & Kaminski, 2002). Competence is measured by the number of words read correctly per minute and the corresponding total score. However, it is important

to note that while many of these tests assess oral reading fluency, our flash reading test specifically focuses on word reading fluency. Unlike tests such as DIBELS, which assess the fluency of connected texts, the flash reading test targets the speed and accuracy of lexical retrieval in isolation. This distinction is important because it means that the flash reading test does not capture the broader aspects of reading fluency, such as comprehension and prosody, which are central to oral reading assessments. With the exception of the MASI-R, all tests are offered as closed diagnostic instruments, with no guidance on how to link the results to support materials. The MASI-R, however, follows the recommendations of Hasbrouck and Tindal (Hasbrouck & Tindal, 2006), who suggested that students solving fewer than 50% of the items correctly may require additional support.

Common assessment instruments pose challenges for school practice as they are not economical or accurate in inclusive classrooms (Jungjohann et al., 2018). In general, tests measure most accurately when examinees fall within a normal range (Heine, 2023). The further a result deviates from the mean, the less accurate it becomes, which is particularly problematic when working with students with disabilities. Additionally, some of these status diagnostics for reading for students, especially students with intellectual disabilities, have been criticized for not matching students' levels, thereby failing to provide sensitive information useful for adaptive teaching (Afacan & Wilkerson, 2022). There is a notable lack of test instruments suitable for informing learning and teaching and monitoring progress at lower levels. This gap hinders the ability to provide data-based reading support for all students, highlighting the need for more effective tools in inclusive education.

To effectively and efficiently use tests in everyday school practice, classroom-based instruments and digital test administration are helpful. Conducting tests that do not test oral reading fluency aloud but that test individual word reading fluency and can be conducted within the classroom setting can save significant time, which can then be used for targeted support. However, traditional reading-aloud tests are problematic as reading aloud in a classroom generally impairs concentration and, consequently, the accuracy of the measurement. Therefore, new test formats that can be administered in group settings are needed. Digital tests can offer a solution, enabling the development of new instructional methods and making use of computer-based algorithms, such as random item selection (Klauer, 2006), adaptive algorithms (Ebenbeck, 2023), or automated scoring, making these tests easier for teachers to administer.

2. Research Questions

Measuring and monitoring reading performance is essential for determining support groups and providing individually tailored support in inclusive classrooms. Test instruments need to be easy to use, economical, evidence-based, accurate for all student groups, and useful for support planning. In line with these requirements, the flash reading test instrument was developed to measure the speed of lexical retrieval. Beyond its screening function, the test is designed to provide differentiated insights into subskills of reading, enabling educators to tailor data-based instructional strategies adaptively. This study aimed to examine the design of the test and its impact on a target group of students in inclusive settings based on the following questions:

- Can the flash reading test accurately measure the performance of students with reading difficulties, especially those with learning and intellectual disabilities?
- Which difficulty-generating-item characteristics are most suitable for students with varying levels of word reading fluency?

3. Materials and Methods

3.1. Instrument

To measure the speed of lexical recall, a well-known task format for practicing reading fluency was adapted for use as a digital test. In the flash reading exercise, the teacher briefly displays a card with a word and students quickly read the displayed word aloud within seconds or milliseconds. This exercise was further developed into a test in two steps.

First, an analog test version was created (Jungjohann et al., 2023), which was conducted as a computer-assisted paper–pencil measurement. In this version, students viewed words of varying lengths displayed for two seconds via a slide presentation. After each word was displayed, students selected from four response options the word they had just seen and read. This test was psychometrically evaluated using the Rasch model (Jungjohann et al., 2023), but it did not exhibit the desired range of item difficulties. Additionally, the implementation of the test was cumbersome. The technical setup required extra effort for a slide presentation using a projector and computer. Furthermore, all students saw the items simultaneously, leading many to read the words aloud, potentially biasing the test results. Lastly, the item pool’s difficulty range was not broad enough, lacking more challenging items.

To address these issues, the second digital version of the test was conducted entirely on tablets. Students continued to see one word per item, which appeared briefly on the screen, and then selected the appropriate word from four possible answers. In addition to varying word lengths, the items also differed in their display duration. While in the first version of the test, all words were displayed for 2 s, the second version varied the display duration between 0.5 s, 1 s, 1.5 s, and 2 s. The test consisted of 30 items and had no time limit but was typically completed within three to five minutes. The words included in the test were selected based on commonly used core vocabulary from elementary schools in Bavaria and North Rhine-Westphalia, which reflected the language encountered in everyday school life. This ensured that the words were familiar to students and actively practiced in their instruction.

3.2. Sample

As samples, students were selected who demonstrated the necessary reading skills for the screening, of which the flash reading test was a part, and who were currently receiving reading instruction at school. To be included in this study, students needed to have completed at least the letter acquisition stage. To assess how well the flash reading test could measure the reading performance of students with very weak skills, participants were drawn from inclusive primary schools as well as students from schools for intellectual development at primary and secondary levels. These schools cater exclusively to students with intellectual disabilities, exhibiting a wide range of performance heterogeneity from mild to severe intellectual disability, with approximately 30% of students learning to read (Ratz & Selmayr, 2023). The students from these schools who participated in this study had mild to moderate intellectual disabilities.

In total, 400 students with and without disabilities from inclusive elementary schools (grades 2 to 4, $M_{age} = 8.51$, $SD_{age} = 1.22$) and special schools for intellectual development (grades 2 to 9, $M_{age} = 11.29$, $SD_{age} = 2.02$) completed the digital flash reading test on tablets. Among these students, 303 students had no disability and 97 students had a disability (11 with speech and language impairment, 12 with dyslexia, 31 with learning disabilities, and 43 with intellectual disabilities). Our sample was deliberately chosen to be very heterogeneous in order to check the suitability of our test for heterogeneous, inclusive classes. In Germany, a strong distinction is still made between the different special educational needs, although we are aware that these cannot always be clearly distinguished

from one another. In order to check the suitability for practice, the composition of our sample was nevertheless based on current school statistics for Germany, whereby pupils with SEN were oversampled for clearer results. Students with learning disabilities, dyslexia, and speech and language impairment were in grades 2 to 4, while students with intellectual disabilities were in grades 2 to 9.

3.3. Methods

For the psychometric analysis, the item pool was calibrated using a one-dimensional dichotomous Rasch model with the R package *pairwise* (Heine, 2023), as this method was also used to calibrate the first version of the test. Andersen's Likelihood Ratio Test (LR test) was employed to examine the global fit and global Differential Item Functioning (DIF) between groups with and without disabilities. DIF was assessed at the item level using the Graphical Model Test (GRM) and the Wald Test. Items that disrupted the model were removed and, subsequently, item difficulties and student abilities of those with and without disabilities were calculated and compared.

The test was based on two variables intended to generate item difficulty (so-called difficulty-generating-item characteristics; DGICs): word length and item display duration. To measure the influence of these variables on item difficulty, word length in letters, word length in syllables, and item display duration in milliseconds were analyzed using regression models. Initially, these three variables were examined in linear simple regressions. Subsequently, hierarchical multiple regression models were formed using forward selection. In this process, the variables were sequentially added to the model, with the variable that correlated most strongly with the dependent variable, item difficulty, being included first. Then, the remaining variables were hierarchically added to the model, and the variance in model explanation (R^2) as well as their difference in comparison were analyzed.

In the regression models, the variables were treated as numerical values (word length in letters, syllable count of words, and display duration in milliseconds). However, it is possible that it was not the specific value but rather a categorical classification of the items that was relevant. To test this hypothesis, Linear Logistic Test Models (LLTMs) were created. LLTMs are an extension of the Rasch model, where DGICs can be predefined for each item. This allowed for defining which DGICs influenced item difficulty at the item level.

Two DGICs were defined: for word length, the median word length (6 letters) was assumed and used as a splitting criterion. Thus, items were categorized as either shorter or longer than 6 letters. For display duration, the items were also split into two categories, where items with 500 ms and 1000 ms formed a category of short display duration and items with 1500 ms and 2000 ms formed a category of long display duration.

Three LLTMs similar to the regressions were formed. The first LLTM assumed only word length as a DGIC, the second LLTM assumed only display duration as a DGIC, and the third LLTM assumed both categories as DGICs. The LLTMs were compared to a conventional one-dimensional Rasch model without predefined categorization using the Akaike Information Criterion (AIC) and Bayesian Information Criterion (BIC) as well as conditional log-likelihood to assess model fit.

After the psychometric analysis, the scores of students with disabilities, without disabilities, and for each specific type of disability were descriptively analyzed and compared for the overall test and each individual DGIC category.

4. Results

4.1. Rasch Model Fit

The LR test showed no significant difference, indicating a good overall fit of the Rasch model for the flash reading test. In the Wald test with a disability split, one item stood out as being more difficult for students with disabilities than for students without disabilities. This item was removed from the pool. After its removal, the Wald test showed no deviations in the estimated item parameters between students with and without disabilities (Figure 1A). This suggested that the test measured consistently across all groups and was well suited to accurately detecting weak reading performance, even among students with SEN.

Overall, the item pool had a broad range of item difficulties ($Min = -0.91$, $Max = 2.19$, $M = 0$, $Md = -0.30$, $SD = 0.74$). The measured sample ($Min = -1.90$, $Max = 4.44$, $M = 2.41$, $Md = 2.67$, $SD = 1.44$) showed a performance distribution skewed toward the upper range, indicating a ceiling effect, which was expected given the measured ability (Figure 1B). Students with disabilities were distributed across the ability range but constituted a higher proportion of individuals in the low-achieving range.

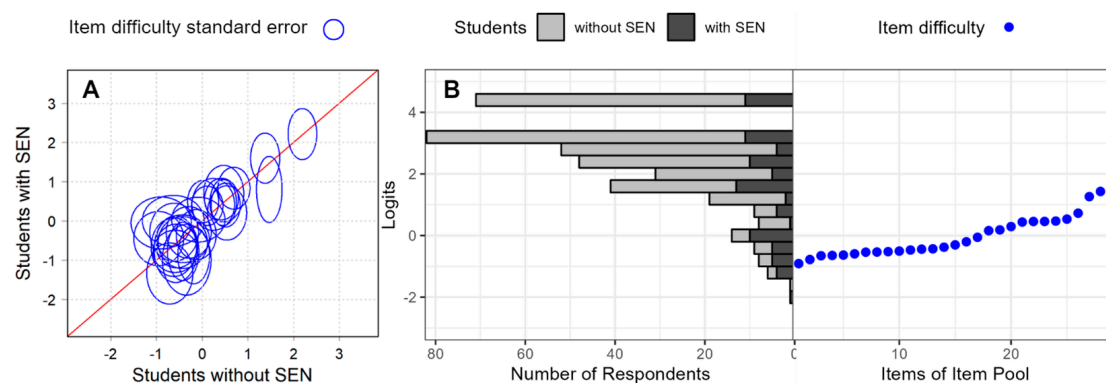


Figure 1. Graphical Model Test (A) and Item–Person Map (B) of the flash reading test item pool.

4.2. Difficulty-Generating-Item Characteristics

The display duration, word length in letters, and word length in syllables showed a different distribution in relation to the difficulty of the items in the item pool (Figure 2). Items with longer display durations tended to be easier than those with shorter display durations. Specifically, items with a display duration of 500 ms were among the most difficult, while items with display durations of 1500 ms and 2000 ms were mainly among the easiest. Items with a display duration of 1000 ms were evenly distributed across the entire difficulty range (Figure 2A).

In terms of word length in letters, very long words with more than seven letters were predominately found in the upper range of item difficulty, whereas very short words with less than five letters were situated in the lower range of difficulty. However, words of moderate length show a wide spread across the center of the difficulty (Figure 2B). The distribution of syllables spanned the entire range of item difficulties, with items containing three syllables primarily occupying the higher end of the difficulty scale (Figure 2C). Overall, the pattern indicated that longer words with shorter display durations tended to constitute the more challenging items in the pool. Conversely, shorter words with longer display durations tended to form the easier part of the item pool, while items with mixed characteristics covered the middle range of difficulty.

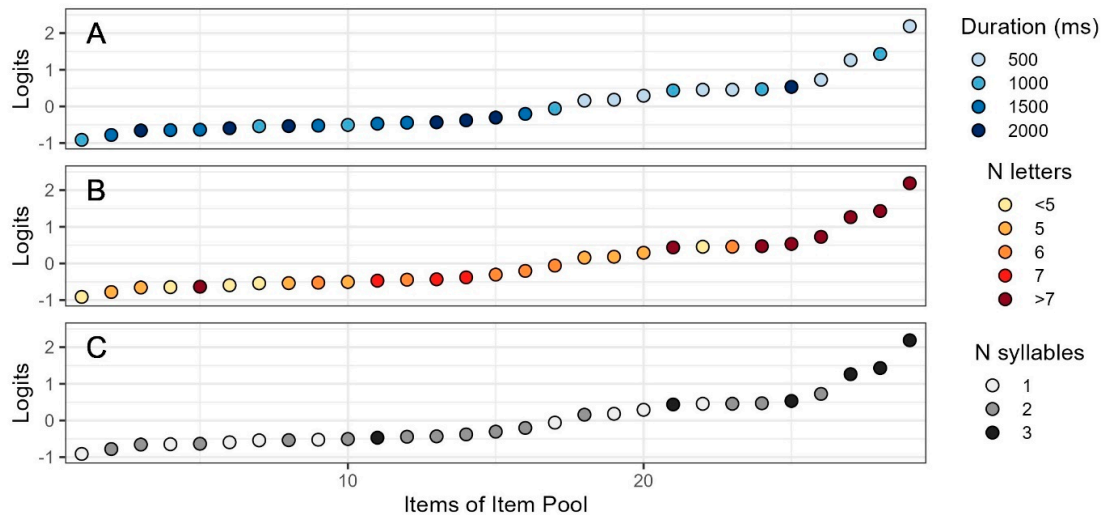


Figure 2. Display duration in milliseconds [duration (ms)] (A), word length in letters [N letters] (B), and word length in syllables [N syllables] (C) of the items in the item pool in relation to their difficulty in logits.

4.3. Regression Analysis

The correlation analysis revealed no multicollinearity among item difficulty, display duration, number of syllables, and word length in letters, as none of the correlations exceeded 0.8. Therefore, all variables could be included in the regression analysis (Table 1). The strongest correlations were observed between word length in syllables and letters, which was consistent with the structural characteristics of the German language.

Table 1. Means, standard deviations, and correlations with confidence intervals.

Variable	M	SD	1	2	3
1. Item difficulty	−0.00	0.74			
2. Duration	1224.14	576.10	−0.59 ** [−0.79, −0.29]		
3. N syllables	1.90	0.72	0.52 ** [0.18, 0.74]	0.06 [−0.32, 0.42]	
4. N length	6.21	1.92	0.66 ** [0.39, 0.83]	−0.08 [−0.43, 0.30]	0.76 ** [0.55, 0.88]

Note. M and SD are used to represent mean and standard deviation, respectively. Values in square brackets indicate the 95% confidence interval for each correlation. ** indicates $p < 0.01$.

In the linear regressions (Table 2), word length in letters accounted for the highest variance in item difficulty, explaining 44.2% of the variance ($F(1,27) = 21.39$, $p < 0.001$, Table 2). Display duration of the items also significantly predicted item difficulty, explaining 35.3% of the variance ($F(1,27) = 14.75$, $p < 0.001$). On the other hand, word length in syllables explained the lowest variance of 26.6% ($F(1,27) = 9.80$, $p < 0.01$).

Table 2. Linear regression results using sigma as the criterion.

Predictor	<i>b</i>	<i>b</i> 95% CI [LL, UL]	<i>beta</i>	<i>Beta</i> 95% CI [LL, UL]	Fit
(Intercept) N Letters	−1.59 ** 0.26 **	[−2.33, −0.85] [0.14, 0.37]	0.66	[0.37, 0.96]	$R^2 = 0.442$ ** 95% CI [0.15, 0.62]
(Intercept) Duration	0.93 ** −0.00 **	[0.38, 1.48] [−0.00, −0.00]	−0.59	[−0.91, −0.28]	$R^2 = 0.353$ ** 95% CI [0.08, 0.56]
(Intercept) N Syllables	−1.00 ** 0.53 **	[−1.70, −0.30] [0.18, 0.87]	0.52	[0.18, 0.85]	$R^2 = 0.266$ ** 95% CI [0.03, 0.49]

Note. M and SD are used to represent mean and standard deviation, respectively. Values in square brackets indicate the 95% confidence interval for each correlation. ** indicates $p < 0.01$.

In hierarchical multiple regressions (Table 3), the combination of word length and display duration together explained 74% of the variance in item difficulty ($F(2,26) = 36.92$, $p < 0.001$), representing an increase of approximately 30% and 40% compared to the respective linear regressions. Word length exhibited a significant positive effect on item difficulty, indicating that longer words contributed to greater item difficulty. Conversely, display duration showed a significant negative effect on item difficulty, indicating that shorter display durations made items more difficult.

Table 3. Multiple regression results using sigma as the criterion.

Predictor	<i>b</i>	<i>b</i> 95% CI [LL, UL]	<i>beta</i>	<i>beta</i> 95% CI [LL, UL]	Fit	Difference
(Intercept) N Letters	−1.59 ** 0.26 **	[−2.33, −0.85] [0.14, 0.37]	0.66	[0.37, 0.96]	$R^2 = 0.442$ ** 95% CI [0.15, 0.62]	
(Intercept) N Letters Duration	−0.63 * 0.24 ** −0.00 **	[−1.26, −0.00] [0.16, 0.32] [−0.00, −0.00]	0.62 −0.55	[0.42, 0.83] [−0.75, −0.34]	$R^2 = 0.740$ ** 95% CI [0.50, 0.82]	$\Delta R^2 = 0.298$ ** 95% CI [0.07, 0.52]
(Intercept) N Letters Duration N Syllables	−0.62 0.19 ** −0.00 ** 0.18	[−1.24, 0.01] [0.06, 0.31] [−0.00, −0.00] [−0.15, 0.51]	0.49 −0.57 0.18	[0.16, 0.81] [−0.78, −0.36] [−0.14, 0.50]	$R^2 = 0.752$ ** 95% CI [0.50, 0.82]	$\Delta R^2 = 0.013$ 95% CI [−0.03, 0.05]

Note: A significant b-weight indicates that the beta-weight and semi-partial correlation were also significant. *b* represents unstandardized regression weights. *beta* indicates the standardized regression weights. LL and UL indicate the lower and upper limits of a confidence interval, respectively. * indicates $p < 0.05$; ** indicates $p < 0.01$.

Therefore, the most challenging items were those with long words and short display durations, while the easiest items were those with short words and long display durations. The influence of these two variables on item difficulty appeared to be equally strong across the entire item pool, suggesting that both variables exerted similar levels of influence.

Including the number of syllables as a predictor did not substantially improve the variance explained by the model. There was no significant difference compared to the two-factor model. The influence of the number of syllables on item difficulty was also not significant in the model. As the number of syllables increased, its effect on item difficulty diminished, while the effect of the display duration became more pronounced.

Since these effects were not statistically significant, they did not contribute significantly to further model clarification. Overall, there was only a redistribution of effects within the model, partly due to the relatively high correlation among predictors. Models with fewer predictors are generally preferred and, thus, the three-factor model was rejected in favor of the two-factor model.

4.4. LLTM Fitting and Analysis

The three modeled LLTMs and the estimated Rasch model were compared regarding their fit to the data and ability to represent the structure of the item pool (Table 4). Results based on conditional log-likelihood, BIC, and AIC indicated that the Rasch model exhibited the best fit, followed by LLTM with two DGICs (Model 3). Compared to the LLTM that assumed and defined only one DGIC (Model 1 and 2), the combination of both DGICs performed better.

Table 4. Log-likelihood and difficulty-generating-item characteristics of the models.

Model	CLL	N Parameters	BIC	AIC	Estimate	
					DGIC 1	DGIC 2
RM	−2609.40	28	5386.56	5274.80	-	-
1	−2911.47	1	5828.93	5824.94	−0.94	-
2	−2906.33	1	5818.64	5814.65	-	−0.98
3	−2795.83	2	5603.65	5595.67	−0.93	−0.97

Note: CLL represents the conditional log-likelihood. A higher CLL indicates a better model fit. AIC represents the Akaike information criterion and BIC represents the Bayesian information criterion. A smaller AIC and BIC indicate a better model fit. DGIC 1 represents the length of words in letters; DGIC 2 represents the duration of items in milliseconds.

Overall, the Rasch model demonstrated the superior model fit. However, it included parameters for each item in the item pool, totaling 28 parameters for modeling, in contrast to the LLTMs, which assumed one or two parameters due to the DGICs. Models with numerous parameters often showed better fit but could overfit with smaller samples, potentially leading to inflated values for the Rasch model due to its complexity. The estimated parameters (“Estimates”) indicated the magnitude and direction of each parameter’s effect in the models. As expected, all LLTMs showed negative estimates, suggesting that the likelihood of a correct answer decreased with increasing difficulty of the item within each DGIC.

4.5. Performance Analysis

In the flash reading test, students with disabilities generally showed lower performance compared to their peers without disabilities (Figure 3). Particularly striking were the significant ceiling effects observed among students without disabilities, with speech and language impairment, or with dyslexia. Although occasional individual students without disabilities achieved lower scores, the majority tended to perform well above average. Students with learning disabilities achieved high scores overall, yet there was notable variability in performance within this group. Notably, students with intellectual disabilities demonstrated the lowest performance levels, accompanied by considerable variance in performance outcomes.

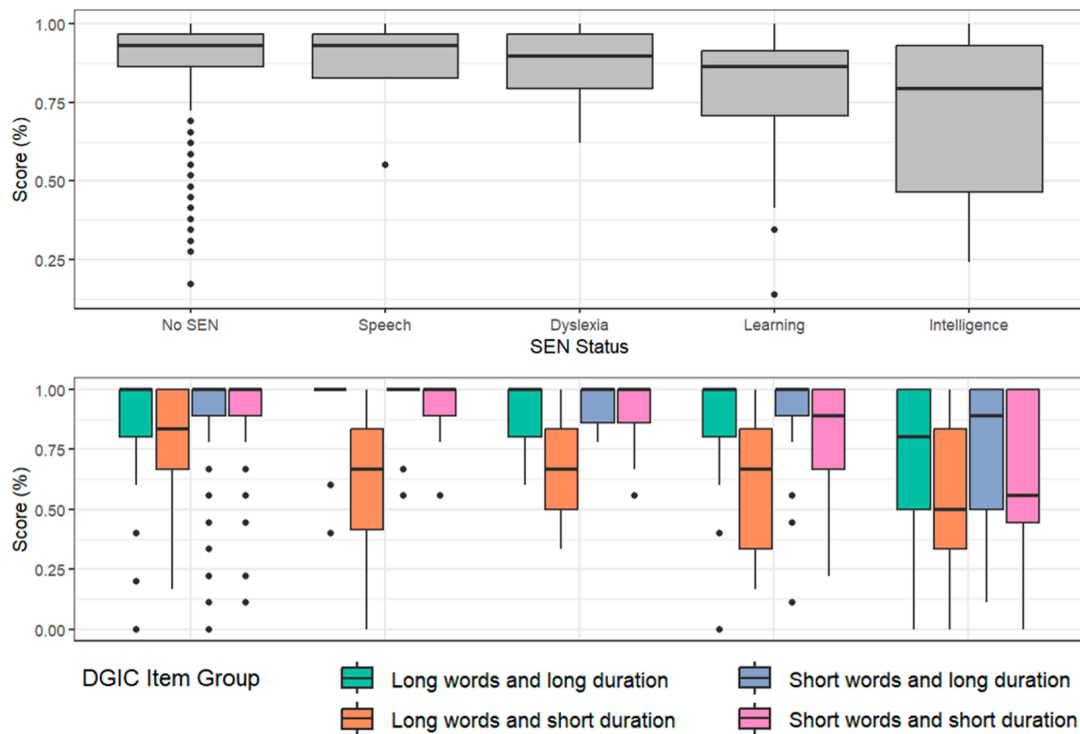


Figure 3. Percentage of correct answers of students with and without SEN in the complete test and different item groups defined by the DGICs of the LLTM.

Long words with a short display duration were consistently the most challenging item category for all student groups, with the lowest mean scores observed across all groups. Conversely, items featuring short words and long display durations represented the easiest category, evidenced by consistent ceiling effects for all student groups except those with intellectual disabilities. For the item categories involving long words with long display durations and short words with short display durations, similar difficulty levels were observed across all student groups, with only minor differences. Specifically, mean scores were identical for students without disabilities and those with dyslexia, although there was slightly broader performance heterogeneity indicated by a larger *SD* for long words with a long display duration. Students with speech and language impairment exhibited a larger *SD* for short words with short display durations compared to long words with long display durations, despite having the same mean score. Students with learning and intellectual disabilities achieved higher mean scores for long words with long display durations than for short words with short display durations. Consistently, students with intellectual disabilities displayed the lowest performance results across all comparisons.

For students with learning and intellectual disabilities, the display duration emerged as the primary DGIC. In contrast, for all other students, word length played a dominant role, although this distinction was somewhat less pronounced. The easiest items, such as short words with long display durations and long words with short display durations, were distinctly identifiable as the easiest and most difficult items, respectively. The results showed that the length and duration of the words shown were important for the different student groups in order to measure them in a differentiated way.

5. Discussion

The findings of this study illuminate the performance disparities observed among student groups with and without disabilities in the newly designed flash reading test, with a particular focus on the influence of word length and display duration as DGICs. The

flash reading test emerges as a valuable educational assessment tool for assessing the speed of lexical recall in inclusive classrooms, effectively capturing the heterogeneous nature of students' reading abilities (Watkins, 2007). Unlike traditional tests such as DIBELS, which focus on oral reading fluency in connected texts, the flash reading test is specifically designed to assess word reading fluency, offering a more targeted measurement of lexical retrieval speed that can inform individualized support planning. The test demonstrates accuracy and fairness in measuring reading performance across all students, where difficulty is equally influenced by letter length and display duration. Several key aspects warrant discussion, including variability in performance across student groups, the role of DGICs, the quality of the test, and its educational implications.

5.1. Performance Variability Across Student Groups

We observed significant performance gaps between students with different types of disabilities and those without disabilities. Students with disabilities generally exhibited lower performance levels compared to their peers without disabilities, which aligns with findings by Schwab and Gasteiger-Klicpera (2014) and Gebhardt et al. (2015). Particularly, students with intellectual disabilities demonstrated weaker performance in the flash reading test. These results are consistent with the findings of Di Blasi et al. (2019), who also highlighted reading fluency as a challenge for students with intellectual disabilities compared to other disability groups.

However, it is important to note that there was substantial overlap in the performance of students with and without disabilities, underscoring that categorizing students solely based on disabilities may not always be conducive to effective educational interventions. Both groups of students can experience reading difficulties, emphasizing the need for comprehensive reading support for all students, especially within inclusive classroom settings. Systematic identification tools, such as the one presented in this study, are critical for recognizing struggling learners early on, ensuring timely and targeted support. By linking assessment to widely used instructional approaches, these tools provide a foundation for effective interventions while highlighting the need for additional, evidence-based methods to address diverse learning needs comprehensively. This underscores the importance of tailored educational approaches that address individual learning needs regardless of disability status.

5.2. Impact of Difficulty-Generating Characteristics

One significant finding of this study was the varying influence of DGICs across different disability categories. For students with learning and intellectual disabilities, the display duration emerged as the primary predictor of item difficulty. This suggests that these students may face greater challenges with items presented briefly, potentially due to slower processing speeds or difficulties in maintaining sustained attention (Ebenbeck et al., 2023). On the other hand, students without disabilities and those with speech and language impairment and dyslexia showed a stronger association between item difficulty and word length, although the clarity of this association varied.

To effectively measure students with intellectual disabilities, it was crucial to include short words in the item pool to appropriately differentiate difficulty levels. By combining display duration and word length, the test effectively identified the weakest 10% of students, supporting its potential use as a screening tool. Using the test for screening purposes can enable the early identification of reading fluency difficulties, particularly in the realm of weak literacy skills, facilitating timely intervention and targeted support allocation (Torgesen, 2002). This differential impact underscores the necessity for customized instructional strategies tailored to addressing specific cognitive and processing needs across

diverse student populations. Moreover, we recommend validating test formats for inclusive settings across various student groups to further elucidate such relationships.

5.3. Ceiling Effects and Item Difficulty

Our results showed significant ceiling effects of the test, particularly prominent among students without disabilities. These students frequently demonstrated high performance levels, particularly on items involving short words with long display durations. The test is designed to assess foundational reading abilities, serving as a prerequisite for developing adequate reading fluency (Ennemoser et al., 2012). Once students achieve sufficient reading fluency, even the most challenging items in the flash reading test can be confidently completed, resulting in the observed ceiling effect.

However, the presence of a ceiling effect does not undermine the test's efficacy for its intended purpose, which is to accurately identify students who have not yet mastered reading fluency. The test effectively distinguishes between students performing at lower levels, making it suitable for pinpointing those in need of additional support and intervention.

5.4. Implications for Data-Based Decision-Making in Education

Criteria-oriented testing aims to pinpoint students who exhibit critical performance levels requiring tailored interventions distinct from those for high-achieving students. In this context, students achieving a hit rate below 75% of all items were classified as having critical performances. This threshold aligns with curriculum standards and the test's design, emphasizing the importance of speed in lexical recall as a fundamental skill of reading fluency.

While occasional errors are expected, consistent inability to master difficult items indicates a need for targeted support to achieve mastery across all item components. This approach ensures that students with critical performance levels are identified promptly, enabling the implementation of personalized interventions aimed at enhancing their reading fluency and comprehension skills effectively (Ennemoser et al., 2012). While the flash reading test provides valuable insights into word reading fluency, it is important to acknowledge that it does not assess the broader aspects of oral reading fluency, such as prosody and comprehension, which are critical for understanding reading performance in connected texts.

The flash reading test offers a distinct advantage by transforming a typical remedial intervention into a diagnostic tool capable of guiding data-driven decisions for intervention and the possibility of progress monitoring (Schildkamp et al., 2017). This dual functionality enables its efficient use in educational settings, supporting both instructional improvements and diagnostic assessments (Watkins, 2007). To further enhance its effectiveness, variations of flash reading exercises can be implemented. For instance, incorporating signal groups, frequently encountered words, and familiar vocabulary can help enhance reading fluency and the speed of lexical recall. Practice with common initial and final syllables can also be beneficial. However, it is important to note that the test itself does not automatically translate into customized instructional strategies. The ability to derive adaptive, child-specific interventions requires teachers to possess trained diagnostic competencies. Without these skills, the potential benefits of the test, such as the early identification of reading fluency difficulties, may not be fully realized. Future research should explore the extent to which the test results can reliably inform targeted instructional planning and intervention design. However, the adaptive nature of the support measures and diagnostics ensures that students are familiar with the procedures, making the test feel more like a beneficial exercise. The test measures only one robust indicator and is therefore also suitable for use for progress monitoring purposes due to its one-dimensionality and random item

selection. This close integration of diagnostics and intervention, when paired with appropriate teacher competencies, has been shown to contribute to improved student outcomes (Carlson et al., 2011).

The digital implementation of this diagnostic tool facilitates the precise and comprehensive assessment of student performance throughout the learning process. Unlike traditional analog methods, digital tools offer detailed insights beyond just sum scores. The use of millisecond time measurements provides nuanced information about performance levels that would be impossible to capture analogously. Moreover, digital platforms allow for test repetition through random item selection (Klauer, 2006) or adaptive algorithms (Ebenbeck, 2023), making them suitable for repeated measurements and long-term learning observations. By eliminating the need for reading aloud, the digital test ensures fairness and inclusivity, making it particularly well-suited for diverse classroom settings. Its application in heterogeneous learning groups enables educators to pinpoint which interventions are effective for individual students and which are not (Praetorius et al., 2013; Ready & Wright, 2011). This streamlined approach underscores the efficacy of a single diagnostic instrument across diverse educational contexts.

5.5. Limitations and Future Research

Acknowledging the potential ceiling effects is crucial from a statistical perspective and should be highlighted as a limitation of the current study. Additionally, it is important to note that our sample included only specific types of disabilities, omitting others such as behavioral or motor disabilities. Furthermore, the relatively small sample sizes in certain subgroups, such as students with speech and language impairment or dyslexia, may limit the generalizability of findings for these groups and contribute to greater variability in the results. This limitation underscores the need for a cautious interpretation of the data and suggests that future studies should aim to include larger and more diverse subgroups to enhance statistical robustness. Future studies should explore the influence of DGICs on students in inclusive settings, particularly those with disabilities in behavior or motor development, to provide a comprehensive understanding. Further research should also delve into the integration of diagnostic and fostering tools, aiming to develop additional remedial interventions tailored for educational settings. The synergy between targeted diagnostics and personalized support facilitates adaptive teaching practices, ultimately enhancing inclusive education that meets the diverse needs of learners. This exploration promises to advance our understanding and implementation of effective educational strategies in inclusive classrooms.

Author Contributions: Conceptualization, J.Z., N.E. and M.G.; methodology, J.Z., N.E. and M.G.; data analysis, N.E. and J.Z.; writing—original draft preparation, J.Z.; writing—review and editing, J.Z., N.E. and M.G.; administration and supervision, M.G. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by German Federal Ministry of Education and Research: 01NV2116D.

Institutional Review Board Statement: 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.

Informed Consent Statement: Informed consent was obtained from all subjects involved in the study.

Data Availability Statement: Data is available here: <https://osf.io/rbz7u/>, accessed on 20 December 2024.

Conflicts of Interest: The authors declare no conflicts of interest.

References

- Afacan, K., & Wilkerson, K. L. (2022). Reading outcomes of students with intellectual disability on statewide assessments. *Journal of Intellectual Disabilities, 26*(2), 195–210. [CrossRef] [PubMed]
- Allor, J. H., Mathes, P. G., Roberts, J. K., Cheatham, J. P., & Champlin, T. M. (2010). Comprehensive reading instruction for students with intellectual disabilities: Findings from the first three years of a longitudinal study. *Psychology in the Schools, 47*(5), 445–466. [CrossRef]
- Al Otaiba, S., McMaster, K., Wanzek, J., & Zaru, M. W. (2023). What we know and need to know about literacy interventions for elementary students with reading difficulties and disabilities, including dyslexia. *Reading Research Quarterly, 58*(3), 313–332. [CrossRef] [PubMed]
- 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*(11), 2443–2448. [CrossRef]
- Berendes, K., Schnitzler, C. D., Willmes, K., & Huber, W. (2010). Die Bedeutung von Phonembewusstheit und semantisch-lexikalischen Fähigkeiten für Schriftsprachleistungen in der Grundschule. *Sprache Stimme Gehör, 34*(2), e33–e41. [CrossRef]
- Blumenthal, Y. (2017). Ein Rahmenkonzept mit mehreren Förderebenen—Response to Intervention (RTI). In B. Hartke (Ed.), *Handlungsmöglichkeiten schulische Inklusion: Das Rügener Modell kompakt* (pp. 20–32). Verlag W. Kohlhammer. ISBN 978-3-17-033540-0.
- Carlson, D., Borman, G. D., & Robinson, M. (2011). A multistate district-level cluster randomized trial of the impact of data-driven reform on reading and mathematics achievement. *Educational Evaluation and Policy Analysis, 33*(3), 378–398. [CrossRef]
- Coltheart, M., Rastle, K., Perry, C., Langdon, R., & Ziegler, J. (2001). DRC: A dual route cascaded model of visual word recognition and reading aloud. *Psychological Review, 108*(1), 204–256. [CrossRef] [PubMed]
- Di Blasi, F. D., Buono, S., Cantagallo, C., Di Filippo, G., & Zoccolotti, P. (2019). Reading skills in children with mild to borderline intellectual disability: A cross-sectional study on second to eighth graders. *Journal of Intellectual Disability Research, 63*(9), 1023–1040. [CrossRef]
- Ebenbeck, N. (2023). *Computerized adaptive testing in inclusive education* [Doctoral dissertation, The University of Regensburg].
- Ebenbeck, N., Jungjohann, J., Mühlhng, A., & Gebhardt, M. (2023). Die Bearbeitungsgeschwindigkeit von Kindern mit Lernschwierigkeiten als Grundlage für die Testentwicklung von Lernverlaufsdagnostik. *Zeitschrift für Heilpädagogik, 74*(1), 29–37.
- Ennemoser, M., Marx, P., Weber, J., & Schneider, W. (2012). Spezifische Vorläuferfertigkeiten der Lesegeschwindigkeit, des Leseverständnisses und des Rechtschreibens. *Zeitschrift für Entwicklungspsychologie und Pädagogische Psychologie, 44*(2), 53–67. [CrossRef]
- Espin, C. A., van den Bosch, R. M., van der Liende, M., Rippe, R. C. A., Beutick, M., Langa, A., & Mol, S. E. (2021). A systematic review of CBM professional development materials: Are teachers receiving sufficient instruction in data-based decision-making? *Journal of Learning Disabilities, 54*(4), 256–268. [CrossRef] [PubMed]
- Galuschka, K., & Schulte-Körne, G. (2015). Evidenzbasierte Interventionsansätze und forschungsbasierte Programme zur Förderung der Leseleistung bei Kindern und Jugendlichen mit Lesestörung—Ein systematischer Review. *Zeitschrift für Erziehungswissenschaft, 18*(3), 473–487. [CrossRef]
- Gebhardt, M., Heine, J.-H., & Sälzer, C. (2015). Schulische Kompetenzen von Schülerinnen und Schülern ohne sonderpädagogischen Förderbedarf im gemeinsamen Unterricht. *Vierteljahresschrift für Heilpädagogik und ihre Nachbargebiete, 84*(3), 246–258. [CrossRef]
- Gebhardt, M., Jungjohann, J., & Schurig, M. (2021). *Lernverlaufsdagnostik im förderorientierten Unterricht: Testkonstruktionen, Instrumente, Praxis*. Ernst Reinhardt Verlag. ISBN 978-3-497-03053-8.
- Good, R. H., & Kaminski, R. A. (2002). *Dynamic indicators of basic early literacy skills* (6th ed.). Institute for the Development of Educational Achievement.
- Grubb, J., & Young, E. L. (2024). Using screening data: Educators' perceptions of a structured data review. *Frontiers in Education, 9*, 1306385. [CrossRef]
- Hasbrouck, J., & Tindal, G. A. (2006). Oral reading fluency norms: A valuable assessment tool for reading teachers. *The Reading Teacher, 59*(7), 636–644. [CrossRef]
- Heine, J.-H. (2023). *Pairwise: Rasch model parameters by pairwise algorithm*. (R package version 0.6.1-0). Zugriff am.
- Howell, K. W. (1982). *MASI multilevel academic skills inventory: Math examiner's manual*. Charles E. Merrill Pub. Co.
- 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. [CrossRef]
- Jungjohann, J., Gegenfurtner, A., & Gebhardt, M. (2018). Systematisches Review von Lernverlaufsmessung im Bereich der frühen Leseflüssigkeit. *Empirische Sonderpädagogik, 10*(1), 100–118. [CrossRef]
- Keuning, T., van Geel, M., & Visscher, A. (2017). Why a data-based decision-making intervention works in some schools and not in others. *Learning Disabilities Research & Practice, 32*(1), 32–45. [CrossRef]
- Keuning, T., van Geel, M., Visscher, A., & Fox, J. P. (2019). Assessing and validating effects of a data-based decision-making intervention on student growth for mathematics and spelling. *Journal of Educational Measurement, 56*(4), 757–792. [CrossRef]

- Klauer, K. J. (2006). Erfassung des Lernfortschritts durch curriculumbasierte Messung. *Heilpädagogische Forschung*, 32(1), 16–26.
- Lenhard, W. (2019). *Leseverständnis und Lesekompetenz: Grundlagen—Diagnostik—Förderung* (2nd ed.). Verlag W. Kohlhammer. ISBN 978-3-17-035020-5.
- 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. [CrossRef]
- Morrison, T. G., & Wilcox, B. (2020). Assessing expressive oral reading fluency. *Education Sciences*, 10(3), 59. [CrossRef]
- Mullis, I. V. S., & Martin, M. O. (2021). *PIRLS 2021 assessment frameworks*. International Association for the Evaluation of Educational Achievement.
- Perfetti, C. (2007). Reading ability: Lexical quality to comprehension. *Scientific Studies of Reading*, 11(4), 357–383. [CrossRef]
- Pinnel, G. S. (1995). *Listening to children read aloud: Data from NAEP's Integrated Reading Performance Record (IRPR) at grade 4*. National Center for Education Statistics, U.S. Department of Education, Office of Educational Research and Improvement. ISBN 978-0886851675.
- 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. [CrossRef]
- Ratz, C., & Lenhard, W. (2013). Reading skills among students with intellectual disabilities. *Research in Developmental Disabilities*, 34(5), 1740–1748. [CrossRef] [PubMed]
- Ratz, C., & Selmayr, A. (2023). Schriftsprachliche Kompetenzen. In D. Baumann, W. Dworschak, M. Kroschewski, C. Ratz, A. Selmayr, & M. Wagner (Eds.), *Schülerschaft mit dem Förderschwerpunkt geistige Entwicklung II (SFGE II)* (1st ed., pp. 117–134). W. Bertelsmann Verlag. ISBN 9783763967834.
- Ready, D. D., & Wright, D. L. (2011). Accuracy and inaccuracy in teachers' perceptions of young children's cognitive abilities. *American Educational Research Journal*, 48(2), 335–360. [CrossRef]
- Rosebrock, C., & Nix, D. (2017). *Grundlagen der Lesedidaktik und der systematischen schulischen Leseförderung* (4th ed.). Schneider Verlag Hohengehren. ISBN 978-3-8340-0314-0.
- Röthlisberger, M., Schneider, H., & Juska-Bacher, B. (2021). Lesen von Kindern mit Deutsch als Erst- und Zweitsprache—Wortschatz als limitierender Faktor. *Zeitschrift für Grundschulforschung*, 14(3), 359–374. [CrossRef]
- Schabmann, A., & Schmidt, B. (2009). Sind Lehrer gute Lese-Rechtschreibdiagnostiker? Der Einfluss von problematischem Schülerverhalten auf die Einschätzungen der Lesekompetenz durch Lehrkräfte. *Heilpädagogische Forschung*, 35(3), 133–145.
- Schildkamp, K. (2019). Data-based decision-making for school improvement: Research insights and gaps. *Educational Research*, 61(3), 257–273. [CrossRef]
- Schildkamp, K., Karbautzki, L., & Vanhoof, J. (2014). Exploring data use practices around Europe: Identifying enablers and barriers. *Studies in Educational Evaluation*, 42, 15–24. [CrossRef]
- Schildkamp, K., Poortman, C., Luyten, H., & Ebbeler, J. (2017). Factors promoting and hindering data-based decision making in schools. *School Effectiveness and School Improvement*, 28(2), 242–258. [CrossRef]
- 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. [CrossRef] [PubMed]
- Schwab, S., & Gasteiger-Klicpera, B. (2014). Förderung der Lesekompetenzen bei Kindern der zweiten Schulstufe—Evaluierung eines differenzierten Sprach- und Leseförderprogramms im Rahmen des Grundschulunterrichts. *Zeitschrift für Bildungsforschung*, 4(1), 63–79. [CrossRef]
- Silverman, R. D., Speece, D. L., Harring, J. R., & Ritchey, K. D. (2013). Fluency has a role in the simple view of reading. *Scientific Studies of Reading*, 17(2), 108–133. [CrossRef]
- Torgesen, J. K. (2002). The prevention of reading difficulties. *Journal of School Psychology*, 40(1), 7–26. [CrossRef]
- Voß, S. (2017). Datenbasierte Förderentscheidungen. In B. Hartke (Ed.), *Handlungsmöglichkeiten schulische Inklusion: Das Rügener Modell kompakt* (1st ed., pp. 33–56). Verlag W. Kohlhammer. ISBN 978-3-17-033540-0.
- Watkins, A. (2007). *Assessment in inclusive settings: Key issues for policy and practice*. European Agency for Development in Special Needs Education. ISBN 9788790591809.

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.