Journal of Research in Reading
UKLA The United Kingdom Literacy Association
J. Res. Read., ISSN 0141-0423
Volume , Issue , 2023, pp 1–19
DOI:10.1111/1467-9817.12436

Classroom effects are as large as grade-level effects on curriculum-based measurement maze reading scores of secondary school students with and without special educational needs

Jana Jungjohann
Faculty of Human Science, University of Regensburg, Regensburg, Germany

Michael Schurig
Faculty of Rehabilitation Sciences, TU Dortmund University, Dortmund, Germany

Markus Gebhardt
Faculty of Human Science, University of Regensburg, Regensburg, Germany

Background: Previous studies used curriculum-based measurement (CBM) maze scores as an indicator of the reading comprehension level of secondary school students with and without special educational needs in multiple grades, pinpointing a high influence of both student- and context-related variables. However, studies on cumulative influence are necessary for better understanding of data-based decision-making.
Methods: We examined a sample of 1066 secondary school students using four linear mixed-effect models: How much variance in maze scores exists between multiple student characteristics (i.e., gender, immigration background, learning disability and developmental language disorder) and context variables (i.e., classroom, grade and school type) across Grades 5–8?
Results: The intra-class correlation (*ICC*) results show that the influence by the context-related variable classroom (*ICC* = .094) is almost as large as by the variable grade level (*ICC* = .126). School type (i.e., inclusive school vs. special school) has the least influence (*ICC* = .02). In addition, the effects of student-related variables explain only a small proportion of the variance (marginal R^2 = .114).
Conclusions: Maze scores can be used as a screening instrument for students with multiple characteristics across grades; they also show that it makes no difference which type of school students attend. As teachers and further classroom-related variables have almost as much influence as grade level, we discuss that teachers can minimise classroom effects by using maze scores as a formative approach.

Keywords: linear mixed-effect models, reading comprehension, screening, secondary school age, special educational needs

Highlights

What is already known about this topic

- CBM maze scores can be used as indicator for reading level as a screening and formative assessment.
- Individual differences at student and context levels uniquely influence reading.
- Teachers use maze scores for decision-making at both an individual student level and a classroom level to prevent reading difficulties.

What this paper adds

- Student variables only are not sufficient to explain variance in maze scores.
- The impact on maze score by classroom membership is almost as large as by grade level.
- The type of secondary school students attend makes no differences in terms of maze scores.

Implications for theory, policy or practice

- To explain variance in maze scores, a joint consideration of the influence by student- and context-related variables is necessary.
- When planning reading instruction, particularly, classroom influences need to be considered.

Screening and formative assessments of reading are among the critical components of improving instruction for all students at secondary level, especially at-risk students and those with special educational needs (SEN; Jungjohann & Gebhardt, 2023). At secondary level, preventive and good reading instruction is still necessary because, in many countries, more than 20% of 15-year-olds are at-risk students in reading who regularly achieve lower than their grade level (Organisation for Economic Co-Operation and Development [OECD], 2019a). Predominantly in multi-tiered systems of support, but also in different school settings (Fuchs & Fuchs, 2006), the curriculum-based measurement (CBM) maze is commonly used to screen the level of reading comprehension at single measurement points (e.g., fall, winter and spring) in a time-efficient way (Tzivinikou et al., 2020). CBM maze is a group-administered silent reading task with short passages (~250 words) in which individual words are replaced with blanks according to set rules (e.g., every seventh word). Students read the passages for a maximum of 5 minutes and fill in the blanks by selecting one of the offered response words. All correctly filled blanks are added as sum scores (i.e., maze score). Teachers use maze scores for decision-making in the form of evidence-based practices at both an individual student level (i.e., formative evaluation of instruction) and a classroom level (i.e., benchmarking to evaluate instruction; Blumenthal et al., 2021) to prevent serious reading difficulties. Thus, it is necessary to understand both individual and context variances within the maze scores before using it for data-based decision-making.

Despite unequivocal evidence of the unique influence of individual differences at student level (e.g., gender, immigration background and SEN) and context level (e.g., classroom and school type) on reading comprehension (Pfost et al., 2014), little is known about whether and to what extent these variables have a cumulative effect across maze scores (Brown-Chidsey et al., 2003; Kim et al., 2015). A joint consideration of influences on maze scores by student- and context-related variables leads to a better understanding of varying reading performances, as, in everyday school life, all influences occur together. Thus, we address this gap in the present study. We examine what degrees of variance in maze scores in German secondary schools can be explained by multiple variables at both the student level and the context level.

Influence of Students' Characteristics on Reading Comprehension

As the Programme for International Student Assessment (PISA) 2018 demonstrated again (OECD, 2019a), reading comprehension varies due to multiple student characteristics. Across Grades 1–10, the largest share of the variance in reading comprehension performance can be explained by students' underlying cognitive skills (e.g., language proficiency, working memory and word recognition; Görgen et al., 2021; Kieffer & Christodoulou, 2020). However, individual differences in cognitive processes are not sufficient to explain all kinds of variance in reading comprehension at secondary level.

Two subgroups of students who receive SEN services are likely to show serious difficulties in reading comprehension: students with learning disabilities (LD) and those with developmental language disorders (DLD). Students with LD show most often extensive difficulties in basic academic skills and may have a lower intelligence quotient compared with typically developing children. On an international scale, the vast majority of these students perform significantly below their grade level (Gebhardt et al., 2015; National Assessment of Educational Progress, 2022). Students with DLD can have speech, language and communication needs. Based on a discrepancy model, their general cognitive ability is within average range in the absence of other development difficulties that could impact their difficulties (Spreer et al., 2019). As greater attention is paid to the individual support and educational accommodation needs, these students are now being classified as having DLD, previously known as specific language impairment (Curran & Hogan, 2023). Reviewing literature on students with DLD regarding literacy outcomes, Curran and Hogan (2023) concluded that the risk for reading difficulties varies greatly depending on the degree of the language impairment.

Affiliation to gender and ethnic minority background are additionally associated with difficulties in reading comprehension. Gender difference arises for students with and without SEN from the fact that boys are mostly significantly overrepresented in the low reading achievement areas, whereas the proportion of girls is marked in the high reading achievement areas (Lepper et al., 2021; Logan & Johnston, 2009). In Germany, scientific studies with students commonly record a potential immigrant and/or ethnic minority background, meaning that the students or at least one parent were born abroad, or the students were born in Germany, but do not possess the German citizenship (Will, 2019). Immigrant background is perceived to be strongly related to the socioeconomic status of the students' families in Germany. In the German PISA 2018 cohort, about half of the students with an immigration background achieved the lowest proficiency levels and thus an insufficient

level of reading comprehension (OECD, 2019b). This means that these students achieved a similarly low reading performance as students with LD.

Influence of Classrooms and School Types on Reading Comprehension

The conditions in which students learn give rise to context variables that affect reading comprehension. Classroom dynamics are complex and encompass relationships and interactions between teachers and students, among students as well as the perceptions, attitudes and behaviours of students and teachers (Montague & Rinaldi, 2001). This includes the fit between the requirements of the curricular or individualised education plan (i.e., individualised programme to ensure that students receive specialised instruction and related services), students' abilities and needs, and the interpretation of both by teachers. Variances in teachers' interpretation translate into concrete classroom actions, which, however, address students to varying degrees and are effects of teaching experience, treatment adherence, and methods and materials used (Capin et al., 2021). As a consequence, classroom can be understood as an effects' summary of the student group as a social unit together with the teacher's actions. A consideration at classroom level aggregates the effects independent of its individual members, and a view at the level of the individual has the same effect in each case.

The type of school a student attends also differs in many ways such as the degrees of inclusive education, environment or resources. Therefore, cumulative effects on students' reading comprehension can be expected. On an international scale, studies conclude that a placement in a regular school that is also attended by students with SEN (i.e., inclusive school) has neutral or positive influences on the achievement levels of students without SEN (Kalambouka et al., 2007). For students with SEN, school conditions vary even more because there are multiple school types they can attend. Empirical findings suggest that students with SEN achieve higher outcomes in inclusive schools (Lindsay, 2007; Myklebust, 2002). Special schools that are attended only by students with an official diagnosis of their current school support needs still exist in some countries such as Germany (Ebenbeck et al., 2022). They were established assuming that they would provide students with SEN with an optimal learning environment. With regard to academic achievement, there is growing evidence that this assumption can no longer be sustained (Stranghöner et al., 2021).

Cumulative Influence on Maze Scores

CBM maze is a formative assessment approach with the goal to identify at-risk students in reading comprehension (Deno, 2003). It measures reading-related skills as a robust indicator that allows to draw conclusions about the level of reading comprehension without providing further information about specific processes of reading comprehension at word, sentence or text level (Tzivinikou et al., 2020). It is established because of its teacher-friendly and time-efficient administration. Multiple studies showed that maze scores are moderately to strongly related to reading comprehension in all grades with technical adequacy (e.g., Chung et al., 2018; Kim et al., 2015). Contrary to the earlier assumption that CBM maze measures text comprehension, recent research suggests that maze scores can be explained more strongly by code-related skills such as decoding and fluency (Amendum et al., 2021; Muijselaar et al., 2017; Shin & McMaster, 2019) and parsing

processes at sentence level (Anderson et al., 2020; January & Ardoin, 2012). Thus, CBM maze is also referred to as a silent reading fluency measure but does not claim to comprehensively map all components of reading comprehension students require at secondary level. Advanced CBM enhances instruction through the use of strategically created items and/or distractors to increase sensitivity to specific reading skills. Advanced CBM maze uses linguistically selected blanks instead of fixed-ratio blanks to assess core components of text comprehension beyond the word and sentence levels (e.g., Jensen & Elbro, 2022; Jungjohann et al., 2018). For both advanced CBM mazes, correlations were reported indicating a strong relation with an established measure of reading comprehension at text level (Anderson et al., 2020; Jensen & Elbro, 2022). Such approaches offer great potential for reading comprehension instruction as called for by MacKay et al. (2021) because they can provide teachers with information about specific processes in reading comprehension.

The quality of a CBM should be tested using item response theory in addition to classical test theory (Schurig et al., 2021), as it allows for testing multiple parameters such as item homogeneity for a one-dimensional test interpretation and the consideration of the speed component in the students' response behaviour. CBM maze can be administered across multiple grade levels at single measurement points as a screening, or multiple times to monitor learning progress. Both require alternative multiple test forms and sensitivity to student growth. Therefore, CBM maze has to be both short enough to be useable during lessons and long enough to inform about the actual performance level and the rate of progress. CBM mazes are speed tests rather than power tests because of their time limit. Speed test means that students with higher competencies solve more items while weaker students complete fewer items correctly (van Breukelen, 2005). Due to the time limit and to prevent ceiling effects, the item pool of a speed test should contain more items than a target person with the highest competence can solve.

To our knowledge, two studies have researched cumulative variance in reading comprehension attributable to student and context influencing factors using the CBM maze as a screening tool. Both studies used the sum scores of a single measurement point, included secondary level students as participants and indicated that student characteristics and context factors moderate cumulatively the predictiveness of maze scores. The former study highlights that grade level, regardless of student characteristics, is relatively strong at predicting the maze score at secondary school level (Brown-Chidsey et al., 2003). Brown-Chidsey et al. (2003) estimated the variance contributions of students across Grades 5–8 with three different parallel forms. All participants ($N = 476$) attended the same school across six classrooms at each grade level. Seven percent of the participating students received SEN services. In two CBM mazes, most of the variance was resolved by grade level (68–71%), while individual differences (10–14%) and SEN status (9–10%) were accountable for similarly little variance. In contrast, for the third CBM maze, individual differences accounted for 84%, grade level for 6% and SEN status for 4% of the variance. The authors suspect non-homogeneous parallel forms as an explanation for the different degrees of variances. Brown-Chidsey et al. (2003) did not conduct analyses of classroom membership or further student-related factors (i.e., gender or specific SEN), which limits the study results. The more recent study included more student- and context-related factors but did not examine variance across grade levels. Kim et al. (2015) figured out that most of the variance in the maze score of Grades 3–10 originates from student characteristics (48–67%). Participants were more than 1 million students across multiple classrooms, schools and districts in Florida, USA. The lowest proportion of variance was due to school and district levels (3–4%). However, the influence due to classroom varied for grade levels. In Grades 3–5,

variance due to classroom (21–23%) was almost three times lower than variance due to individual differences (66–67%). In Grades 6–10, variance due to classroom (41–46%) and that due to student characteristics (48–51%) were similar. Despite the sample being comprehensive, it still limits the conclusions at student level. In the analyses of Kim et al. (2015), 55% of the students belonged to an ethnic minority and 15% were identified with a primary exceptionality (i.e., students with special gifts or SEN). All students were included, but there was no differentiated consideration of the influence at student level.

Present Study

Previous research has shown both student and context factors to be associated with students' reading comprehension outcome across grades and the CBM maze to be a reliable and valid measure of the level of reading comprehension. Notably, studies that simultaneously examined multiple student characteristics in the context of multiple classrooms across grades are lacking. Such studies could investigate whether CBM maze is more predictive for some students than others and would provide guidance for identifying at-risk students in reading comprehension using maze scores. With such results, teachers should be able to make more informed decisions taking the intertwined nature of context factors and students' characteristics into account for reading instruction. In other words, teachers receive information about how much weight to assign to single influencing factors in their decisions. The purpose of our study is to expand the understanding of the variance of maze scores in secondary schools. Thus, we follow and expand previous research (Brown-Chidsey et al., 2003; Kim et al., 2015) to figure out the sources of variance in fifth through eighth grade students' maze scores. We ask: How much variance in maze scores exists for multiple student characteristics (i.e., gender, immigration background, LD and DLD) and context level (i.e., classroom, grade and inclusive or special type of school) across grades?

Based on the literature, it was hypothesised that the greatest variance would be explained by grade and student characteristics.

Methods

Participants

Participants were a total of 1066 students in Grades 5–8 from northwestern Germany. They were enrolled in 67 classrooms in 16 schools. Of these, 30 classrooms in seven schools were segregated special schools. Seven types of special schools with different focuses on SEN still exist in the German school system in addition to inclusive schools (Ebenbeck et al., 2022). The participants' teachers provided information about students' age (M = 13;1 years), gender (41.7% were female), immigration background (47.19%) and SEN (36.77%). A total of 23.73% of the students were officially diagnosed with LD and 13.04% with DLD. In accordance with the local school law, students with LD failed in all core school subjects, had a lower intelligence quotient and were taught according to an individualised education plan. Students with DLD were taught according to the curriculum of the actual grade level and had extensive needs in speech, language and communication. One hundred forty-four students were not present due to illness or other reasons at the second measurement (see below). Their data were treated as missing and were not

Table 1. Sample characteristics by grade level.

Grade	$n_{Students}$	$n_{Schools}$	$n_{Classrooms}$	% Male	M_{Age}	% Immigration background	% SEN
Fifth grade	207	5	13	57.00	11;8	57.97	32.36
Sixth grade	373	11	23	58.71	12;5	57.64	30.30
Seventh grade	244	7	16	59.02	13;5	36.48	45.49
Eighth grade	242	6	15	61.98	14;5	32.64	41.73

Note: $N = 1066$. The beginning of the school year 2019/2020 was taken as the reference for the average age. Abbreviation: SEN, special educational needs.

included in the analyses. A grade-specific summary of the participants is given in Table 1. Ethical aspects of the research were reviewed and approved by the second author's faculty institutional review board (Ethics Commission of TU Dortmund University, Department of Rehabilitation Sciences). Participation was voluntary and supervised by school staff. All data were collected with the informed consent of the participants, parents, teachers and administrators during any given lesson without relation to the subject or the teacher.

Procedures

Local school administrators were contacted by trained research assistants. All students completed the CBM maze twice, 3 weeks apart. The 3-week interval between the measurement points was chosen to ensure that the knowledge of how to handle the test was still present. The participating schools also opted for the 3-week interval for organisational reasons. Following the guidelines of Hessels (2009) that students with learning difficulties need to be familiarised with a new test to be able to respond to the items in the way that is expected for testing, the first measurement point was used as an exercise to ensure that the students were familiar with the web-based test.

At both measurement points, the research assistants tested participants in classroom groups in the presence of a teacher. Each student worked individually on their android tablet. All research assistants followed the same scripted procedure. First, the participants were given an example item to explain how to use the test. Once participants had answered it correctly, they answered as many items as they could within a 5-minute period. To collect the data for the second measurement, the same research assistants returned to the classrooms. Outcome measure was administered in groups in classrooms by research assistants, and student and context variables were surveyed with classroom teachers.

Outcome Measure

A web-based advanced CBM maze with strategically chosen blanks was used to assess students' silent reading fluency (Jungjohann & Gebhardt, 2021). Test administration takes places online via a web-based platform for CBM monitoring, called www.levumi.de (Jungjohann et al., 2018; Mühling et al., 2019). The platform is university hosted and free from publishers' paywalls. The item pool contains 93 items. Each item is created carefully to consider one of the three reading comprehension-related dimensions: (1) syntactic skills, (2) inference and (3) coherence making at sentence level. In each parallel form, the items from the three dimensions alternate evenly. The deletion pattern follows linguistic rules to

guide comprehension analysis. Students can only identify the target word if they have understood both the accompanying and subordinate clauses and can link their contents. For each item, three distractors are created, with all three of them being contextually meaningless but syntactically possible. Students receive 1 point for each accurate choice made within a time limit of 5 minutes.

This item design controls sentence complexity and strengthens syntactic skills at sentence level as it is critical both for making meaning from text and for the instruction by teachers in line with MacKay et al. (2021). This provides teachers with detailed information about which sentence-related comprehension processes are currently causing students the most difficulty, in addition to the sum score (Jungjohann, 2022). For instruction, this additional information is valuable because it can be easily applied to targeted exercises for understanding complex subordinate clauses, drawing inferences and coherence. The following examples (translated from German) illustrate item and distractor construction. The underlined part represents the blank. The first word is the target word:

Dimension – Complex subordinate clauses: Paul's girlfriend, who loves/hates/teaches/carries him very much, still doesn't want to marry him.
Dimension – Inference making: The orchestra begins the concert. The audience falls silent/chatters/guesses/sings directly.
Dimension – Coherence making: Every Saturday, there is a farmer's market in the city. Many stalls/benches/hours/fields sell fruits and vegetables.

Previous research confirmed the psychometric quality for students ranging in age from 11 to 16 (i.e., fifth to eighth graders; Jungjohann, 2022). Cronbach's alpha was .93. In the sample of this study, the CBM maze had acceptable reliability in the Rasch model, with Warm's weighted likelihood estimates being *WLE* = .741. Therefore, we concluded that the Rasch models fit the data well and that the sum scores can be used unidimensionally for further analyses.

Variables at Student Level

The three variables gender, immigration background and the presence of SEN were collected. Gender was coded as female, male and diverse. Following the procedure in studies with German school students (Will, 2019), it was asked whether there was an immigration background or not. The presence of SEN was divided into no SEN services, LD and DLD.

Variables at Context Level

At context level, the four variables grade level, classroom, school and school type were collected. It was documented which classes from which grade level belonged to which school. The schools were classified into three categories according to whether they were special schools for students with LD, whether they were special schools for students with DLD or whether they offered instruction for students with and without SEN (i.e., inclusive schools). As the concept of the school was evaluated in terms of inclusive teaching, it was well possible that there were classrooms without students with SEN as well.

Data Analysis

Students' answers to the maze tasks were automatically stored by the platform www.levumi.de and coded correct and incorrect. Therefore, no scoring errors were to be expected. Due to there being 93 items in the item pool of the speed test, there was a theoretical maximum of 93 points as a total score. In order to control effects on scores due to random guessing, all answer patterns with the maximum of 93 answers were excluded ($n = 7$). The individual student's sum score was computed using Gnu R (R Core Team, 2020).

To estimate the variance in the sum scores, we followed the guidelines of Finch et al. (2014). The data and syntax are available under https://osf.io/fgh4d/. We conducted four linear mixed-effect models using the package lme4 Version 1.1-26 (Bates et al., 2020). In all model analyses, restricted maximum likelihood estimation (*REML*) was applied (Kreft & de Leeuw, 2011). The influence by the schools was controlled by the context variables school type and classroom. This was done for two reasons. First, the classroom variable explains more variance due to the explicit affiliation of teachers. Second, the school variable explains the influence of the district the schools are located in, and, in Kim et al.'s (2015) study, the influence to be explained by the school variable was marginal. Instead, we included the variable school type, for which larger effects were to be expected.

As Model 1, we specified a simple random intercept model as an unconditional model (Finch et al., 2014). This step provided the information about variance in the sum scores due to each of the clustering levels. The unconditional model was evaluated to determine that differences in students' sum scores warranted the inclusion of the further covariances. From this, we calculated intra-class correlation (*ICC*). *ICC*s reflect the proportion of Level 2 variance in the total variance in the dependent variable, which occurs across different units of clustering. A significant intercept variance and substantial *ICC* are preconditions for using multilevel modelling because, without differences in the dependent variable, there would be no need to explain those differences by further variables. According to Hox et al. (2018), *ICC*s above .05 are sufficient variation to perform a multilevel analysis. Next, we built an individual-level model by adding students' variables at Level 1 as fixed effects (Model 2). Afterwards, 2 two-level models with random intercept and fixed slope were generated including as on Level 1 predictors all variables at student level (i.e., gender, immigration background and SEN). At Level 2, in addition to the school type variable, in Model 3 grade and in Model 4, both grade and classroom variables appear in the aggregate model. Interaction effects between variables were not calculated due to cell frequencies being too low under some conditions.

Results

Descriptive Statistics

A summary of the number of participants, means, standard deviations, minimum and maximum for the maze score is presented in Table 2. The evaluation of the descriptive data shows in most cases that higher grade-level students completed more items within the 5-minute testing time than lower grade-level students. In all cases, this increase is true for students without SEN, for those with LD and for the total number of students of each grade, with the exception of the sixth graders with DLD. They solved fewer items than the fifth graders.

Table 2. Sum scores by grade level and SEN services.

Grade	*n*	*M*	*SD*	Min	Max
Fifth grade					
Students without SEN	134	13.13	6.44	0	37
LD	36	10.06	5.77	0	25
DLD	25	17.68	7.43	7	33
Total	195	13.14	6.76	0	37
Sixth grade					
Students without SEN	232	15.35	6.69	1	32
LD	63	11.08	5.66	2	24
DLD	27	13.37	5.21	4	26
Total	322	14.35	6.59	1	32
Seventh grade					
Students without SEN	96	17.61	6.31	4	34
LD	75	13.01	5.73	1	29
DLD	34	19.97	9.13	2	40
Total	205	16.32	7.14	1	40
Eighth grade					
Students without SEN	108	20.22	6.14	4	38
LD	59	14.63	7.01	1	32
DLD	33	22.09	10.15	2	51
Total	200	18.88	7.69	1	51

Note: $N = 922$. Abbreviations: DLD, developmental language disorders; LD, learning disabilities; *SD*, standard deviation; SEN, special educational needs.

Furthermore, one-way analysis of variance (ANOVA) results indicate that the means' average varies significantly among student groups across grades ($F(2, 918) = 26.74$, $p < .001$). The group of students with LD solved the fewest items on average across all grades. In Grades 6–8, a Bonferroni-adjusted post hoc analysis revealed no significant differences ($p > .05$) in the performance of students with DLD and students without SEN.

Measurement Models

Results from all models are shown in Table 3. The *ICC*s of Model 1 indicate that modelling of covariates was necessary (*ICC* = .161). Model comparison indicates that Model 4 is the best fitting model ($\Delta AIC = 27.4–163.4$, $\Delta BIC = 22.6–134.5$). In addition, the likelihood ratio test confirms that Model 4 describes the data significantly better than the most similar Model 3 ($\chi^2(9) = 29.12$, $p < .001$). Following the parsimony principle, we decided to interpret the variance components in Model 4.

Table 3. Linear mixed-effect models.

	Model 1				Model 2				Model 3				Model 4			
	Estimate	*SE*	95% CI		Estimate	*SE*	95% CI		Estimate	*SE*	95% CI		Estimate	*SE*	95% CI	
Effect			*LL*	*UL*			*LL*	*UL*			*LL*	*CL*			*LL*	*UL*
Fixed effects																
Intercept	15.25*	1.80	11.13	19.36	18.55***	0.84	17.55	19.87	18.48***	1.56	15.38	21.59	18.30***	1.47	15.91	21.42
Sex_2[a]					−0.80	0.46	−1.74	0.06	−0.86	0.44	1.74	0.00	−0.75	0.43	−1.62	0.07
SEN_2[b]					−3.65**	0.83	−5.42	2.66	−3.97***	0.88	−5.82	−2.25	−3.88***	0.83	−5.72	−3.19
SEN_5[b]					0.39	1.08	−1.11	1.86	−0.07	1.13	−2.24	2.09	0.33	1.08	−1.55	1.85
Mig_1[c]					−3.78***	0.49	−4.78	2.88	−2.92***	0.48	−3.91	−2.01	−3.17***	0.54	−4.27	−2.16
	Variance	***SD***	***ICC***		**Variance**	***SD***	***ICC***		**Variance**	***SD***	***ICC***		**Variance**	***SD***	***ICC***	
Random effects																
School type intercept	9.51	3.09	.161		0.75	0.86	.016		1.39	1.18	.032		0.78	0.88	.02	
Grade intercept									5.82	2.41	.121		5.60	2.37	.126	
Classroom intercept													4.03	2.01	.094	
Residual	49.70	7.05			46.05	6.79			42.30	6.50			38.71	6.22		
$N_{\text{Observations}}$	922				922				922				922			
$N_{\text{School type}}$	3				3				3				3			

(*Continues*)

Table 3. (Continued)

Variance	*SD*	*ICC*	Variance	*SD*	*ICC*	Variance	*SD*	*ICC*	Variance	*SD*	*ICC*
Random effects											
N_{Grade}						4			4		
$N_{Classroom}$									67		
Model fit											
AIC	6234.3		6158.9			6098.3			6070.9		
BIC	6248.8		6192.7			6136.9			6114.3		
Marginal R^2	.000		.114			.090			.097		
.161			.128			.222			.289		Conditional R^2

Note: Root mean square error of approximation (*REML*) estimate. Abbreviations: *AIC*, Akaike information criterion; *BIC*, Bayesian information criterion; CI, confidence interval; Conditional R^2, variance explained by fixed and random effects; *ICC*, intra-class correlation; *LL*, lower limit; Marginal R^2, variance explained by fixed effects; *SD*, standard deviation; *SE*, standard error; SEN, special educational needs; *UL*, upper limit.

[a]1 = female, 2 = male.

[b]1 = no SEN service, 2 = learning disabilities, 5 = developmental language disorders.

[c]0 = no migration background, 1 = migration background.

$^{*}p < .05$.

$^{**}p < .01$.

$^{***}p < .001$.

Cumulative Variance

Marginal and conditional R^2 are shown in Table 3 for all models. Both types of pseudo-R^2 can be interpreted following Cohen's (1988) criteria: 0.02–0.12 is considered a small, but practically important effect, 0.13–0.25 is moderate and ≤0.26 is large. To interpret the unique effect of the variables at student level, the individual-level Model 2 must be used. Marginal R^2 = .114 indicates that the student variables only have a small effect on the variance of the maze score. The conditional-explained variance in Model 4 shows a large effect (R^2 = .289), meaning that the cumulative influence by both student and context variables explains almost twice as much percentage of variance compared with Model 2.

Effects at Student Level

The results of Model 4 show that the mean maze score is 18.30. No significant effects are estimated across students with DLD and gender. Significant effects are in intercepts across students with LD ($t(13.36) = -4.65$, $p < .001$) and an immigration background ($t(543.63) = -5.82$, $p < .001$), meaning that there is a −3.88 decrease in students' maze score for LD and a −3.17 decrease for students having an immigration background.

Effects at Context Level

In Model 4, with a random intercept for school type, grade and classroom, *ICC* values were computed as measures of effect size for random effects (Hox et al., 2018). The greatest variance is explained by grade (*ICC* = .126), followed by classroom membership (*ICC* = .094) and school type (*ICC* = .02). Compared with the total amount of variance in this model, grade level and classroom together explain almost all the variance. These results indicate that it makes no difference in terms of maze scores which type of school students attend but that classroom membership has almost as much influence as grade level. To visualise this influence, Figure 1 shows boxplots of the sum scores for each classroom sorted by median and coloured by grade level. In particular, the medians of all six classrooms are striking, because they are spread across the board.

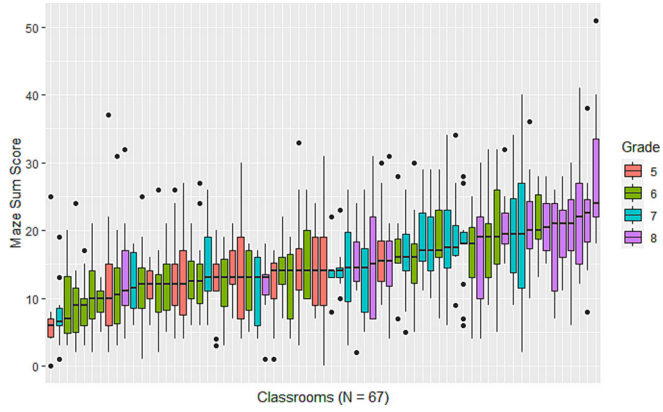

Figure 1. Boxplots of the maze sum scores for each classroom sorted by median and coloured by grade level.

Discussion

This study examined cumulative variances by student- and context-related variables with respect to maze scores in secondary school with the goal of providing a better understanding of the data for classroom use. Following two related studies (Brown-Chidsey et al., 2003; Kim et al., 2015), the maze score from one measurement point was analysed as a screening. As an extension of these previous studies, relevant student- and context-related variances were examined together to account for cumulative effects.

In school-based diagnostics, a consideration of context effects beside student-related variables referring to the level of reading comprehension is necessary, as confirmed by our model comparison and by the results of Kim et al. (2015). In our study, the most complex Model 4, which includes the student-related variables gender, SEN and immigration background at Level 1 and the context-related variables school type, grade and classroom at Level 2, fitted the data best, as it explained the greatest variance (conditional R^2 = .289). Blumenthal et al.'s (2021) study indicates that individual factors are predominantly included in the interpretation of diagnostic data. Especially in special education, needs are mainly determined by students' individual factors. Considering context and individual student variables together is particularly important for countries with varying school conditions such as Germany. German students with SEN can attend a special school focusing on one or more different types of SEN or inclusive classrooms (Ebenbeck et al., 2022). How schooling is realised under these circumstances is highly different, as each state has its own federal education system comprising both special schools and inclusive education. Additionally, German teachers are not obliged to use evidence-based practices in their classrooms. In our study, a large proportion of the variance was explained by classroom effects (*ICC* = .094), whereas the influence by school type was the smallest (*ICC* = .02), implying the strong relation of the individual teacher and further classroom-related variables to students' maze score. One possible explanation is that, in the schools, classrooms were formed based on student achievement levels. Alternatively, the large proportion of the classroom effects could be caused by the quality of the support provided by individual teachers. Through the use of CBM instruments as both a screening approach and a formative approach, teachers can gain insight into classroom effects related to the level or rate of progress of reading comprehension and development by administering easy-to-use tests (Deno, 2003). Capin et al. (2021) report that teachers' instructional quality has a significant impact on reading comprehension. In daily instruction, teachers can use the advanced CBM maze along with other summative and qualitative reading assessments to gain information about individual reading processes, reading motivation and reading behaviour. They can then use this assessment information to make data-based decisions to support individualised reading comprehension instruction. By using more effective instructions, teachers can reduce classroom effects.

A further indication of strong classroom effects is the varying average performance in the sixth grade. The median sum scores of sixth graders scatter across the range of all included grades (see Figure 1). Although most of the variance in reading comprehension is explained by grade level (*ICC* = .126), the medians of the sixth grades are distributed between the medians of all other grades. Of particular interest are the results of Kim et al. (2015) who also observed a break in the influence of grade level for the sixth grade with American students. Beginning in the sixth grade, grade level had a stronger influence on reading achievement (*ICC* = .41–.46) than in Grades 3–5 (*ICC* = .12–.23). Both Kim

et al. (2015) and our studies suggest that, at this age, students are particularly susceptible to influences exerted by teachers and classroom variables.

The CBM maze was successfully completed by students of all grade levels because the effects of student variables accounted for only a small proportion of the variance in the sum scores (Model 2; marginal R^2 = .114). As expected, in line with previous research findings (e.g., Gebhardt et al., 2015; OECD, 2019b), significantly lower intercepts were observed in the maze scores of students with LD and immigrant backgrounds. With respect to students with DLD, the research is inconclusive (Curran & Hogan, 2023; Spreer et al., 2019) but rather suggests significantly lower intercepts. In our study, no significant differences were observed in comparison with students without SEN, indicating a particularly high level of achievement among the participating students with DLD. In Germany, about half of the students with DLD are taught in special schools (Kultusministerkonferenz, 2020). In our sample, the majority of them attended a special school only for students with DLD so that a sufficient number of them with this particular student variable could participate in the study. One possible explanation is that reading instruction in general has a particularly high priority in the special school, as all students have speech, language and communication needs leading to intensive support in reading, resulting in above-average reading achievements among the students. On the one hand, our observations confirm that research on reading skills and reading support for students with DLD should be intensified. On the other hand, our results are a further indication that the CBM maze can be used for students with multiple characteristics as a reading screening tool.

Limitations and Future Directions

The study design and the recruitment of the sample limit the validity of the study results. Based on Kim et al. (2015) and Brown-Chidsey et al. (2003), we used a cross-sectional design with one measurement time point and CBM maze. As a result, high influences of classroom membership on the maze score as a screening tool were observed. It remains unclear whether the results can be replicated in longitudinal use for progress monitoring. Additionally, the design does not provide deeper insights into the validity of the response format of the CBM maze as an indicator of students' skills. For this purpose, a parallel use of a standardised reading test that validly measures various reading comprehension processes would be necessary. The sample was recruited in only one German state through voluntary participation by schools and students. Due to the complex regulations for school studies in Germany, it was not possible to draw the sample randomly. Therefore, the study results cannot be generalised. Nevertheless, large-scale studies such as PISA show comparable *ICC*s and study results (e.g., OECD, 2019b). Due to the limited sample and too low cell frequencies, no further investigation of random effects could be conducted within the framework of this paper. A joint investigation of the influence of SEN and the school type would have been useful. For such an investigation, a larger sample of students in inclusive classrooms is needed. One methodological limitation is that Level 1 and Level 2 effects were considered separately; all correlations were linearly modelled so that non-linear correlations were not controlled.

Another limitation is the lack of information on further variables in classrooms. Due to previous findings (e.g., Capin et al., 2021), a particularly high influence exerted by the teacher is expected. However, further influences (e.g., teaching experience, treatment adherence, methods and materials used, and effective reading time) need to be investigated

in the context of reading comprehension and could be explored through qualitative observations in a mixed-method approach along with maze score. Based on Kim et al.'s (2015) findings, we decided not to include the variable school level into the models because lower effects were expected. This variable should be included in future studies.

Conclusions

In summary, our study highlights the usefulness of applying the CBM maze as a screening tool in secondary education. Furthermore, our results show the importance of considering classroom influences when planning and implementing reading instruction. In Germany, individual learning support is only granted if a student has an officially diagnosed SEN status. But preventive reading support in classroom is necessary for all students to improve their learning progress and to balance classroom effects. For instance, preventive support could be provided at classroom level through the implementation of a multi-tiered system of appropriate measures. This would ensure that students receive appropriate support regardless of whether SEN was diagnosed or not, also in higher grades and in inclusive classrooms.

Acknowledgements

The authors would like to thank all participants, teachers and schools for taking part in this research. Open Access funding enabled and organized by Projekt DEAL.

Conflict of Interest Statement

The authors declare no conflicts of interest.

Data Availability Statement

Data and code for R, Version 1.3.1093, are available in the project 'Variance in Reading Comprehension' on OSF.io (https://osf.io/fgh4d/).

References

Amendum, S. J., Conradi Smith, K., & Liebfreund, M. D. (2021). Explaining reading variance by student subgroup: Should we move beyond oral reading fluency? *Journal of Research in Reading*, *44*(4), 757–786. https://doi.org/10.1111/1467-9817.12371

Anderson, S., Jungjohann, J., & Gebhardt, M. (2020). Effects of using curriculum-based measurement (CBM) for progress monitoring in reading and an additive reading instruction in second classes. *Zeitschrift für Grundschulforschung*, *51*(1), 1–166. https://doi.org/10.1007/s42278-019-00072-5

Bates, D., Maechler, M., Bolker, B., & Walker, S. (2020). *Linear mixed-effects models using 'Eigen' and S4*. https://github.com/lme4/lme4/

Blumenthal, S., Blumenthal, Y., Lembke, E. S., Powell, S. R., Schultze-Petzold, P., & Thomas, E. R. (2021). Educator perspectives on data-based decision making in Germany and the United States. *Journal of Learning Disabilities*, *54*(4), 284–299. https://doi.org/10.1177/0022219420986120

Brown-Chidsey, R., Davis, L., & Maya, C. (2003). Sources of variance in curriculum-based measures of silent reading. *Psychology in the Schools*, *40*(4), 363–377. https://doi.org/10.1002/pits.10095

Capin, P., Roberts, G., Clemens, N. H., & Vaughn, S. (2021). When treatment adherence matters: Interactions among treatment adherence, instructional quality, and student characteristics on reading outcomes. *Reading Research Quarterly*, *57*(2), 753–774. https://doi.org/10.1002/rrq.442

Chung, S., Espin, C. A., & Stevenson, C. E. (2018). CBM maze-scores as indicators of reading level and growth for seventh-grade students. *Reading and Writing*, *31*(3), 627–648. https://doi.org/10.1007/s11145-017-9803-8

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences*. L. Erlbaum Associates.

Curran, M., & Hogan, T. P. (2023). Developmental language disorder: What it is and why it matters. In S. Q. Cabell, S. B. Neuman, & N. P. Terry (Eds.), *Handbook on the science of early literacy* (pp. 325–335). The Guilford Press.

Deno, S. L. (2003). Curriculum-based measures: Development and perspectives. *Sage Journal*, *28*(3–4), 3–12. https://doi.org/10.1177/073724770302800302

Ebenbeck, N., Rieser, J., Jungjohann, J., & Gebhardt, M. (2022). How the existence of special schools affects the placement of students with special needs in inclusive primary schools. *Journal of Research in Special Educational Needs*, *22*(3), 274–287. https://doi.org/10.1111/1471-3802.12565

Finch, W. H., Bolin, J. E., & Kelley, K. (2014). *Multilevel modeling using R*. Chapman and Hall/CRC.

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

Gebhardt, M., Sälzer, C., Mang, J., Müller, K., & Prenzel, M. (2015). Performance of students with special educational needs in Germany. Findings from Programme for International Student Assessment 2012. *Journal of Cognitive Education and Psychology*, *14*(3), 343–356. https://doi.org/10.1891/1945-8959.14.3.343

Görgen, R., de Simone, E., Schulte-Körne, G., & Moll, K. (2021). Predictors of reading and spelling skills in German: The role of morphological awareness. *Journal of Research in Reading*, *44*(1), 210–227. https://doi.org/10.1111/1467-9817.12343

Hessels, M. G. P. (2009). Estimation of the predictive validity of the hart by means of a dynamic test of geography. *Journal of Cognitive Education and Psychology*, *8*(1), 5–21. https://doi.org/10.1891/1945-8959.8.1.5

Hox, J. J., Moerbeek, M., & van de Schoot, R. (2018). *Multilevel analysis: Techniques and applications* (Third ed.). Quantitative methodology series. Routledge, Taylor & Francis Group.

January, S.-A. A., & Ardoin, S. P. (2012). The impact of context and word type on students' maze task accuracy. *School Psychology Review*, *41*(3), 262–271. https://doi.org/10.1080/02796015.2012.12087508

Jensen, K. L., & Elbro, C. (2022). Clozing in on reading comprehension: A deep cloze test of global inference making. *Reading and Writing*, *35*(5), 1221–1237. https://doi.org/10.1007/s11145-021-10230-w

Jungjohann, J. (2022). Komplexe Nebensätze, Kohärenz- oder Inferenzbildung: Unterschiede im satzübergreifenden Leseverständnis von Jugendlichen mit sonderpädagogischem Unterstützungsbedarf im Bereich Sprache. [Complex subordinate clauses, coherence or inference making: Differences in sentence-related reading comprehension among adolescents with developmental language disorders]. *Forschung Sprache*, *10*(2), 19–33. https://doi.org/10.5283/epub.53198

Jungjohann, J., DeVries, J. M., Gebhardt, M., & Mühling, A. (2018). Levumi: A web-based curriculum-based measurement to monitor learning progress in inclusive classrooms. In K. Miesenberger & G. Kouroupetroglou (Hrsg.) (Eds.), *Computers helping people with special needs. ICCHP 2018. Lecture notes in computer science* (pp. 369–378). Springer International Publishing. https://doi.org/10.1007/978-3-319-94277-3_58

Jungjohann, J., & Gebhardt, M. (2021). *SinnL-Levumi N6 - Tests zum sinnkonstruierenden Lesen als Lernverlaufsdiagnostik – "Sinnkonstruierendes Satzlesen" der Onlineplattform www.levumi.de [SinnL-Levumi N6 - Tests for sentence comprehension as learning progress monitoring]*. Universität Regensburg. https://doi.org/10.5283/epub.47877

Jungjohann, J., & Gebhardt, M. (2023). Dimensions of classroom-based assessments in inclusive education. *International Journal of Special Education (IJSE)*, *38*(1), 131–144. https://doi.org/10.52291/ijse.2023.38.12

Jungjohann, J., DeVries, J. M., Mühling, A., & Gebhardt, M. (2018). Using theory-based test construction to develop a new curriculum-based measurement for sentence reading comprehension. *Frontiers in Education*, *3*(1), 115. https://doi.org/10.3389/feduc.2018.00115

Kalambouka, A., Farrell, P., Dyson, A., & Kaplan, I. (2007). The impact of placing pupils with special educational needs in mainstream schools on the achievement of their peers. *Educational Research*, *49*(4), 365–382. https://doi.org/10.1080/00131880701717222

Kieffer, M. J., & Christodoulou, J. A. (2020). Automaticity and control: How do executive functions and reading fluency interact in predicting reading comprehension? *Reading Research Quarterly*, *55*(1), 147–166. https://doi.org/10.1002/rrq.289

Kim, Y.-S., Petscher, Y., & Foorman, B. (2015). The unique relation of silent reading fluency to end-of-year reading comprehension: Understanding individual differences at the student, classroom, school, and district levels. *Reading and Writing*, *28*(1), 131–150. https://doi.org/10.1007/s11145-013-9455-2

Kreft, I., & de Leeuw, J. (2011). *Introducing multilevel modeling*. Sage.

Kultusministerkonferenz. (2020). *Sonderpädagogische Förderung in Schulen 2009 bis 2018 [Special education support in schools 2009 to 2018]*. https://www.kmk.org/dokumentation-statistik/statistik/schulstatistik/sonderpaedagogische-foerderung-an-schulen.html

Lepper, C., Stang, J., & McElvany, N. (2021). Gender differences in text-based interest: Text characteristics as underlying variables. *Reading Research Quarterly*, *57*(2), 537–554. https://doi.org/10.1002/rrq.420

Lindsay, G. (2007). Educational psychology and the effectiveness of inclusive education/mainstreaming. *The British Journal of Educational Psychology*, *77*(1), 1–24. https://doi.org/10.1348/000709906X156881

Logan, S., & Johnston, R. (2009). Gender differences in reading ability and attitudes: Examining where these differences lie. *Journal of Research in Reading*, *32*(2), 199–214. https://doi.org/10.1111/j.1467-9817.2008.01389.x

MacKay, E., Lynch, E., Sorenson Duncan, T., & Deacon, S. (2021). Informing the science of reading: Students' awareness of sentence-level information is important for reading comprehension. *Reading Research Quarterly*, *56*(S1). https://doi.org/10.1002/rrq.397

Montague, M., & Rinaldi, C. (2001). Classroom dynamics and children at risk: A followup. *Learning Disability Quarterly*, *24*(2), 75–83. https://doi.org/10.2307/1511063

Mühling, A., Jungjohann, J., & Gebhardt, M. (2019). Progress monitoring in primary education using Levumi: A case study. In H. Lane, S. Zvacek, & J. Uhomoihi (Hrsg.) (Eds.), *Proceedings of the 11th International Conference on Computer Supported Education* (pp. 137–144. Science and Technology Publications. https://doi.org/10.5220/0007658301370144

Muijselaar, M. M. L., Kendeou, P., de Jong, P. F., & van den Broek, P. W. (2017). What does the CBM-maze test measure? *Scientific Studies of Reading*, *21*(2), 120–132. https://doi.org/10.1080/10888438.2016.1263994

Myklebust, J. O. (2002). Inclusion or exclusion? Transitions among special needs students in upper secondary education in Norway. *European Journal of Special Needs Education*, *17*(3), 251–263. https://doi.org/10.1080/08856250210162158

National Assessment of Educational Progress. (2022). *The nation's report card: Reading 2022*. https://www.nationsreportcard.gov/reading/nation/groups/?grade=4

OECD. (2019a). *PISA 2018 results (volume I): What students know and can do*. OECD Publishing and Centre for Educational Research and Innovation. https://doi.org/10.1787/5f07c754-en

OECD. (2019b). Programme for International Student Assessment (PISA). Results from PISA 2018. Country Note Germany. https://www.oecd.org/pisa/publications/PISA2018_CN_DEU.pdf

Pfost, M., Hattie, J., Dörfler, T., & Artelt, C. (2014). Individual differences in reading development. *Review of Educational Research*, *84*(2), 203–244. https://doi.org/10.3102/0034654313509492

R Core Team. (2020). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing. https://www.R-project.org/

Schurig, M., Jungjohann, J., & Gebhardt, M. (2021). Minimization of a short computer-based test in reading. *Frontiers in Education*, *6*, 684595. https://doi.org/10.3389/feduc.2021.684595

Shin, J., & McMaster, K. (2019). Relations between CBM (oral reading and maze) and reading comprehension on state achievement tests: A meta-analysis. *Journal of School Psychology*, *73*, 131–149. https://doi.org/10.1016/j.jsp.2019.03.005

Spreer, M., Glück, C. W., & Theisel, A. (2019). Sprachliche Fähigkeiten und Schulleistungen von Grundschulkindern mit sonderpädagogischem Förderbedarf Sprache im Längsschnitt [Language abilities and academic achievement of primary-school pupils with special educational needs for language in a longitudinal perspective]. *Empirische Sonderpädagogik*, *11*(4), 318–338.

Stranghöner, D., Wild, E., & Schwinger, M. (2021). Identifying predictors of performance in reading and writing in primary students with mild learning difficulties. *European Journal of Special Needs Education*, *1–15*, 804–818. https://doi.org/10.1080/08856257.2021.1954345

Tzivinikou, S., Tsolis, A., Kagkara, D., & Theodosiou, S. (2020). Curriculum based measurement maze: A review. *Psychology*, *11*(10), 1592–1611. https://doi.org/10.4236/psych.2020.1110101

van Breukelen, G. J. P. (2005). Psychometric modeling of response speed and accuracy with mixed and conditional regression. *Psychometrika*, *70*(2), 359–376. https://doi.org/10.1007/s11336-003-1078-0

Will, A.-K. (2019). The German statistical category "migration background": Historical roots, revisions and shortcomings. *Ethnicities*, *19*(3), 535–557. https://doi.org/10.1177/1468796819833437

Dr Jana Jungjohann is postdoctoral researcher and has the Chair of Pedagogy for Learning Disabilities Including Inclusive Pedagogy at the University of Regensburg, Germany. Areas of interest are development of educational assessment and progress monitoring in reading and spelling, intervention materials for inclusive school practice, teacher training and special education.

Dr Michael Schurig is postdoctoral researcher at Faculty of Rehabilitation Science, TU Dortmund University, Germany. Areas of interest are school development research, quantitative research methodology and special education.

Dr Markus Gebhardt is Professor and has the Chair of Pedagogy for Learning Disabilities Including Inclusive Pedagogy at the University of Regensburg, Germany. Areas of interest are inclusive education, special education, educational assessment and teacher education. In this line, he developed an online platform for progress monitoring (www.levumi.de) and constructed the free tests according to Open Science standards.

Received 24 September 2021; revised version received 24 July 2023.

Address for correspondence: Jana Jungjohann, Chair of Pedagogy for Learning Disabilities Including Inclusive Pedagogy, Faculty of Human Science, University of Regensburg, Sedanstrasse 1, Regensburg, Bavaria D-93055, Germany. Email: jana.jungjohann@ur.de