frontiers in Education

ORIGINAL RESEARCH
published: 21 December 2018
doi: 10.3389/feduc.2018.00115

Using Theory-Based Test Construction to Develop a New Curriculum-Based Measurement for Sentence Reading Comprehension

Jana Jungjohann[1]*, *Jeffrey M. DeVries*[1], *Andreas Mühling*[2] and *Markus Gebhardt*[1]

[1] Faculty of Rehabilitation Science, Research in Inclusive Education, Technische Universität Dortmund, Dortmund, Germany,
[2] Department of Computer Science, Computer Science Education Research, Kiel University, Kiel, Germany

Reading comprehension at sentence level is a core component in the students' comprehension development, but there is a lack of comprehension assessments at the sentence level, which respect the theory of reading comprehension. In this article, a new web-based sentence-comprehension assessment for German primary school students is developed and evaluated using a curriculum-based measurement (CBM) framework. The test focuses on sentence level reading comprehension as an intermediary between word and text comprehension. The construction builds upon the theory of reading comprehension using CBM-Maze techniques. It is consistent on all tasks and contains different syntactic and semantic structures within the items. This paper presents the test development, a description of the item performance, an analysis of its factor structure, and tests of measurement invariance, and group comparisons (i.e., across gender, immigration background, over two measurement points, and the presence of special educational needs; SEN). Third grade students ($n = 761$) with and without SEN finished two CBM tests over 3 weeks. Results reveal that items had good technical adequacy, the constructed test is unidimensional, and it is valid for students both with and without SEN. Similarly, it is valid for both sexes, and results are valid across both measurement points. We discuss our method for creating a unidimensional test based on multiple item difficulties and make recommendations for future test construction.

Keywords: curriculum-based measurement, reading assessment, web-based assessment, sentence reading comprehension, special educational needs, primary school age, progress monitoring

OPEN ACCESS

Edited by:
Mustafa Asil,
University of Otago, New Zealand

Reviewed by:
Lisa Zimmerman,
University of South Africa, South Africa
Haci Bayram Yilmaz,
Ondokuz Mayıs University, Turkey

***Correspondence:**
Jana Jungjohann
jana.jungjohann@tu-dortmund.de

Specialty section:
This article was submitted to Assessment, Testing and Applied Measurement, a section of the journal Frontiers in Education

Received: *29 August 2018*
Accepted: *10 December 2018*
Published: *21 December 2018*

Citation:
Jungjohann J, DeVries JM, Mühling A and Gebhardt M (2018) Using Theory-Based Test Construction to Develop a New Curriculum-Based Measurement for Sentence Reading Comprehension. Front. Educ. 3:115. doi: 10.3389/feduc.2018.00115

INTRODUCTION

Reading acquisition, particularly reading comprehension, is one of the most important academic skills (Nash and Snowling, 2006); however, multiple student groups [i.e., students with special educational needs (SEN), or students with immigration background] have disadvantages to reach a high reading comprehension competency level. In particular, reading comprehension is a challenge for students with SEN resulting in problems in both primary and secondary school (Gersten et al., 2001; Cain et al., 2004; Berkeley et al., 2011; Cortiella and Horowitz, 2014; Spencer et al., 2014). Many of students with SEN have language difficulties that accompany a high risk for difficulties in literacy. In the United States and the United Kingdom, the vast majority of students with SEN struggles in this area (Kavale and Reese, 1992; NAEP, 2015; Lindsay and Strand, 2016). Similar

results were reported for learners speaking in phonologically consistent languages such as German (Gebhardt et al., 2015). Additionally, in most western countries, including Germany, students with an immigration background (i.e., they were born in another country, have at least one parent were born in another country, or are a non-citizen but born in the target country; Salentin, 2014) have greater academic challenges than native students (OECD, 2016). Consequently, they also tent to have a lower reading competency (Schnepf, 2007; OECD, 2014; Lenkeit et al., 2018) which partially related to a discrepancy in language skills for the host country language (Kristen et al., 2011). Recently, Lindsay and Strand's (2016) large scale and longitudinal study emphasized the importance of identifying children with reading problems and their individual needs as early as possible. Alarmingly, students with reading comprehension problems during primary school demonstrate reading difficulties well into adolescence (Landerl and Wimmer, 2008; Taylor, 2012). Additionally, it is widely agreed that there is a large need for adequate assessments to achieve positive outcomes for children with SEN (OECD, 2005; Hao and Johnson, 2013; Lindsay, 2016).

One possible assessment type is curriculum-based measurement (CBM), which involves short, time-limited, and frequent tests to visualize the learning progress of low achieving students (e.g., students with SEN; Deno, 1985, 2003). For instance, the CBM-Maze task is a reading comprehension task where students receive a complete text with multiple blanks. They fill in each blank with one word from a few choices (Fuchs et al., 1992). CBM-Maze was designed to monitor the growth of intermediate and secondary students' reading comprehension. More recent studies showed that CBM-Maze measures early language skills, such as sentence level comprehension and code-related skills rather than higher language skills, such as inference-making, text comprehension, and knowledge about text structures (Wayman et al., 2007; Graney et al., 2010; Muijselaar et al., 2017). Because CBM-Maze assesses earlier reading skills, it may be adapted for younger students, including low achieving students. In this paper, we develop a new assessment for reading comprehension at the sentence level for primary school students, in the lines of reading comprehension theory and CBM framework. Our goal is to create an assessment that considers different structures of language and is highly suitable for both researchers and practitioners.

Reading Comprehension

Reading comprehension is "the ultimate goal of reading" (Nation, 2011, p. 248); it is necessary for everyone to be successful in school and society. In general, reading comprehension is a system of cognitive skills and processes (Kendeou et al., 2014). Multiple underlying skills, such as rapid naming, phonological, and orthographic processing, fluency, vocabulary, and working memory, need to interlock to allow for good comprehension performance (e.g., Cain et al., 2004; Cain and Oakhill, 2007; Kendeou et al., 2012; Keenan and Meenan, 2014). According to the simple view of reading (Gough and Tunmer, 1986; Hoover and Gough, 1990), reading comprehension is defined as a product of code-decoding skills and linguistic comprehension. Furthermore, it is divided into three related levels: word, sentence, and text. Comprehension at the word level is the lowest-tier ability. Readers visually identify a word as a single unit and compare it with their mental representation (Coltheart et al., 2001). It includes subskills such as phonological awareness, decoding, and written word recognition. The sentence represents the middle tier. When readers connect several elements of a sentence, such as words, and phrases, they frame a local representation at this level. Sentence comprehension builds a fundamental bridge between the lower (word) and upper (text) levels using syntactic parsing and semantic integration (Frazier, 1987; Ecalle et al., 2013). At the top tier is text comprehension. In order to understand connected texts, reading learners need to establish additional complex cognitive processes, such as inference-making, coherence-making, and background knowledge (Van Dijk and Kintsch, 1983). In contrast to the simple view of reading, the hierarchical construction-integration model of text comprehension (originally by Kintsch, 1998), is divided into two mental representations: textbase and situation (Kintsch and Rawson, 2011). In textbase representations, readers combine the low level reading skills on word and sentence levels (i.e., microstructure) to build a local, coherent representation of the macrostructure of a text. The situation model relates to text content (i.e., integration of further information and knowledge) and is comparable with the text level comprehension. Both the textbase and the situation model of a text are fundamentally connected (Perfetti, 1985). Without understanding single words or sentence meaning, no reader is able to make inferences, or coherences. Consequently, in both the simple view of reading and the construction-integration model, sentence comprehension is a critical skill for advanced text comprehension.

Processes of Sentence Comprehension

General syntactic and semantic comprehension processes influence the comprehension at the word and sentence level. Additionally, word recognition is linked with the syntactic and semantic analysis (Oakhill et al., 2003; Frisson et al., 2005). All these processes can be understood as parallel, modular, or dominated by one process (Taraban and McClelland, 1990; McRae et al., 1998; Kennison, 2009). Even though research results are not consistent, relevant factors about sentence comprehension are known. For instance, word classes guide syntactic parsing. Kennison (2009) found that verb information affected syntactical parsing in undergraduate and graduate students. In contrast, Cain et al. (2005) studied primary school students' ability to choose correct conjunctions in tasks similar to CBM-Maze tasks (specifically Cloze tasks). Results showed that filling in additive and adversative conjunctions is easier than temporal and causal ones for 8–10 years old. Thus, syntactical information influences the extraction of the sentence meaning, and semantic information influences the comprehension of individual words. More precisely, their results show that the part of speech causes different difficulty levels in the children's sentence comprehension. Therefore, reading comprehension assessments need to include these possible difficulties in the test structure to be able to measure these core skills. In line with earlier claims, these results confirm that reading comprehension assessments should reveal whether students are struggling in

lower- or upper-level subskills, and where the difficulties lie (Catts et al., 2003).

Reading Difficulties and Sentence Comprehension

Comprehension problems are very heterogeneous. They can be caused by lexical processes (e.g., phonological and semantic skills or visual word recognition), by the capacity of the working memory, or by higher text processes (see Nation, 2011). Students might develop word recognition difficulties, isolated comprehension difficulties, or combined problems in both areas (Leach et al., 2003; Nation, 2011; Catts et al., 2012; Kendeou et al., 2014). Some studies suggest that younger readers' problems can be mostly ascribed to poor word recognition skills (Vellutino et al., 2007; Tilstra et al., 2009). Accordingly, less than one percent of early primary students who perform well in decoding, and vocabulary show isolated comprehension problems (Spencer et al., 2014). However, poor readers also differ from good readers in the efficiency of related cognitive processes (Perfetti, 2007). At the sentence level, poor comprehenders use sentence content for word recognition, which is especially challenging in complex semantic and syntactic structures. West and Stanovich (1978) showed that sentence content supports word recognition processes of fourth graders more than the process of advanced adult readers. Consequently, all readers can develop greater difficulties if the sentence content is not congruent to the word meaning. Martohardjono et al. (2005) investigated the relevance of the syntax of single sentences in reading comprehension on bilingual students. Results revealed that the participants could not use the syntactic structure to support their word recognition. Relatedly, poor readers in third grade struggle to extract the meaning of syntactically complex sentences when they were presented verbally (Waltzman and Cairns, 2000). Besides the semantic and syntactic deficits, poor readers take longer to read complex sentences (Graesser et al., 1980; Chung-Fat-Yim et al., 2017). The more complex the sentence, the more time is needed to process the syntax and semantics of the sentence. Thus, poor readers need more cognitive resources to read and understand single sentences.

Furthermore, sentence level comprehension tests can measure general reading comprehension. Ecalle et al. (2013) examined the role of sentence processing as a mediatory skill within reading comprehension in second through ninth graders. First, they presented the students semantically similar sentences with different complexity and vocabulary. The students had to judge whether the contents were similar or not. Second, they examined the impact of these skills on expository text comprehension. The results confirmed that sentence processing increases over age and that sentence comprehension "could constitute a good indicator of a more general level of reading comprehension irrespective of the type of text" (Ecalle et al., 2013, p. 128).

Overall, these results show that even small changes (i.e., semantical and syntactic ones) within a sentence can influence students' comprehension. Especially important is that the comprehension ability of poor readers (i.e., students with SEN or with immigration background) is sensitive to both word meaning and sentence structure. Additionally, these results affirm the need for specific reading comprehension assessments at sentence level in contrast to word recognition or fluency tests (Cutting and Scarborough, 2006).

Curriculum-Based Measurements

Curriculum-based measurement (CBM) is a problem-solving approach for assessing the learning growth of low achieving students (e.g., students with SEN) in basic academic skills, such as reading, writing, spelling, or mathematic competencies (Deno, 1985, 2003). The main idea of CBM is to monitor the children's development with short and very frequent tests during regular lessons. This allows teachers to graphically view the slope of the individual learning growth and to evaluate the effectiveness of the instruction. Teachers can then link the results with their decision-making and lesson planning.

Curriculum-Based Measurements as an Assessment of Reading Comprehension

A lot of research relating to CBM has been conducted (Fuchs, 2017), especially in reading tasks. In primary schools, two kinds of CBM instruments are ordinarily used to measure reading competencies: CBM-R and CBM-Maze (Graney et al., 2010). CBM-R involves individual students reading aloud from a word list or a connected text, and it measures oral reading fluency and accuracy (i.e., word recognition). CBM-Maze was developed to compensate for the disadvantages of CBM-R, namely individual administration and teacher distrust of CBM-R as a reading comprehension measurement. CBM-Maze is a group administered silent reading task that measures general text reading comprehension. However, both tests provide valid measurements of students' reading comprehension skills (Ardoin et al., 2004; Marcotte and Hintze, 2009). For both types of CBM, many studies report technical adequacy, strong alternate-form reliability, moderate to strong criterion-related validity, and predictive validity (e.g., Shin et al., 2000; Graney et al., 2010; Espin et al., 2012; Ardoin et al., 2013). Furthermore, correlations between CBM-R and CBM-Maze were found (Wayman et al., 2007). CBM-R is typically used for measurements within the first three grades and CBM-Maze for fourth and higher graders.

Principles of CBM-Maze

In traditional CBM-Maze tasks, students read a timed short passage (∼250 words), in which different words are deleted. For each blank, one target word and two or more distractors are presented. The students choose one word for each blank. In general, the first and the last sentence of each passage are kept intact to allow the context to guide comprehension. The number of correctly filled blanks is then used as the competence measure for reading comprehension. The CBM-Maze task is based on the Cloze (Louthan, 1965; Gellert and Elbro, 2013) test, where the students fill out the blanks without any time limit or word suggestions. Since then, many studies examined CBM-Maze test construction and administration issues.

Test administration

Fuchs and Fuchs (1992) were among the first researchers to describe the CBM-Maze task. In contrast to the CBM-R, they

highlighted the higher classroom usability because the CBM-Maze can be administered in groups and at computers. Recently, Nelson et al. (2017) found that computer adaptive tests—as a practice of CBM—are valid for progress monitoring with fourth and fifth graders. They reported that with frequent data collection, computer testing systems can examine the overall learning growth of individuals and student groups. One main feature of CBM tests is that they can be administered multiple times, which requires both alternative test forms and sensitivity to student growth (Fuchs, 2004).

In their meta-analysis, García and Cain (2014) determined general reading comprehension assessment characteristics. They observed significant differences for the linguistic material and the administration procedure (i.e., reading aloud or silently, and test time). However, this was not upheld for CBM-Maze tasks because research has not found significant differences for primary students between reading silently or aloud (Hale et al., 2011). Accordingly, the CBM-Maze assessments are mostly administered silently for higher practicability in group settings. The CBM-Maze test time is usually short, from 1 to 10 min (e.g., Fuchs and Fuchs, 1992; Brown-Chidsey et al., 2003; Wiley and Deno, 2005). While no specific limit has been agreed upon, Brown-Chidsey et al. (2003) suggested that a test time of 10 min is too long for a 250-word passage.

Item construction

Traditionally, the items of CBM-Maze tasks are connected passages. Depending on the age of the students, different kinds of passages are used, such as fables (Förster and Souvignier, 2011), newspaper articles (Tichá et al., 2009), and historical texts (Brown-Chidsey et al., 2003). Outside the CBM approach, maze tasks are commonly used to measure semantic and syntactic skills within sentence processing (Forster, 2010). In these cases, single sentences are mostly used instead of connected passages (e.g., Witzel and Witzel, 2016). January and Ardoin (2012) examined differences in the construction of the CBM-Maze probes with third, fourth, and fifth graders. They examined whether the students' performance is contingent on the content of the passages. The findings indicated that primary school students performed well on both intact (i.e., sentences in order) and scrambled (i.e., sentences out of order) CBM-Maze passages. Additionally, they concluded that the CBM-Maze task measures reading comprehension at the sentence level because the students did not need the context to perform well. Taken together with the results of Ecalle et al. (2013), these results suggest that CBM-Maze could also be administered with single sentences instead of connected passages.

Furthermore, item construction depends on the deletion pattern and on the linguistic material. Some test designers use a fixed (i.e., every seventh word) or a lexical (e.g., deletion of nouns, verbs, adjectives, or conjunctions) deletion pattern, however, Kingston and Weaver (1970) showed that the lexical deletion pattern had similar results as fixed deletion. Similarly, January and Ardoin (2012) could not find significant differences in the students' accuracy based upon different lexical deletion patterns.

While not explicitly tested, results that indicated similar accuracy for different types of lexical deletion suggest unidimensionality. This is important for teachers and researchers because items that fall on the same dimension are easier to interpret (Gustafsson and Åberg-Bengtsson, 2010). This does not mean that the underlying construct of reading comprehension is unidimensional, but that the results are interpretable along a single dimension of reading comprehension (Reise et al., 2013). Furthermore, individual test performance differs based upon factors such as SEN and immigration background (Cortiella and Horowitz, 2014; Spencer et al., 2014; OECD, 2016). Thus, in order for a test to be fair for all test takers, the linguistic material should be similar in construction and the items need to be appropriate for diverse groups of students, such as learners with SEN, those with an immigration background, and learners of both genders (i.e., measurement invariance; Good and Jefferson, 1998; Steinmetz, 2013).

Distractors

Other studies discussed the influence of distractors on a correct answer. Early studies indicated two types of distractors: semantically plausible with incorrect syntactic structure or semantically meaningless with correct syntactic structure (Guthrie et al., 1974; Gillingham and Garner, 1992). Resulting suggestions for distractors include a similar look as the target word, a lack of contextual sense, words with a related, incompatible contextual meaning, or nonsense words (Fuchs and Fuchs, 1992). McKenna and Miller (1980) found that syntactically correct distractors are more difficult for students to exclude in comparison to similar looking words. Meanwhile Conoyer et al. (2017) found that tests using content-based and part of the speech-based distractors were similar.

The Present Study

Because reading comprehension at the sentence level is a necessary skill (Ecalle et al., 2013), we developed a new web-based test to measure this competence. Our new assessment focuses on sentence reading ability within a CBM framework for primary school students (i.e., third graders). Our study details test development and analyzes item difficulty. It also assesses dimensionality with an analysis of the factorial structure and tests measurement invariance across several relevant groups. Additionally, we track the performance of our participants across two measurement points and examine the effect of subject variables including SEN, immigration background, and gender. Accordingly, we developed two groups of research questions. The first group of questions relates to technical evaluation of the test and the second group relates to the overall performance of our participants.

The first three questions relate to the technical aspects of test construction and interpretation including item difficulty, unidimensionality, and measurement invariance:

1. *What are the item difficulties and do they relate to different item types*? We use multiple word types to create different difficulty levels and we expect that some word types will be more difficult than others (see Cain et al., 2005; Frisson et al., 2005; Kennison, 2009).

2. *Can the instrument results be interpreted unidimensionally?* A unidimensional test structure would allow for easy test interpretation for both researchers and educators, because they only need to consider overall performance on the test. We use a consistent and straightforward sentence structure with age-related words to create a test structure that is applicable to both good and poor comprehenders (e.g., students with SEN). Thus, we hypothesize that all items fit on a unidimensional test structure because the item structure is consistent and all items represent the same underlying reading competence.
3. *Does the test possess measurement invariance relating to SEN, immigration background, gender, and measurement points?* Test construction followed guidelines for CBM test construction and evaluation (Fuchs, 2004; Wilbert and Linnemann, 2011). This includes multiple alternate test forms of equal difficulty, integration of several subskills (e.g., word recognition, syntactic parsing, and semantic integration). Because we adopt these established recommendations and combine them with CBM-Maze praxis (e.g., Brown-Chidsey et al., 2003; January and Ardoin, 2012; Conoyer et al., 2017), we hypothesize that our test will be invariant over student groups and measurement points.

The last two questions focus on the performance of the participants in relation to classroom and individual factors:

4. *What is the intraclass correlation?* Although the test was given to different classrooms, we expect the results to be comparable in each classroom. Therefore, we expect a relatively low intraclass correlation, meaning that the test performed similarly across all classrooms. For comparison, Hedges and Hedberg (2007) use the guideline of 0.05–0.15 in their large-scale assessment.
5. *Did performance improve over time, and did subject variables, including SEN, immigration background, and gender influence performance?* We compare the sum scores of our participants across two measurement points. We expect that there will be an improvement in performance from measurement point one to two and learners with SEN will perform worse (see Gebhardt et al., 2015; Lindsay and Strand, 2016).

METHODS

Test Administration

The new reading assessment is administered online via a German web-based platform for CBM monitoring, called Levumi (www.levumi.de). The code for the Levumi platform is published on Github, and all tests, test materials, and teacher handbooks will be published with a creative commons license. This means that this test is free of charge for teachers and researchers (Jungjohann et al., 2018). The platform runs on all major browsers and only requires an internet connection. It records each student's response and reaction time for every item.

The test can be administered in groups or individually. Each student has his or her individual student account where the activated tests become available. Teachers or researchers activate the test for each measurement point for the participants (e.g., students) in the test-taker's individual account. A computer or tablet is required for each simultaneous test-taker.

At the beginning of each test, a simple interactive example is shown. This prevents an accidental test start or other interface problems. After the example, students have 8 min to answer as many items as possible. On the screen, the items appear one after another. The students see a single sentence with a blank. Underneath the sentences, the target word and the distractors are displayed in a random order. When a student chooses one of the possible words, it appears in the blank. The students can change their minds by clicking on another response, and afterwards they confirm their answer by clicking on the "next" button. When the time limit runs out, the students can finish the current item, and then the test closes automatically.

At each measurement point, the item order is different, allowing for alternate test forms for frequent measurements over a school year. The items have a fixed item order for the first measurement (see **Table 1**), with items alternating between each different word-category (as described below). This fixed item order creates a baseline for comparison for the first measurement point. Random orders are generated for the second measurement point on to allow for a large number of alternative test forms. In these orders, a category is randomly selected, and then an item from each category, but the ratio of items from each category is kept equal.

Item and Distractor Construction

The overall process of item creation followed the CBM-Maze's principles; however, some modifications were made according to reading comprehension theory in order to create a test using sentences rather than connected passages. First, all items were carefully created as individual sentences. The entire pool of 60 items can be seen in **Table 1**. To ensure that every sentence is appropriate for third graders, all important words were collated from curricula within the German primary school systems (e.g., lists of frequent words based on grade level). Every sentence is a sentoid, meaning it is semantically explicit including the distractors. All sentences are constructed in the active voice and with age-appropriate syntactic structures (i.e., avoiding sentences with multiple clauses).

Second, a lexical deletion pattern was chosen to set different item difficulties within one test. Research results showed that the use of single word types can affect a different sentence comprehension. To adopt these results for the test construction, all items were classified by the lexical deletion pattern. The essential hypothesis is that difficulty is determined by the type of word deleted (i.e., part of speech). Because all items were set with a similar sentence structure, they relate to the same competence (i.e., reading comprehension at sentence level). To build up the variation of the German language, word types were summarized in multiple categories. This new assessment considers only possibilities relevant for third graders and not all possibilities within the German language. Therefore, three categories were set. The first word-category included nouns ($n = 20$) used as both subjects and objects. The second category included verbs and adjectives ($n = 21$). The third category included conjunctions and prepositions ($n = 19$).

TABLE 1 | Overview of the items*.

Item number	Item category	Item order	Item	Item difficulty (%)
1	1	1	A face has two eyes/*fingers/books/cars*.	86
2	1	17	My Mum sleeps in a bed/*picture/cage/table*.	94
3	1	38	Lasse draws nice pictures/*bits/air/cold*.	88
4	1	28	A lama has four legs/*thumbs/camels/books*.	88
5	1	7	My friend moves into a new house/*shirt/exercise book/flowerbed*.	84
6	1	35	I tie my shoes/*keys/chest/nature*.	84
7	1	58	Dad locks the front door/*bottle/bridge/blackboard* in the evening.	88
8	1	42	Jutta goes shopping with her sister/*shower/cottage/pain*.	87
9	1	60	My Dad works in a(n) office/*July/ruler/fun*.	90
10	1	51	The buoy floats in the water/*bed/lions/cloth*.	90
11	1	24	The bird/*dog/club/father* flies to the nest.	91
12	1	48	The ducks/*bears/houses/hair* quack in the lake.	87
13	1	4	The flowers/*bids/boys/chairs* bloom in the field.	87
14	1	22	The friends/*pens/shoes/lights* are up to no good.	82
15	1	57	The rabbit/*sand/skirt/tooth* runs across the field.	91
16	1	6	The scissors/*pizzas/onions/foreheads* cut paper.	92
17	1	14	The baby/*radio/packet/puzzle* cries for its mother.	91
18	1	32	The sun/*meadow/clock/doll* shines every day.	92
19	1	41	The frogs/*mushrooms/fruits/teeth* jump across the street.	89
20	1	44	The bee/*tree bark/nose/flower* sits on a blossom.	77
21	2	40	The lemonade is sweet/*quiet/wealthy/sandy*.	89
22	2	21	An apple is a round/*long/blue/warm* fruit.	91
23	2	54	A needle is sharp/*guilty/lovely/loud*.	88
24	2	13	The police arrests the bad/*square/flat/warm* robbers.	91
25	2	23	With my glasses, I am able to see well/*fresh/loud/young.*	92
26	2	52	My Dad buys a new/*round/cold/high* car.	91
27	2	2	My sister always studies hard/*greenly/flatly/thinly*.	89
28	2	43	The fat/*high/round/square* man shouts loudly.	81
29	2	26	Lisa tells me a funny/*late/wet/deep* joke.	91
30	2	30	At dinnertime, I am often hungry/*close/big/far*.	90
31	2	11	The fast/*full/close/fresh* car races down the street.	87
32	2	36	Your friend bakes/*builds/asks/studies* a large loaf of bread.	83
33	2	47	The girl eats/*runs/ drives/rotates* the soup.	84
34	2	16	Lukas talks/*opens/packs/bakes* with Frida about the holidays.	70
35	2	19	The sun shines/*rains/melts/shouts* often in the summer.	82
36	2	50	You wait/*meet/love/remove* out the thunderstorm.	57
37	2	34	My grandma sleeps/*hits/rinses/glances* late on Saturdays.	82
38	2	56	Paula buys/*fries/chews/races* a present for me.	83
39	2	5	The cake tastes/*sniffs/drinks/chooses* very good to us.	80
40	2	29	In the forest, we collect/*calculate/close/place* leaves.	89
41	2	10	I live/*want/know/let* with my family.	85
42	3	31	Paul has neither/*that/although/because* of a pen nor a notebook.	61
43	3	12	I have a good grade, but/*that/because/or* Sina unfortunately doesn't.	82
44	3	39	If/*Before/Instead/Than* water freezes, it turns into ice.	78
45	3	45	While/*So/Except/Neither* my Mum paints the fence, I play in the garden.	74
46	3	53	Jutta doesn't like to eat fruit, except for/*from which/so/because of* cherries.	70
47	3	20	After/*As if/But rather/Neither* we have moved, Mum and Dad buy new furniture.	67
48	3	9	I brush my teeth, before/*after/soon/from* I go to bed.	74
49	3	55	The gull is able to fly, because/*before/except/from* it is a bird.	72
50	3	25	Instead of/*Because/Right after/From* looking, I would rather ask my Mom.	62

(Continued)

TABLE 1 | Continued

Item number	Item category	Item order	Item	Item difficulty (%)
51	3	3	As soon as/*But/And/Therefore*, I'm ready, I'll let you know.	73
52	3	37	You climb up/*to/from/amid* the tree.	50
53	3	49	Lots of people sit in/*through/on/out* the plaza.	81
54	3	59	The car pulls the other car only with/*between/for/to* difficulty.	66
55	3	27	Through/*Over/Inside/To* the telescope, I see a tower.	66
56	3	46	The present is for/*on /under /next to* my sister.	77
57	3	15	My notebook is in between/*from/inside/out* of the pillows.	73
58	3	8	We can't visit during/*behind/over/for* the bad weather	78
59	3	33	We play under/*through/inside/from* the table.	85
60	3	18	This pen is from/*under/over/in* my grandpa.	82
	1		Category 1 (Nouns)-average difficulty	88
	2		Category 2 (Adjective/verbs)–average difficulty	85
	3		Category 3 (Conjunction/prepositions)–average difficulty	72
			All Categories–Average Difficulty	82

Italicized words denote distractors.

Third, three distractors were created for every target word. Every distractor was contextually meaningless but syntactically possible. Three rules guided the distractor creation: one distractor had a similar look, another distractor was related to the contextual sense of the target word or sentence but lacked the correct meaning, and the last distractor made no contextual sense. Because of vocabulary limitations, these rules could not be implemented precisely in each item. In these cases, words from the other rules were adapted. In every case, the same number of words were presented. The following example illustrates these rules:

(Item 4—German) Ein Lama hat vier Beine/*Bücher/Daumen/Kamele*.
(Item 4—English) A llama has four legs/*books/thumbs/camels*.

Participants

Participants were third grade students attending regular elementary schools in the northwest of Germany ($n = 761$). Approximately half of the participants were female (46.5%). The participants' teachers were asked about the immigration background ($n = 344$) and SEN. SEN were listed as learning ($n = 37$), cognitive development ($n = 2$), and other (i.e., speech and language impairments, emotional disturbed, or functional disability; $n = 67$). Participants with low German language competence ($n = 40$) were also categorized as SEN.

Procedure

Trained research assistants (i.e., university students) contacted local elementary school administrators and teachers to recruit participants. All data was collected with the informed consent of participants, parents, teachers, and administrators. The research assistants tested participants in small groups. Each participant worked individually on a single school computer. After the first measurement, researchers returned 3 weeks later to collect the data for the second measurement. Some students were not available due to illness or other reasons at the second measurement ($n = 94$). In this case, their data for the second measurement was treated as missing, but their data from the first measurement was used where appropriate.

During both measurements, research assistants followed the same scripted procedure. Children were told that the little dragon Levumi has brought many sentences with it, but that each of the sentences were missing a word. The child was asked if he or she could find the correct word in each sentence. Next, the participants were given an example item. Once they gave the correct answer for the example item, the research assistant showed the participant how to give this answer on the test with the mouse. After the example item, participants answered as many items as they could in an 8 min period.

Analysis

Item Difficulties

Average item difficulty was obtained by averaging the percentage of correct responses across both measurement points. Sufficient numbers of participants completed all test items ($n = 87$, 11.4%) to allow for a power-analysis, and a further 25% of participants completed the vast majority of the test (50 items). Therefore, missing scores were treated as not yet reached in our analyses. A repeated measures one-way ANOVA compared the average difficulties based on word-category (noun, adjective/verb, and conjunction/preposition).

Dimensionality

To assess dimensionality of the instrument, the factor structure was tested via three separate confirmatory factor analyses (CFAs). All factor analyses were conducted in Mplus 7.4 (Muthén and Muthén, 1998–2015) using a weighted least squares with mean and variance adjusted (WLSMV) estimation method. The WLSMV estimator is more appropriate for categorical data

(Muthén et al., 1997; Flora and Curran, 2004). Factor structures were based on the word type category and function in the sentence. The 3-factor model mirrored the three word-categories: nouns, verbs/adjectives, and conjunctions/prepositions. In the 2-factor model, the verbs/adjectives, and conjunctions/prepositions factors were combined. Finally, in the 1-factor structure, all items were placed on the same factor. We appraised the model fits via root mean squared error of approximation (RMSEA), CFI, and gamma-hat. We considered RMSEA < 0.08, CFI > 0.90, and gamma-hat > 0.90 acceptable fits. Meanwhile, we considered RMSEA < 0.05, CFI > 0.95, and gamma-hat > 0.95 good fits (Hu and Bentler, 1998).

Next, we compared the fits of the separate models in order of increasing complexity. We compared the 2-factor model to the 1-factor model and the 3-factor model to the 2-factor model. We examined changes in CFI (ΔCFI) and gamma-hat (ΔGamma-hat). We set a threshold of 0.01 for ΔCFI and ΔGamma-hat as a significantly better fit (Cheung and Rensvold, 2002; Dimitrov, 2010).

Measurement Invariance

We examined the measurement invariance of each of the three models based on presence of SEN, immigration background, gender, and measurement point. We constrained or freed thresholds and lambda together. In other words, we tested the scalar (strong) model directly against the configural (base) model, as recommended when using WLSMV analysis by Muthén and Muthén (1998–2015). As described above, differences in ΔCFI and ΔGamma-hat greater than 0.01 were considered significant.

Intraclass Correlation

Next, the intraclass correlation was calculated using the proportion of variance explained by classroom compared to overall variance. Values were calculated based on the sum of squares in a one way ANOVA of class on average percent correct at the first measurement point.

Change Over Time and the Influence of Subject Variables

Finally, to assess the differential performance on the test by our target groups, we conducted a repeated measures ANOVA including SEN, immigration background, and gender across both measurement points on the sum score of each participant.

RESULTS

Item Difficulties

Table 1 lists all items and item difficulties across both measurement points. ANOVA results confirmed that difficulty varied across word-categories, $F_{(2,57)} = 25.215$, $p < 0.001$. Tukey's honestly significant difference test revealed that items in the easier two categories (noun and adjective/verb) were similar in difficulty, $p > 0.05$, meanwhile items in the conjunction/preposition group were significantly harder than the other two groups, $p < 0.05$.

TABLE 2 | Model Fits.

	RMSEA (90% CI)	CFI	Gamma-hat	ΔCFI	ΔGamma-hat
1-Factor	0.013 (0.009–0.016)	0.988	0.991	–	–
2-Factor	0.012 (0.007–0.015)	0.991	0.992	0.003	0.001
3-Factor	0.011 (0.006–0.015)	0.991	0.993	0.000	0.001

ΔCFI and ΔGamma-hat represent the change in model fit from the less complex to the more complex model.

Dimensionality

Fit metrics for all three models surpassed our criteria for good fits, as described in **Table 2**. Fit metrics were only slightly worse in the 2-factor model than in the 1-factor model, and virtually identical between the 2-factor and 3-factor model. None of the model comparisons exceeded the critical value of ΔCFI or ΔGamma-hat of 0.01. Therefore, we conclude that all models fit equally well, and on the grounds of parsimony, we prefer the simpler 1-factor model. Thus, the instrument can be considered unidimensional.

Measurement Invariance

Measurement invariance test results are shown in **Table 3**. In each case, ΔCFI and ΔGamma-hat are below the threshold of 0.01, meaning that the scalar model fit similar to the metric model. Therefore, we conclude that all three models possessed strong measurement invariance across presence of SEN, gender, immigration background, and measurement point. Because invariance was upheld for all models, the simpler 1-factor model is still preferable to other models. Thus, a unidimensional interpretation is equally valid for all subgroups within our data.

Intraclass Correlation

The intraclass correlation coefficient, as measured by proportion of total variance, indicated the test functioned similarly across all classrooms in our data, ICC = 0.15. This is relatively high, but still in the guidelines used in previous work (see Hedges and Hedberg, 2007).

Change Over Time and the Influence of Subject Variables

Table 4 shows the results of the sum score analysis. The repeated measures ANOVA revealed that students performed better on the test at measurement point 2, $F_{(1,658)} = 93.32$, $p < 0.001$. Additionally, learners with SEN performed worse overall than those without, $F_{(1,658)} = 89.01$, $p < 0.001$. Furthermore, a significant interaction indicated that learners with SEN did not improve from measurement point 1 to measurement point 2, $F_{(1,658)} = 5.45$, $p < 0.05$. No other interactions or main effects were found, all $ps > 0.05$.

DISCUSSION

Overview of Findings and Theoretical Implication

This study developed and evaluated a new theory-based formative assessment that measures reading comprehension at sentence level and that follows the CBM approach for

TABLE 3 | Measurement Invariance for 1-, 2-, and 3-factors.

Grouping	Model	RMSEA	CFI	Gamma-hat	ΔCFI	ΔGamma-hat
Special education needs	**1-Factor**					
	Configural	0.012	0.982	0.984	–	–
	Scalar	0.012	0.982	0.984	0.000	0.000
	2-Factor					
	Configural	0.012	0.984	0.984	–	–
	Scalar	0.012	0.984	0.984	0.000	0.000
	3-Factor					
	Configural	0.011	0.985	0.987	–	–
	Scalar	0.011	0.986	0.987	−0.001	0.000
Immigration background	**1-Factor**					
	Configural	0.012	0.989	0.984	–	–
	Scalar	0.012	0.988	0.984	0.001	0.000
	2-Factor					
	Configural	0.011	0.990	0.987	–	–
	Scalar	0.011	0.990	0.987	0.000	0.000
	3-Factor					
	Configural	0.011	0.991	0.987	–	–
	Scalar	0.011	0.990	0.987	0.001	0.000
Gender	**1-Factor**					
	Configural	0.013	0.981	0.981	–	–
	Scalar	0.012	0.987	0.984	−0.006	−0.003
	2-Factor					
	Configural	0.012	0.988	0.984	–	–
	Scalar	0.011	0.989	0.987	−0.001	−0.003
	3-Factor					
	Configural	0.012	0.989	0.985	–	–
	Scalar	0.011	0.989	0.987	0.000	−0.002
Measurement point	**1-Factor**					
	Configural	0.010	0.990	0.989	–	–
	Scalar	0.011	0.989	0.986	0.001	0.003
	2-Factor					
	Configural	0.009	0.992	0.910	–	–
	Scalar	0.010	0.991	0.989	0.001	0.001
	3-Factor					
	Configural	0.009	0.992	0.991	–	–
	Scalar	0.010	0.991	0.989	0.001	0.002

ΔCFI and ΔGamma-hat denote the difference between the configural and scalar models.

practical use in inclusive primary schools. The main goal was to create a unidimensional test structure with different item difficulties to allow for easy interpretation and a high usability for heterogeneous classrooms. Within our theory-based test construction, we linked common reading comprehension models at the sentence level with the principles of the CBM-Maze task. In addition, guidelines were set to assure that all finalized items from the same word-category were equivalent in construction and difficulty.

In general, the evaluation of the test construction revealed a 1-factor model with items of varying difficulty. Our results indicated significant differences between the three deletion pattern categories (e.g., word-categories) of the single items. In this study, the German third graders had fewer problems identifying correct target words for nouns (e.g., category 1), verbs, or adverbs (e.g., category 2) compared to conjunctions, or prepositions (e.g., category 3). These results are in line with previous results from Kennison (2009) and Cain et al. (2005), which indicated that different word types affect the syntactical parsing in different ways. Our results showed that these previous results could be generalized to the German language. Additionally, the different item difficulty between the word-categories can assist teachers to precisely screen problems in reading development. Förster and Souvignier (2011) argued that precise identification is an important feature of CBM assessments. Furthermore, the CFA demonstrated a unidimensional test structure, which allows for a simple interpretation from educators. Overall, our theory-based test construction demonstrated both adherence to reading comprehension theory and technical adequacy, making it useful

TABLE 4 | Comparison of sum scores.

	Measurement point 1		Measurement point 2		Significance
	M	(*SD*)	*M*	(*SD*)	*p*-value
Special educational need					0.001***
Yes	21.4	(13.1)	21.2	(11.9)	–
No	34.4	(13.2)	39.9	(13.4)	–
Immigration background					0.083
Yes	30.0	(13.8)	34.2	(14.6)	–
No	34.0	(14.0)	38.8	(14.8)	–
Gender					0.171
Male	31.8	(14.2)	36.3	(15.4)	–
Female	32.8	(13.8)	37.3	(14.3)	–
All groups	32.2	(14.0)	36.8	(14.9)	0.001***

The significance column denotes the between subjects results of the ANOVA for the special education needs, immigration background, and gender rows, but for the all groups row, it denotes the within subjects variable of measurement point.
****Significant at $p < 0.001$.*

to both teachers and researchers. Additionally, these results indicated that it is possible to set item difficulty by language related rules without creating several test types (i.e., letter, sentence, and text-based tests).

The assessment's practicability for inclusive classrooms was verified by the results of the measurement invariance tests, the intraclass correlation, and the analysis of the sum scores in three key ways. First, the intraclass correlation showed that our test performed similarly within different inclusive classroom settings. Additionally, the sum score comparison showed that changes in the students' ability were detectable without any changes in the test administration for a specific student group. This indicates that no specific class or student characteristics are required for test administration. Second, the alternative test forms were invariant across both measurement points. Meaning, that the random drawing order created good multiple alternative test forms, which can ease the test handling for both researchers and teachers. Within big classrooms, teachers do not have to track which student completed which test form or remember test dates manually, meaning that Levumi can reduce teachers' workload in multiple ways. Similarly, researchers can create a large number of alternative test forms. Thus, our web-based Levumi platform demonstrates one of the key benefits of computer-based formative assessment (see Russell, 2010). Third, especially in inclusive classrooms, the students are heterogeneous in their academic performance (Gebhardt et al., 2015). While students with SEN performed significantly lower, our Levumi reading comprehension test was invariant for different student groups (i.e., SEN, gender, and immigration background). Meaning, that teachers can use the Levumi test for all these students because the test assesses competence fairly across these groups. Furthermore, students with and without SEN can use the same test system over multiple measurement points in inclusive classrooms. Again, this reduces the teachers' workload because teachers do not have to use other materials for special student groups within one classroom (Jungjohann et al., 2018).

Our Levumi reading comprehension test includes the test administration benefits of the CBM-Maze, such as group administration, and silent reading (see Graney et al., 2010). Additionally, it is suitable for early readers and for readers with low reading abilities (e.g., students with SEN). In particular, our test uses a sentence-based item pool, rather than a complete text-based item pool. Fuchs et al. (2004) and Good et al. (2001) argued that complete text-based tests can cause floor effects on low performance readers. Because of this, the Levumi test may be more suitable for even younger students and those who might have difficulties with complete texts. Correspondingly, these test characteristics demonstrate that it is possible to expand the established CBM assessment types in new ways. Additionally, we expand the existing techniques of evaluating CBM assessments with intraclass correlations, factor analyses, and measurement invariance tests. These three evaluation techniques are well established on other fields of test evaluation, and their use can help to rigorously evaluate existing and future CBM assessments. This demonstrated technique of theory-based test construction and evaluation provides an essential template for other researchers who may be developing a diverse range of CBM assessments.

Limitations and Future Work

Nonetheless, this study has some limitations. The findings still need to be replicated with a more varied participant pool and with a larger sample from other regions inside and outside of Germany. This actual study also focused on third graders, but reading difficulties can appear earlier in first and second grade, when students start to develop an understanding of written words and sentences (Richter et al., 2013). Therefore, further studies should also include first and second graders to expand adequate CBM assessment for this age.

Besides a broader participant pool, future longitudinal research is necessary to validate our hypotheses. One main CBM characteristic is the ability to track the students' learning growth across multiple measurement points over a long period (e.g., one school year; Deno, 2003). In our study, we confirmed that our test is invariant over two measurement points for a period of 3 weeks. For classroom use, it is necessary to analyze the test's ability to measure the students learning slope over a larger period with more than two measurement points per student.

Additionally, we have not established a concrete indicator of criterion validity yet (see Fuchs, 2004). Besides the Levumi reading comprehension test, participants could complete additional CBM assessments with a complete text-based item pool, and other established reading comprehension screenings. This would establish if the Levumi reading comprehension test relies more on to code-related skills (e.g., reading fluency) than on language related skills (e.g., reading comprehension), as suggested by Muijselaar et al. (2017). This can indicate which reading problems our test is effective at identifying. This is particularly important because established reading

comprehension tests do not agree with each other in the identification of reading problems (Keenan and Meenan, 2014).

Our last key limitation relates to the item language. Our research is limited by only using the German items. Both, the English item translation and the general theory-based test construction need to be evaluated in additional languages. At first, studies should evaluate the translated items with native speakers to test the quality of the items. Results of these studies would confirm the usability of our theory-based guidelines for CBM test construction in other languages. The original items can be translated into additional languages based upon these studies. This procedure will expand CBM offerings into new languages and regions.

Further work should also focus on instructional utility. This study followed Fuchs's (2004) recommendations for CBM test construction and examined the technical adequacy for formative learning growth monitoring (e.g., stage 1 and 2; Fuchs, 2004). Instructional utility means that teachers can include the CBM test system in their actual lessons, that they can understand, and interpret the results, and that they can link the students' learning slopes with their reading instructions. In Germany, the CBM approach is still unknown by many teachers. Therefore, the three main aspects of the instructional utility (Fuchs, 2004) are also concerns for the Levumi reading comprehension test and should be investigated in further research. First, the acceptance and the application of the Levumi platform needs to be evaluated within the school context. For the practical use of the Levumi platform, teachers need access to a computer, or a tablet and an internet connection. Some German schools already have good technical equipment, while others do not. Even assuming access to good technical equipment, teachers must be willing to use a web- and computer-based assessment. To that end, a clear user interface and teacher-focused supporting material will encourage user adoption. Second, further studies should examine how the teachers can handle the Levumi test results. Recently, studies revealed that preservice and in-service teachers can have problems in understanding CBM graphs (Van den Bosch et al., 2017; Zeuch et al., 2017). As one example, pre-service teachers estimate fictitious future student achievements lower than can be expected by a linear regression model (Klapproth, 2018). In all these studies, the participants were not able to adjust the layout, visualize additional information from the tests (e.g., correct, and incorrect answers), or receive statistical help (e.g., trend line, or goal line). Consequently, future studies need to test these interpretation difficulties in order to adapt the specific Levumi output (i.e., CBM graphs, and further information). Third, especially, the web-based test system brings a high potential for automatized support in CBM graph interpretation and instruction making. For instance, the Levumi platform could highlight at risk students, automatize, or add additional information into the graph, such as statistical trend lines, instruction phases, or students' moods. Furthermore, the Levumi platform could learn typical problem patterns and suggest relevant instruction materials. Lastly, work needs to be done to identify the connection between specific aspects of reading competency and performance on individual items and overall results. All these possibilities should be implemented carefully so that teachers are not confused by CBM data.

CONCLUSION

We created a new CBM assessment using theory-based test construction and evaluation. Evaluations indicated the test to be of high use for further research and praxis. This provides three key implications. First, researchers can adapt our approach as guidelines for further CBM assessments (i.e., further language, learning domains) to enhance the CBM research and evaluation field in new ways. Second, the web-based Levumi platform and the reading comprehension assessment are suitable for inclusive classrooms and their use can reduce the teachers' workload in multiple ways. And third, our procedure demonstrates the development of a test with multiple item difficulties which can be interpreted along a single dimension.

ETHICS STATEMENT

Permission for this study was granted through dean of the Faculty of Rehabilitation Science, Technical University of Dortmund. Following the requirements of the ministry of education of the federal state North Rhine-Westphalia (Schulgesetz für das Land Nordrhein-Westfalen, 2018), school administrators decided in co-ordination with their teachers about participation in this scientific study. An additional ethics approval was not required for this study as per Institution's guidelines and national regulations. Parents obtained written information about the study and any potential benefits. They gave their written consent for each child. Participation was supervised by school staff. Participation was voluntary and participants were free to withdraw at any time.

AUTHOR CONTRIBUTIONS

JJ developed the Levumi reading comprehension assessment and coordinated the study. As the primary author, JJ did most of the writing, research, and some of the analyses. JD did most of the data analyses. AM programmed the Levumi platform, realized the Levumi reading comprehension assessment based on the test specification of JJ, and edited the manuscript. MG gave the initial study design, supervised the entire research process, and edited the manuscript.

ACKNOWLEDGMENTS

We acknowledge financial support by Deutsche Forschungsgemeinschaft and Technische Universität Dortmund/ TU Dortmund Technical University within the funding programme Open Access Publishing.

REFERENCES

Ardoin, S. P., Christ, T. J., Morena, L. S., Cormier, D. C., and Klingbeil, D. A. (2013). A systematic review and summarization of the recommendations and research surrounding Curriculum-Based Measurement of oral reading fluency (CBM-R) decision rules. *J. Sch. Psychol.* 51, 1–18. doi: 10.1016/j.jsp.2012.09.004

Ardoin, S. P., Witt, J. C., Suldo, S. M., Connell, J. E., Koenig, J. L., Resetar, J. L., et al. (2004). Examining the incremental benefits of administering a maze and three versus one curriculum-based measurement reading probes when conducting universal screening. *School Psych. Rev.* 33, 218–233.

Berkeley, S., Mastropieri, M. A., and Scruggs, T. E. (2011). Reading comprehension strategy instruction and attribution retraining for secondary students with learning and other mild disabilities. *J. Learn. Disabil.* 44, 18–32. doi: 10.1177/0022219410371677

Brown-Chidsey, R., Davis, L., and Maya, C. (2003). Sources of variance in curriculum-based measures of silent reading. *Psychol. Sch.* 40, 363–377. doi: 10.1002/pits.10095

Cain, K., and Oakhill, J. (2007). "Reading comprehension difficulties: Correlates, causes, and consequences," in *Children's Comprehension Problems in Oral and Written Language*, eds K. Cain and J. Oakhill (London: Guilford Press), 41–75.

Cain, K., Oakhill, J., and Bryant, P. (2004). Children's reading comprehension ability. Concurrent prediction by working memory, verbal ability, and component skills. *J. Educ. Psychol.* 96, 31–42. doi: 10.1037/0022-0663.96.1.31

Cain, K., Patson, N., and Andrews, L. (2005). Age- and ability-related differences in young readers' use of conjunctions. *J. Child. Lang.* 32, 877–892. doi: 10.1017/S0305000905007014

Catts, H. W., Compton, D., Tomblin, J. B., and Bridges, M. S. (2012). Prevalence and nature of late-emerging poor readers. *J. Educ. Psychol.* 104, 166–181. doi: 10.1037/a0025323

Catts, H. W., Hogan, T. P., and Fey, M. E. (2003). Subgrouping poor readers on the basis of individual differences in reading-related abilities. *J. Learn. Disabil.* 36, 151–164. doi: 10.1177/002221940303600208

Cheung, G. W., and Rensvold, R. B. (2002). Evaluating goodness-of-fit indexes for testing measurement invariance. *Struct. Equation Model.* 9, 233–255. doi: 10.1207/S15328007SEM0902_5

Chung-Fat-Yim, A., Peterson, J. B., and Mar, R. A. (2017). Validating self-paced sentence-by-sentence reading. Story comprehension, recall, and narrative transportation. *Read. Writ.* 30, 857–869. doi: 10.1007/s11145-016-9704-2

Coltheart, M., Rastle, K., Perry, C., Langdon, R., and Ziegler, J. (2001). DRC. A dual route cascaded model of visual word recognition and reading aloud. *Psychol. Rev.* 108, 204–256. doi: 10.1037/0033-295X.108.1.204

Conoyer, S. J., Lembke, E. S., Hosp, J. L., Espin, C. A., Hosp, M. K., and Poch, A. L. (2017). Getting more from your maze. Examining differences in distractors. *Read. Writ. Q.* 33, 141–154. doi: 10.1080/10573569.2016.1142913

Cortiella, C., and Horowitz, S. H. (2014). *The State of Learning Disabilities: Facts, Trends and Emerging Issues*. New York, NY: National Center for Learning Disabilities.

Cutting, L. E., and Scarborough, H. S. (2006). Prediction of reading comprehension. relative contributions of word recognition, language proficiency, and other cognitive skills can depend on how comprehension is measured. *Sci. Stud. Reading* 10, 277–299. doi: 10.1207/s1532799xssr1003_5

Deno, S. L. (1985). Curriculum-based measurement: the emerging alternative. *Exc. Child.* 52, 219–232. doi: 10.1177/001440298505200303

Deno, S. L. (2003). Developments in curriculum-based measurement. *J. Spec. Educ.* 37, 184–192. doi: 10.1177/00224669030370030801

Dimitrov, D. M. (2010). Testing for factorial invariance in the context of construct validation. *Meas. Eval. Counsel. Dev.* 43, 121–149. doi: 10.1177/0748175610373459

Ecalle, J., Bouchafa, H., Potocki, A., and Magnan, A. (2013). Comprehension of written sentences as a core component of children's reading comprehension. *J. Res. Read.* 36, 117–131. doi: 10.1111/j.1467-9817.2011.01491.x

Espin, C., Rose, S., McMaster, K., and Wayman, M. (2012). *A Measure of Success: The Influence of Curriculum Based Measurement on Education.* Minneapolis, MN: University of Minnesota Press.

Flora, D. B., and Curran, P. J. (2004). An empirical evaluation of alternative methods of estimation for confirmatory factor analysis with ordinal data. *Psychol. Methods* 9, 466–491. doi: 10.1037/1082-989X.9.4.466

Forster, K. (2010). Using a maze task to track lexical and sentence processing. *Mental Lexicon* 5, 347–357. doi: 10.1075/ml.5.3.05for

Förster, N., and Souvignier, E. (2011). Curriculum-based measurement. developing a computer-based assessment instrument for monitoring student reading progress on multiple indicators. *Learn. Disabil.* 9, 65–88.

Frazier, L. (1987). "Sentence processing: A tutorial review," in *Attention and Performance*, ed M. Coltheart (Hove: Erlbaum), 559–586.

Frisson, S., Rayner, K., and Pickering, M. J. (2005). Effects of contextual predictability and transitional probability on eye movements during reading. *J. Exp. Psychol. Learn. Memory Cogn.* 31, 862–877. doi: 10.1037/0278-7393.31.5.862

Fuchs, L. S. (2004). The past, present, and future of curriculum-based measurement research. *School Psychol. Rev.* 2, 188–192.

Fuchs, L. S. (2017). curriculum-based measurement as the emerging alternative. Three decades later. *Learn. Disabil. Res. Pract.* 32, 5–7. doi: 10.1111/ldrp.12127

Fuchs, L. S., and Fuchs, D. (1992). Identifying a measure for monitoring student reading progress. *School Psych. Rev.* 21, 45–58.

Fuchs, L. S., Fuchs, D., and Compton, D. L. (2004). Monitoring early reading development in first grade. Word identification fluency versus nonsense word fluency. *Exc. Child.* 71, 7–21. doi: 10.1177/001440290407100101

Fuchs, L. S., Fuchs, D., Hamlett, C. L., and Ferguson, C. (1992). Effects of expert system consultation within curriculumbased measurement using a reading maze task. *Except. Child.* 58, 436– 450. doi: 10.1177/001440299205800507

García, J. R., and Cain, K. (2014). Decoding and reading comprehension. *Rev. Educ. Res.* 84, 74–111. doi: 10.3102/0034654313499616

Gebhardt, M., Sälzer, C., Mang, J., Müller, K., and Prenzel, M. (2015). Performance of students with special educational needs in germany. findings from programme for international student assessment 2012. *J. Cogn. Educ. Psychol.* 14, 343–356. doi: 10.1891/1945-8959.14.3.343

Gellert, A. S., and Elbro, C. (2013). Do experimental measures of word learning predict vocabulary development over time? A study of children from grade 3 to 4. *Learn. Individ. Differ.* 26, 1–8. doi: 10.1016/j.lindif.2013.04.006

Gersten, R., Fuchs, L. S., Williams, J. P., and Baker, S. (2001). Teaching reading comprehension strategies to students with learning disabilities. A review of research. *Rev. Educ. Res.* 71, 279–320. doi: 10.3102/00346543071002279

Gillingham, M. G., and Garner, R. (1992). Readers' comprehension of mazes embedded in expository texts. *J. Educ. Res.* 85, 234–241. doi: 10.1080/00220671.1992.9941121

Good, R. H., and Jefferson, G. (1998). "Contemporary perspectives on curriculum-based measurement validity," in *Advanced Applications of Curriculum-Based Measurement*, ed M. R. Shinn (New York, NY: Guilford Press), 61–88.

Good, R. H., Simmons, D. C., and Kame'enui, E. J. (2001). The importance and decision-making utility of a continuum of fluency-based indicators of foundational reading skills for third-grade high-stakes outcomes. *Sci. Stud. Read.* 5, 257–288. doi: 10.1207/S1532799XSSR0503_4

Gough, P. B., and Tunmer, W. E. (1986). Decoding, reading, and reading disability. *Remed. Spec. Educ.* 7, 6–10. doi: 10.1177/074193258600700104

Graesser, A. C., Hoffman, N. L., and Clark, L. F. (1980). Structural components of reading time. *J. Verbal Lear. Verbal Behav.* 19, 135–151. doi: 10.1016/S0022-5371(80)90132-2

Graney, S. B., Martínez, R. S., Missall, K. N., and Aricak, O. T. (2010). Universal screening of reading in late elementary school. *Remed. Spec. Educ.* 31, 368–377. doi: 10.1177/0741932509338371

Gustafsson, J.-E., and Åberg-Bengtsson, L. (2010). "Unidimensionality and interpretability of psychological instruments," in *Measuring Psychological Constructs: Advances in Model-Based Approaches*, ed S. E. Embretson (Washington, DC: American Psychological Association), 97–121.

Guthrie, J. T., Seifert, M., Burnham, N. A., and Caplan, R. I. (1974). The maze technique to assess, monitor reading comprehension. *Read. Teach.* 28, 161–168.

Hale, A. D., Hawkins, R. O., Sheeley, W., Reynolds, J. R., Jenkins, S., Schmitt, A. J., et al. (2011). An investigation of silent versus aloud reading comprehension of elementary students using Maze assessment procedures. *Psychol. Schools* 48, 4–13. doi: 10.1002/pits.20543

Hao, S., and Johnson, R. L. (2013). Teachers' classroom assessment practices and fourth-graders' reading literacy achievements. An international study. *Teach. Teach. Educ.* 29, 53–63. doi: 10.1016/j.tate.2012.08.010

Hedges, L. V., and Hedberg, E. C. (2007). Intraclass correlation values for planning group-randomized trials in education. *Educ. Eval. Policy Anal.* 29, 60–87. doi: 10.3102/0162373707299706

Hoover, W. A., and Gough, P. B. (1990). The simple view of reading. *Read. Writ.* 2, 127–160. doi: 10.1007/BF00401799

Hu, L.-T., and Bentler, P. M. (1998). Fit indices in covariance structure modeling. Sensitivity to underparameterized model misspecification. *Psychol. Methods* 3, 424–453. doi: 10.1037/1082-989X.3.4.424

January, S.-A. A., and Ardoin, S. P. (2012). The impact of context and word type on students' maze task accuracy. *School Psych. Rev.* 41, 262–271.

Jungjohann, J., DeVries, J. M., Gebhardt, M., and Mühling, A. (2018). "Levumi: A Web-Based Curriculum-Based Measurement to Monitor Learning Progress in Inclusive Classrooms," in *Computers Helping People with Special Needs*. 16th International Conference, eds K. Miesenberger, G. Kouroupetroglou, and P. Penaz (Wiesbaden: Springer), 369–378.

Kavale, K. A., and Reese, J. H. (1992). The character of learning disabilities. An Iowa profile. *Learn. Disabil. Q.* 15, 74–94. doi: 10.2307/1511010

Keenan, J. M., and Meenan, C. E. (2014). Test differences in diagnosing reading comprehension deficits. *J. Learn. Disabil.* 47, 125–135. doi: 10.1177/0022219412439326

Kendeou, P., Papadopoulos, T. C., and Spanoudis, G. (2012). Processing demands of reading comprehension tests in young readers. *Learn. Instr.* 22, 354–367. doi: 10.1016/j.learninstruc.2012.02.001

Kendeou, P., van den Broek, P., Helder, A., and Karlsson, J. (2014). A cognitive view of reading comprehension. Implications for reading difficulties. *Learn. Disabil. Res. Prac.* 29, 10–16. doi: 10.1111/ldrp.12025

Kennison, S. M. (2009). The use of verb information in parsing. Different statistical analyses lead to contradictory conclusions. *J. Psychol. Res.* 38, 363–378. doi: 10.1007/s10936-008-9096-9

Kingston, A. J., and Weaver, W. W. (1970). Feasibility of Cloze techniques for teaching and evaluating culturally disadvantaged beginning readers. *J. Soc. Psychol.* 82, 205–214. doi: 10.1080/00224545.1970.9919952

Kintsch, W. (1998). *Comprehension: A Paradigm for Cognition*. New York, NY: Cambridge University Press.

Kintsch, W., and Rawson, K. A. (2011). "Comprehension," in *The Science of Reading: A Handbook*, ed M. Snowling and C. Hulme (Malden, MA: Blackwell Publ.), 209–226.

Klapproth, F. (2018). Biased predictions of students' future achievement. An experimental study on pre-service teachers' interpretation of curriculum-based measurement graphs. *Stud. Educ. Eval.* 59, 67–75. doi: 10.1016/j.stueduc.2018.03.004

Kristen, C., Edele, A., Kalter, F., Kogan, I., Schulz, B., Stanat, P., et al. (2011). The education of migrants and their children across the life course. *Zeitschrift Für Erziehungswissenschaft* 14, 121–137. doi: 10.1007/s11618-011-0194-3

Landerl, K., and Wimmer, H. (2008). Development of word reading fluency and spelling in a consistent orthography. An 8-year follow-up. *J. Educ. Psychol.* 100, 150–161. doi: 10.1037/0022-0663.100.1.150

Leach, J. M., Scarborough, H. S., and Rescorla, L. (2003). Late-emerging reading disabilities. *J. Educ. Psychol.* 95, 211–224. doi: 10.1037/0022-0663.95.2.211

Lenkeit, J., Schwippert, K., and Knigge, M. (2018). Configurations of multiple disparities in reading performance: longitudinal observations across France, Germany, Sweden and the United Kingdom. *Assess. Educ.* 25, 52–86. doi: 10.1080/0969594X.2017.1309352

Lindsay, G. (2016). Grand challenge: priorities for research in special educational needs. *Front. Educ.* 1:1. doi: 10.3389/feduc.2016.00001

Lindsay, G., and Strand, S. (2016). Children with language impairment. prevalence, associated difficulties, and ethnic disproportionality in an english population. *Front. Educ.* 1:2. doi: 10.3389/feduc.2016.00002

Louthan, V. (1965). Some systematic grammatical deletions and their effects on reading comprehension. *English J.* 54, 295–299. doi: 10.2307/811113

Marcotte, A. M., and Hintze, J. M. (2009). Incremental and predictive utility of formative assessment methods of reading comprehension. *J. Sch. Psychol.* 47, 315–335. doi: 10.1016/j.jsp.2009.04.003

Martohardjono, G., Otheguy, R., Gabriele, A., DeGoeas-Malone, M., Rivero, S., Pyrzanowski, S. et al. (2005). "The role of syntax in reading comprehension: a study of bilingual readers," in *Proceedings of the 4th International Symposium on Bilingualism* (Somerville, MA: Cascadilla Press), 1522–1544. Available online at: http://www.lingref.com/isb/4/119ISB4.PDF

McKenna, M. C., and Miller, J. W. (1980). "The Effects of Age and Distractors Type on Maze Performance," in *Perspectives on Reading Research and Instruction: 29th Yearbook of the National Reading Conference*, ed K. L. Kamil (Washington, DC: National Reading Conference), 288–292.

McRae, K., Spivey-Knowlton, M. J., and Tanenhaus, M. K. (1998). Modeling thematic fit (and other constraints) within an integration competition framework. *J. Mem. Lang.* 38, 283–312. doi: 10.1006/jmla.1997.2543

Muijselaar, M. M. L., Kendeou, P., Jong, P. F., and van den Broek, P. W. (2017). What does the CBM-maze test measure? *Sci. Stud. Read.* 21, 120–132. doi: 10.1080/10888438.2016.1263994

Muthén, B. O., Du Toit, S., and Spisic, D. (1997). *Robust Inference using Weighted Least Squares and Quadratic Estimating Equations in Latent Variable Modeling with Categorical and Continuous Outcomes.* Available online at: https://www.statmodel.com/download/Article_075.pdf

Muthén, L. K., and Muthén, B. O. (1998–2015). *Mplus User's Guide, 7th Edn.* Los Angeles: CA: Muthén & Muthén.

NAEP (2015). *National Assessment of Educational Progress (NAEP; 2015). NAEP Reading Report Card*. Available online at: https://www.nationsreportcard.gov/reading_2017/#?grade=4

Nash, H., and Snowling, M. (2006). Teaching new words to children with poor existing vocabulary knowledge. A controlled evaluation of the definition and context methods. *Int. J. Lang. Commun. Disord.* 41, 335–354. doi: 10.1080/13682820600602295

Nation, K. (2011). "Childrens's reading comprehension difficulties," in *The Science of Reading: A Handbook*, eds M. Snowling and C. Hulme (Malden, MA: Blackwell Publ.), 248–265.

Nelson, P. M., Van Norman, E. R., Klingbeil, D. A., and Parker, D. C. (2017). Progress monitoring with computer adaptive assessments. the impact of data collection schedule on growth estimates. *Psychol. Schools* 54, 463–471. doi: 10.1002/pits.22015

Oakhill, J. V., Cain, K., and Bryant, P. E. (2003). The dissociation of word reading and text comprehension. Evidence from component skills. *Lang. Cogn. Proces.* 18, 443–468. doi: 10.1080/01690960344000008

OECD (2005). *Formative Assessment:Improving Learning in Secondary Classrooms*. Paris: OECD Publishing.

OECD (2014). *PISA 2012 Results: What Students Know and Can Do – Student Performance in Mathematics, Reading and Science,* Vol. 1,. Paris: OECD Publishing.

OECD (2016). *PISA 2015 Results in Focus, PISA in Focus, No. 67*. Paris: OECD Publishing.

Perfetti, C. A. (1985). *Reading Ability*. New York, NY: Oxford University Press.

Perfetti, C. A. (2007). Reading ability: lexical quality to comprehension. *Sci. Stud. Read.* 11, 357–383. doi: 10.1080/10888430701530730

Reise, S. P., Bonifay, W. E., and Haviland, M. G. (2013). Scoring and modeling psychological measures in the presence of multidimensionality. *J. Pers. Assess.* 95, 129–140. doi: 10.1080/00223891.2012.725437

Richter, T., Isberner, M.-B., Naumann, J., and Neeb, Y. (2013). Lexical Quality and reading comprehension in primary school children. *Sci. Stud. Read.* 17, 415–434. doi: 10.1080/10888438.2013.764879

Russell, M. K. (2010). "Technology-aided formative assessment of learning: New developments and applications," in *Handbook of Formative Assessment*, eds H. L. Andrade and G.J. Cizek (New York, NY: Routledge), 125–138.

Salentin, K. (2014). Sampling the ethnic minority population in germany. the background to "Migration Background". *Methods Data Anal.* 8, 25–52. doi: 10.12758/mda.2014.002

Schnepf, S. V. (2007). Immigrants' educational disadvantage: an examination across ten countries and three surveys. *J. Popul. Econ.* 20, 527–545. doi: 10.1007/s00148-006-0102-y

Schulgesetz für das Land Nordrhein-Westfalen (2018). *(Schulgesetz NRW – SchulG) vom 15.Februar 2005 (GV.NRW.S.102) Zuletzt Geändert Duruch Gesetz Vom 21.Juli 2018 (SGV.NRW.223). [Engl. School Law of the Federal State North Rhine-Westphalia]*. Available online at: https://www.schulministerium.nrw.de/docs/Recht/Schulrecht/Schulgesetz/Schulgesetz.pdf

Shin, J., Deno, S. L., and Espin, C. (2000). Technical adequacy of the maze task for curriculum-based measurement of reading growth. *J. Spec. Educ.* 34, 164–172. doi: 10.1177/002246690003400305

Spencer, M., Quinn, J. M., and Wagner, R. K. (2014). Specific reading comprehension disability. major problem, myth, or misnomer? *Learn. Disabil. Res. Pract.* 29, 3–9. doi: 10.1111/ldrp.12024

Steinmetz, H. (2013). Analyzing observed composite differences across groups. *Methodology* 9, 1–12. doi: 10.1027/1614-2241/a000049

Taraban, R., and McClelland, J. L. (1990). "Parsing and comprehension: A multiple constraint view," in *Comprehension Processes in Reading*, eds D. A. Balota, G. B. Flores D'Arcais, and K. Rayner (Hillsdale, NJ: Erlbaum), 231–264.

Taylor, C. R. (2012). Engaging the struggling reader: focusing on reading and success across the content areas. *Natl. Teach. Educ. J.* 5, 51–58.

Tichá, R., Espin, C. A., and Wayman, M. M. (2009). Reading progress monitoring for secondary-school students. Reliability, validity, and sensitivity to growth of reading-aloud and maze-selection measures. *Learn. Disabil. Res. Prac.* 24, 132–142. doi: 10.1111/j.1540-5826.2009.00287.x

Tilstra, J., McMaster, K., van den Broek, P., Kendeou, P., and Rapp, D. (2009). Simple but complex. Components of the simple view of reading across grade levels. *J. Res. Read.* 32, 383–401. doi: 10.1111/j.1467-9817.2009.01401.x

Van den Bosch, R. M., Espin, C. A., Chung, S., and Saab, N. (2017). Data-based decision-making. teachers' comprehension of curriculum-based measurement progress-monitoring graphs. *Learn. Disabil. Res. Prac.* 32, 46–60. doi: 10.1111/ldrp.12122

Van Dijk, T. A., and Kintsch, W. (1983). *Strategies of Discourse Comprehension*. New York, NY: Academic Press.

Vellutino, F. R., Tunmer, W. E., Jaccard, J. J., and Chen, R. (2007). Components of reading ability. multivariate evidence for a convergent skills model of reading development. *Sci. Stud. Read.* 11, 3–32. doi: 10.1207/s1532799xssr1101_2

Waltzman, D. E., and Cairns, H. S. (2000). Grammatical knowledge of third grade good and poor readers. *Appl. Psychol.* 21, 263–284. doi: 10.1017/S014271640000206X

Wayman, M. M., Wallace, T., Wiley, H. I., Tichá, R., and Espin, C. A. (2007). Literature synthesis on curriculum-based literature synthesis on curriculum-based. *J. Spec. Educ.* 41, 85–120. doi: 10.1177/00224669070410020401

West, R. F., and Stanovich, K. E. (1978). Automatic contextual facilitation in readers of three ages. *Child Dev.* 49, 717–727. doi: 10.2307/1128240

Wilbert, J., and Linnemann, M. (2011). Kriterien zur analyse eines tests zur lernverlaufsdiagnostik [engl. criteria for analysizing a test to measure Learning Progress]. *Empirische Sonderpädagogik* 3, 225–242.

Wiley, H. I., and Deno, S. L. (2005). Oral reading and maze measures as predictors of success for english learners on a state standards assessment. *Remed. Spec. Educ.* 26, 207–214. doi: 10.1177/07419325050260040301

Witzel, J., and Witzel, N. (2016). Incremental sentence processing in japanese. A maze investigation into scrambled and control sentences. *J. Psychol. Res.* 45, 475–505. doi: 10.1007/s10936-015-9356-4

Zeuch, N., Förster, N., and Souvignier, E. (2017). Assessing teachers' competencies to read and interpret graphs from learning progress assessment. Results from Tests and interviews. *Learn. Disabil. Res. Pract.* 32, 61–70. doi: 10.1111/ldrp.12126

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.