Learning Disabilities: A Contemporary Journal 9(2), 21-44, 2011

Curriculum-Based Measurement: Developing a Computer-Based Assessment Instrument for Monitoring Student Reading Progress on Multiple Indicators

Natalie Förster
Elmar Souvignier
University of Muenster

The purpose of this study was to examine the technical adequacy of a computer-based assessment instrument which is based on hierarchical models of text comprehension for monitoring student reading progress following the Curriculum-Based Measurement (CBM) approach. At intervals of two weeks, 120 third-grade students finished eight CBM tests. To examine the construct validity, group-administered standardized achievement tests of reading, mathematics, and intelligence were applied. Results indicate that the technical adequacy of the concept is comparable to that of established measures of reading progress. In addition, the new assessment tool provides differentiated diagnostic information about single components of reading proficiency and directly assesses higher processes of comprehension. Benefits of a multidimensional measurement of reading progress for students with special educational needs are discussed.

Keywords: Reading Comprehension, Curriculum-Based Measurement, Progress Monitoring

Achieving reading competence is one of the main goals of education in elementary school, since reading competence is very important for the acquisition of knowledge and skills in academic settings (Daneman, 1991; Mullis, Martin, Gonzalez, & Kennedy, 2003). During the past two decades, one focus of reading research has been the development of assessment methods that monitor student progress in reading. The objectives of measuring student progress are the identification of students who struggle with reading and the planning and monitoring of instructional programs that address individual needs. A measurement procedure that could be used to efficiently monitor student reading progress and to evaluate the effectiveness of instructional interventions is Curriculum-Based Measurement (CBM) (Deno, 1985; Fuchs & Fuchs, 1998; Reschly, Busch, Betts, Deno, & Long, 2009; Stecker, Fuchs, & Fuchs, 2005). A substantial body of research has documented the validity of CBM reading measures (for a review, see Wayman, Wallace, Wiley, Ticha, & Espin, 2007). Yet, as Wayman et al. (2007) point out, little is known about teachers' understanding of CBM progress data and their ability to connect progress-monitoring data to instructional decisions.

The most common strategy to monitor reading progress is the measurement of reading ability by indicators that strongly correlate to standardized reading tests, namely reading fluency and maze tasks (Fuchs, 2004). Even if the predictive validity for reading difficulties is high for both indicators, they do not provide teachers with information on how to adjust the instructional program to the specific needs of a poor reader: So far, there is no analysis of component processes of reading as supposed in hierarchical models of text comprehension (Kintsch, 1998; van Dijk & Kintsch, 1983). Consequently, the aim of our study was to develop a test concept based on hierarchical models of text comprehension and to examine its use as a progress measure for reading achievement.

Assessing Reading Progress

Understanding that learners' responses to the same instructions vary (Snow, 1977), Deno (1990) suggests a formative evaluation approach to individualize instruction. In CBM, performance data are collected regularly across time to formatively evaluate the effectiveness of the intervention and eventually adjust the educational program. Since measures simultaneously integrate the various skills required for competent year-end performance and each test is an alternate form comparable in difficulty and conceptualization, slope can be used to quantify rate of learning. Taking these considerations into account, three requirements for an assessment strategy can be deduced.

First, tests have to identify those students who should be supported. They need to efficiently and accurately diagnose both students´ actual reading skills and their reading progress to help teachers identify students for special education. Although research has been supportive of the assumption that teacher judgment provides a generally accurate prediction of student achievement (Begeny, Eckert, Montarello, & Storie, 2008; Demaray & Elliot, 1998; Hoge & Coladarci, 1989), teachers showed lower levels of accuracy for lower achieving students (Demaray & Elliot, 1998; Feinberg & Shapiro, 2009) and had difficulties in identifying low-progress readers (Madelaine & Wheldall, 2005). The treatment validity approach (Fuchs & Fuchs, 1998) actually requires special education to be considered only when a child's performance reveals a dual discrepancy of substandard performance level and substandard learning rate compared to classmates. Moreover, perception of student growth might be even more difficult than judgment of student achievement.

Second, tests should provide information about how to support students. Since collection of data does not promote student achievement until instructional modifications on the basis of diagnostic information are made (Stecker et al., 2005), tests should not only identify poor achievers but also give information that can be used in the instructional decision-making process. One application to provide extensive feedback to teachers that appeared helpful to plan modifications is skills analysis (Fuchs, Fuchs, Hamlett, & Stecker, 1990; Stecker et al., 2005).

Third, after identifying students in need for special services and implementation of instructional modifications, the effects of individual adaptations have to be assessed to eventually adjust the educational program (Deno, 1990; Fuchs & Fuchs, 1998). This evaluation process requires repeated assessments, so tests must be sensitive to change in performance.

Whereas standardized tests provide information regarding the first two requirements on assessment, they are not suitable for repeated measurement, a precondition of formative evaluation. Thus, a key challenge in assessing (reading) progress is the identification of measurement tasks that (a) simultaneously integrate the various skills required for competent year-end performance, (b) are sensitive to student growth, and (c) are designed to be given on a frequent and repeated basis in school.

Two approaches are used to design CBM-tasks (Fuchs, 2004). One approach, termed *curriculum sampling*, involves systematic sampling of the skills constituting the annual curriculum in such a way that each test represents the curriculum equivalently. This method is often used in math (Foegen, Jiban, & Deno, 2007), whereby each test includes the same sub skills in the same proportion; for instance, addition, subtraction, multiplication, and division problems. For the second approach, termed *robust indicators*, measures are identified that represent broadly defined proficiency and correlate robustly with the various component skills that constitute the academic domain. In the area of reading, oral reading fluency and the maze task have been identified as valid indicators for reading progress (Reschly et al., 2009; Wayman et al., 2007).

Oral Reading Fluency

Most research of all CBM measures has been about the read-aloud measure (R-CBM) (Reschly et al., 2009). In R-CBM, students read aloud for 1 minute from a reading passage, typically a passage of their grade or instructional level. The number of words read correctly is scored (Deno, 1985). Omissions, insertions, substitutions, hesitations, and mispronunciations are marked as errors, unless they are immediately corrected. Overall, research demonstrates great support for the technical adequacy of the reading-aloud measure (Reschly et al., 2009; Wayman et al., 2007). A current meta-analysis of the correlational evidence for R-CBM as an indicator of reading achievement by Reschly et al. (2009) found a moderately high (weighted average $r = .67$) association between R-CBM probes and standardized tests of reading achievement. However, heterogeneous results have been found regarding the relation between the single measure of reading aloud and reading comprehension, especially for older students (Hosp & Fuchs, 2005; Jenkins & Jewell, 1993; Kranzler, Miller, & Jordan, 1999).

Furthermore, there are concerns for feasibility, since reading aloud can be realized only in one-by-one-assessment settings. If assessments are to be realized every one or two weeks, more economical settings as group-administered tests and computer-based measurements are necessary in regular education, both unfeasible with reading aloud.

Maze Selection

The second indicator of reading proficiency considers these barriers. In maze selection, passages are read in which usually every seventh word has been deleted and replaced with three word choices—the correct missing word and two distractors. Students read silently for 1 to 3 minutes, making selections while they read. The number of correct selections is scored. In addition to practical advantages, the maze task appears to be more of a reading comprehension measure than the reading aloud measure because conclusions must be drawn for selection. Jenkins and Jewell (1993) conducted a direct comparison of reading aloud and maze selection across Grades 2 to 6 and found moderate to strong within-grade correlations for both measures, ranging from $r = .58$ to $r = .88$. However, while correlations between reading aloud and reading comprehension declined from Grades 2 to 6, they remained consistent across grade levels for the maze task. Beyond its criterion validity, the maze task has been proven to be sensitive to change in performance over time (Fuchs & Fuchs, 1992) and thus serves well for progress monitoring in reading achievement.

Hence, for the assessment of reading progress, two valid indicators of student achievement exist. Unfortunately, both measures fail to provide differentiated diagnostic information about single components of reading achievement that can be used in the instructional decision-making process, which was one requirement of assessment. Especially with respect to different aspects of reading comprehension, it seems worthwhile to attend to hierarchical models of text comprehension.

Component Processes of Reading

Comprehension is the most important goal of reading. It is widely accepted that text comprehension can be described as the process of constructing a mental representation of the information that is contained in the text (Kintsch, 1998; van Dijk & Kintsch, 1983). In hierarchical models of text comprehension, two mental representations are differentiated: the textbase and the situation model. To form a textbase representation, the reader needs to connect words, phrases and sentences, and build semantic relations between these text elements to reach a local, coherent representation of the text. These microstructure connections can be assigned to lower-level reading processes. In contrast, higher-level reading processes that relate to global, macrostructure connections between the text content and prior knowledge constitute the situation model. Thus, the textbase, as the product of the lower-level reading processes which contains only propositions that were explicit in the text, is the basis of the situation model. It is assumed that for a skilled reader, the lower-level processes of reading are effortless, automated, and unconscious. Generating a situation model, on the other hand, involves strategy-oriented active construction of meaning.

Comparing the current R-CBM measures with hierarchical models of text comprehension, it becomes clear that both tasks address only lower-level reading processes. Reading aloud from a text provides information about the ease with which the reader recognizes the printed words and thus to which degree sub-processes of reading are automated. Although some connections between text elements are necessary to make selections in the maze task, the construction of a situation model is not required.

There is considerable agreement that individual differences in reading achievement are not only ascribed to lower-level reading processes but also to higher-level comprehension processes with the latter becoming more meaningful as reading experience and reading performance increase (Daneman, 1991). Thus, it seems worthwhile to provide teachers not only with information about reading rate and reading accuracy but also with information about whether students are able to master the construction of a deeper understanding of the macrostructure of a text and to generate a situation model.

Going beyond R-CBM measures, some standardized tests of reading achievement are based on hierarchical models of text comprehension (Adam-Schwebe, Souvignier, & Gold, 2009; Mullis et al., 2003). They share the differentiation of two processing depths for the textbase and the situation model proposed by Kintsch (1998) and give a basis for the development of our test concept that differentiates between lower-level and higher-level reading processes.

The Test Construction

Following the requirements on assessment and considering assumptions of hierarchical models of text comprehension, we developed a computer-based measurement system to monitor student reading progress on both basal reading skills as reading rate and reading accuracy and higher processes of reading comprehension.

All tests (for third graders) are based upon Aesop's fables, which were chosen for several reasons: Fables are short, self-contained stories with different characters and a real moral. Thus, they are demanding for children because the generation of a situation model is required for overall text comprehension. In addition, children are typically unfamiliar with this reading material, since fables, in contrast to fairy tales, are usually not read at home. Individual differences in reading comprehension therefore should not be affected by previous knowledge about the text in a way that is not dependent on text comprehension. Besides, fables are listed in the curriculum for third-graders.

The test concept consists of two parts: Since the maze task has proven to be adequate and practicable for monitoring student reading progress on basal reading skills, a maze task is presented first. In this maze task every seventh word has been deleted and replaced with three word choices—one correct choice and two distractors. Distractors were chosen from the same lexical items (nouns, verbs, adverbs, etc.) as the correct word and had the same number of syllables. Thus, text-level processes that compute the syntactic and semantic relationships among successive text elements are assessed. Two pieces of information are recorded with the maze task: the number of correct selections (maximum 17) as a measure of reading accuracy and the time needed to complete the text as reading rate (words read per minute). After finishing the maze task, questions are presented to evaluate the depth of comprehension. At the same time, the correct fable is visible. Following the differentiation of two processing depths for the textbase and the situation model (Kintsch, 1998), two different kinds of questions have been developed: Questions asking for information explicitly contained in the text (text-based questions) and inference questions (knowledge-based questions) asking for overall text comprehension (e. g., "What saying belongs to the fable?"). The former assess if a propositional representation of the text (textbase) is constructed, while the latter examines the students' ability to make appropriate inferences, to integrate information with prior knowledge, and to generate a situation model. The number of correct answers (maximum 6, respectively) is recorded.

A total of eight tests have been developed following this test concept. To ensure parallelism of the tests, the following criteria have been considered: number of words, number of sentences, number of words per sentence, number of different words, and number of syllables. Resulting Flesch-indices (Flesch, 1948) vary between 75 and 91, indicating similar difficulty.

Research Questions

Since research on the appropriateness of a measure for the purpose of progress monitoring always starts with an investigation of the technical features of the static score (Fuchs, 2004), the aim of this study was to examine whether the new computer-based concept for monitoring student reading progress meets the demands of CBM-measures including reliability, validity, parallelism, and sensitivity. In addition, whether the computer-based assessment of reading progress could be easily implemented into regular classroom settings was explored.

Reliability

Internal consistency is expected to be high for all eight tests. Correlations between consecutive tests are supposed to be high, proving good parallel-forms reliability. Additionally, these correlations should be stronger for consecutive tests than for more distant tests.

Validity

To examine the construct validity of the concept, a group-administered standardized test of reading achievement was applied at the beginning and at the end of the study, assessing reading comprehension on the levels of word comprehension, sentence comprehension, and text comprehension (ELFE 1-6) (Lenhard, & Schneider, 2006). Since reasoning is involved especially in higher-level reading processes, reasoning was assessed with a measure of intelligence (CFT 20-R) (Weiß, 2006) to further investigate differences of question type (text-based or knowledge-based). Furthermore, a group-administered standardized achievement test of mathematics was applied to test for discriminant validity (DEMAT 3+) (Roick, Gölitz, & Hasselhorn, 2004).

Drawing on previous research on validity of reading fluency and the maze task (Reschly et al., 2009; Wayman et al., 2007) and considering theoretical assumptions of hierarchical models of text comprehension (Kintsch, 1998; van Dijk & Kintsch, 1983), we assume the correlation pattern to be as follows: Correlations between reading rate and the word comprehension subtest, and reading accuracy and the sentence comprehension subtest respectively, are expected to be strong (Adam-Schwebe et al., 2009), since both former measures assess the ease of word identification, while the latter assess the generation of local, microstructure connections that connect preceding text elements. Likewise, correlations between text-based and knowledge-based processes of reading comprehension and the text-level subtest (ELFE) are expected to be strong. On the other hand, correlations are expected to be lower between the CBM reading measures and the standardized achievement test in mathematics (DEMAT 3+) compared to the correlation between the CBM reading measures and the ELFE-test. Again, strong correlations are expected between knowledge-based questions and the measure of fluid intelligence since the generation of a situation model requires reasoning processes.

Parallelism

Development of the tests followed clear rules for construction, concerning a number of criteria with respect to texts, distractors and questions. Hence, no differences between tests are expected regarding overall difficulty of items (distractors and questions) and time needed to finish tests.

Sensitivity

All measures are expected to be sensitive to student growth, showing linear trends of development for the whole sample of students.

Method

Participants and Design

120 third-grade students (65 female, 55 male) from four elementary schools in a medium-sized German town participated in the study. Students were approximately 8 years old (*M* = 8.3 years, SD = .6 years, range = 7 –10 years). All schools are public schools. The primary language of the children was German (81%). Participation in the study was voluntary.

The study involved a pre- and posttest design. Data was collected between autumn 2008 and summer 2009, starting with the ELFE-pretest and ending with three posttests (ELFE, DEMAT, CFT-20R). Group-administered standardized achievement tests were given by trained university student assistants. Students finished eight computer-based CBM tests at intervals of two weeks between pretest and posttests. Except for the first CBM test, sequence of tests was varied for different groups of children to balance test and time of measurement. In doing so it was possible to examine parallelism of tests and the sensitivity of measures (see Table 1 for design). Classes were split into halves and these groups were randomly assigned to one sequence of tests. CBM tests were finished during regular education settings. Pre- and posttest materials were administered within two weeks before and after the CBM tests.

Measures

At pretest and posttest, all students completed the standardized reading comprehension test ELFE 1-6 (Lenhard & Schneider, 2006). The test collects information on word-, sentence-, and text-level. In word comprehension, pictures are presented and students have to select the correct word that matches the picture out of four choices. Distractors graphemically and phonemically resemble the correct word and consist of the same number of syllables. Students are given three minutes time to answer as many items as possible from a total of 72 pictures.

The second subtest records sentence comprehension with a maze task. Students have to decide which word fits into a sentence. Five word choices are presented—one correct choice and four distractors. For the 28 sentences, targets can be nouns, verbs, adjectives, conjunctions or prepositions. Time is again limited to three minutes.

The final subscale provides information about text comprehension. Students read short texts and have to find the correct answer to a question among four choices. This subscale consists of 20 items. Answers require retrieval of information given in the text. The number of correct answers given within seven minutes is recorded.

Two parallel forms of the test are available with the sequence of choices changed between the versions. Thus, both forms were applied in pre- and posttest giving different forms to students sitting next to each other. For third-graders, internal consistency and odd-even-split-half-reliability range from .86 to .96 (Cronbachs α) and .83 to .93 (r_{tt}). Lenhard and Schneider (2006) state the criterion validity to be high for word and sentence comprehension (.77 and .79) and moderate for text comprehension (.55).

Table 1. *Design*

	Pretest	CBM1	CBM2	CBM3	CBM4	CBM5	CBM6	CBM7	CBM8	Posttests
sequence 1		1	2	3	4	5	6	7	8	
sequence 2		1	8	2	3	4	5	6	7	
sequence 3		1	7	8	2	3	4	5	6	ELFE
sequence 4	ELFE	1	6	7	8	2	3	4	5	DEMAT
sequence 5		1	5	6	7	8	2	3	4	CFT20-R
sequence 6		1	4	5	6	7	8	2	3	
sequence 7		1	3	4	5	6	7	8	2	

Note. Fables: 1 = The Fox and the Stork; 2 = The Tortoise and the Hare; 3 = The Lion and the Bear; 4 = The Ant and the Cricket; 5 = The Mouse and the Frog; 6 = The Fox and the Crow; 7 = The Lion and the Mouse; 8 = The Dove and the Ant.

DEMAT 3+. Students´ performance in mathematics was measured with a German mathematics test for third-graders (DEMAT 3+; Roick et al., 2004). Items record achievement in arithmetic, story problems and geometry and thus include all tasks required in the school-curriculum. The split-half reliability is .85, Cronbach´s alpha is .83 and parallel-forms reliability is .83. The correlation with a state-wide mathematical comparative class tests is .66.

CFT 20-R. Like the DEMAT 3+, students completed the CFT 20-R (Weiß, 2006) in summer 2009. To realize test execution within one lesson, students finished the short version of this test, which consists of four subtests: series completion, classification, matrices and topologies. Test time was extended for one minute per subtest as proposed for elementary school students (Weiß, 2006). Reliability is high (.92) and correlations between the CFT 20-R subtests and the "g"-factor are strong (.78 - .83).

CBM-Measures

The concept of the newly developed CBM tests has been presented previously. In each CBM test the following measures were recorded: reading rate (RR), reading accuracy (RA), reading comprehension in text-based questions (RC-TB) and reading comprehension in knowledge-based questions (RC-KB).

Reading rate was defined as time needed to complete the maze task. It was calculated with the number of words contained in the fable, including headline and distractors. Thus, the number of words per minute was included in the analysis. Reading accuracy was measured as the number of correct selections in the maze task with a maximum of 17 right choices. Reading comprehension (RC-TB and RC-KB) was recorded as the number of correct multiple-choice answers with a maximum of six points, respectively. Since the maze task usually combines reading rate and reading accuracy by counting the number of correct selections during limited time, a combined maze-measure was calculated by adding z-transformed data of reading rate and reading accuracy (maze). Likewise, a total score of all four measures was calculated (CBM_t).

Procedure

CBM tests were available via internet with a personal log-in. Depending on the number of computers available in the classroom or in a computer room, students finished one fable during self-study periods or in group sessions within two weeks. Before completing the first CBM test, handling was practised in a short exercise.

After each CBM test, teachers obtained individual data for each student and for the whole class. Also, growth-curves for the class and for every student were put into graphs to illustrate achievement gains. Since the four measures were differentiated, four graphs occurred. After each test, students obtained feedback about their performance on the current test and the former tests.

Data Analysis

Because of some difficulties (illness of a teacher, project week, school trip), two classes could not complete the eighth CBM test. Hence, results of the seventh CBM test were included in relevant analysis for these students. Depending on the research question (validity, alternate-form reliability, sensitivity), time of measurement was the unit of analysis regardless of the respective fable, or the unit of analysis was type of fable (regardless of time of measurement) to check for internal consistency and for parallelism.

To calculate Cronbach's alpha, maze selections and questions were included in the analysis. Alternate-form reliabilities were calculated for total scores on CBM tests (CBM_t) by Pearson *r* correlation coefficients for scores obtained from neighbouring data points (i. e., score from Time 1 was correlated with score from Time 2, Time 2 with Time 3, and so on), disregarding type of fable.

The criterion validity of the CBM measures was examined by calculating Pearson r coefficients between the scores at the first CBM test and the ELFE pretest and the final CBM test and the ELFE posttest, respectively. Thus, correlation between ELFE and CBM 1 shows validity for "The Fox and the Stork" and correlation between ELFE, DEMAT 3+ and CFT 20-R and CBM 8 demonstrates overall validity for the remaining fables (see table 1 for design). In addition to the four CBM measures—reading rate, reading accuracy, reading comprehension-TB, and reading comprehension-KB—the computed maze measure and the total CBM score were included in the analysis of validity.

Since similar difficulty of tests is crucial for progress monitoring, difficulties of items were separately calculated for the maze-task (gaps), text-based questions and knowledge-based questions on each fable. Difficulty was corrected for chance.

To examine the presence and the nature of growth across time, a repeated measures analysis of variance was conducted for each CBM measure (reading rate, reading accuracy, reading comprehension-TB and reading comprehension-KB). Since growth between first and second assessment can be ascribed to experience with test execution, and data of two classes were missing for the eighth test, only six data points (Time 2 to Time 7) were included in this analysis to give a conservative estimation of growth. A post hoc analysis of main effects was conducted with a Bonferroni adjustment for multiple comparisons.

Results

Descriptive statistics of students' performance on all CBM-measures and the standardized achievement tests are reported in Table 2. Except for reading comprehension-TB, all measures show strong differences between first and second measure, as was expected. However, overall increasing patterns of data from autumn to summer were observed only for reading rate and for reading comprehension-KB. No growth-rates could be observed for reading accuracy and reading comprehension-TB.

Table 2. ***Means and Standard Deviations for CBM and Standardized Tests***

	RR		RA		RC-TB		RC-KB	
	M	SD	M	SD	M	SD	M	SD
1	33.08	13.29	12.17	3.22	4.62	1.62	2.61	1.51
2	40.01	16.51	13.91	2.77	4.58	1.48	3.03	1.61
3	37.45	13.60	13.71	2.63	4.67	1.36	3.48	1.51
4	44.43	15.82	13.67	2.94	4.55	1.48	3.62	1.55
5	46.43	14.79	13.85	2.65	4.73	1.45	3.54	1.59
6	46.93	17.51	13.89	2.93	4.65	1.46	3.46	1.69
7	49.87	16.77	13.92	2.87	4.72	1.35	3.62	1.66
8	49.98	15.87	13.73	2.99	4.50	1.50	3.47	1.67
	ELFE 1-6 pretest[a] N = 120		ELFE 1-6 posttest[a] N = 107		DEMAT 3+[b] N = 116		CFT 20-R[c] N = 110	
	56.73	17.11	73.13	18.02	19.54	5.66	31.17	6.10

Note. RR = reading rate; RA = reading accuracy, maze = RR and RA combined; RC-TB = reading comprehension on text-based questions; RC-KB = reading comprehension on knowledge-based questions; a: maximum = 120; b: maximum = 31; c: maximum = 57

Reliability

To check for reliability, internal consistency was calculated according to the respective fables and the parallel-forms reliability was analysed considering the time of measurement. Cronbach´s alpha varies between .77 and .90 (median = .86). Except for the fable "The Tortoise and the Hare", internal consistencies exceed .80 and are thus up to standard for reliability. Alternate-form reliabilities of total test scores with two weeks intervals between testing ranged from .61 to .74 (mean = .68) and were statistically significant ($p < .01$ for all comparisons). The correlation for four weeks intervals ranged from .58 to .69 (mean = .64) (see Table 3).

Validity

The criterion validity of the CBM measures is displayed in Table 4. Overall, correlation pattern between CBM measures and the ELFE subtests at pretest is replicated at posttest. All validity coefficients were statistically significant with most $p < .01$ (for the ELFE subtests and the CBM measures), except for the relation between reading comprehension-KB and the ELFE-word-comprehension subtest at posttest ($r = .15$). Strong correlations were found for total test scores at pre- and posttest (.71, respectively).

For subtests and single CBM measures, correlations between reading rate and word comprehension, reading accuracy and sentence comprehension, and reading comprehension-TB and text comprehension are strong ($.50 \leq r \leq .60$) as expected, except for reading rate and word comprehension at posttest ($r = .36$). In contrast, relations between reading comprehension-KB and text comprehension were only moderate with $r = .34$ at pretest and $r = .42$ at posttest.

As each ELFE subtest is time-limited and hence combines accuracy with speed, strong correlations arise for the maze-measure (.56 to .78) which combines reading rate and reading accuracy on the CBM tests.

Correlations between CBM measures and the DEMAT 3+ were low to moderate, ranging from $r = -.10$ to $r = .39$, indicating discriminant validity. Again, low to moderate correlations were found for the CFT 20-R and the CBM measures ranging from $r = -.10$ to $r = .31$ with the highest correlation between reading comprehension-KB and reasoning as expected with respect to hierarchical models of text comprehension.

Table 3. *Correlations Between Total Test Scores for Two Weeks Intervals With Alternate Forms*

	CBM 1	CBM 2	CBM 3	CBM 4	CBM 5	CBM 6	CBM 7
CBM 2	0.61						
CBM 3	0.66	0.65					
CBM 4	0.65	0.59	0.73				
CBM 5	0.70	0.61	0.69	0.66			
CBM 6	0.63	0.62	0.62	0.58	0.71		
CBM 7	0.63	0.59	0.62	0.62	0.62	0.66	
CBM 8	0.65	0.62	0.69	0.66	0.68	0.67	0.74

Table 4. ***Correlations Between CBM-Measures and Criterion Measures***

	RR	RA	maze	RC-TB	RC-KB	CBM-t
N = 120 ELFE (pretest)	.58**	.65**	.78**	.55**	.28**	.71**
- word-comprehension	.50**	.44**	.60**	.40**	.19*	.53**
- sentence-comprehension	.51**	60**	.71**	.53**	.19*	.63**
- text-comprehension	.53**	.65**	.74**	.52**	.34**	.70**
N = 105 ELFE (posttest)	.36**	.53**	.70**	.60**	.37**	.71**
- word-comprehension	.36**	.36**	.57**	.41**	.15	.49**
- sentence-comprehension	.36**	.50**	.68**	.54**	.35**	.68**
- text-comprehension	.23*	.49**	.56**	.58**	.42**	.66**
N = 114 DEMAT 3+	-.10	.38**	.22*	.39**	.29**	.37**
N = 108 CFT 20-R	-.10	.26**	.12	.23*	.31**	.26**

Note. RR = reading rate; RA = reading accuracy, maze = RR and RA combined; RC-TB = reading comprehension on text-based questions; RC-KB = reading comprehension on knowledge-based questions; CBM-t = RR, RA, RC1 and RC2 combined; *p < .05; **p < .01.

Table 5. ***Mean Item Difficulties of the Reading Tasks***

	1	2	3	4	5	6	7	8
gaps	.57	.76	.68	.71	.72	.74	.80	.73
text-based questions	.69	.67	.63	.69	.71	.75	.82	.73
knowledge-based questions	.24	.27	.41	.50	.31	.55	.59	.51

Note: Fables: 1 = The Fox and the Stork; 2 = The Tortoise and the Hare; 3 = The Lion and the Bear; 4 = The Ant and the Cricket; 5 = The Mouse and the Frog; 6 = The Fox and the Crow; 7 = The Lion and the Mouse; 8 = The Dove and the Ant.

Parallelism

Average item difficulties of gaps, text-based questions and knowledge-based questions are displayed in Table 5. All scores of fable no. 1 “The Fox and the Stork” stem from the first CBM test; the remaining scores reflect average item difficulties of corresponding fables over seven measurement points (see Table 1 for design). Thus, higher difficulty of items in “The Fox and the Stork” can be ascribed to lower reading achievement and are not included in the analysis. Overall, item difficulties are very similar between fables ranging from .68 to .80 for gaps, .63 to .82 for text-based questions and .27 to .59 for knowledge-based questions. Inspection of difficulties revealed necessity to adapt only single distractors to further increase parallelism. Furthermore, fables were compared regarding the time needed to finish the tests. Because of the exceptional position of the first fable “The Fox and the Stork”, it was excluded from analysis. The average time needed to finish the fables regardless on time of measurement was very similar, ranging from 9 min and 10 sec to 9 min and 55 sec (median = 9 min and 47 sec), indicating good parallelism.

Sensitivity

All four CBM measures were examined in terms of their capacity to detect student growth. A repeated-measures analysis of variance was conducted for each measure, including scores from Time 2 to Time 7. Since Mauchly´s Test shows violation of sphericity ($p < .05$), degrees of freedom were corrected using Huynh-Feldt correction ($\varepsilon > 0.75$) (Huynh & Feldt, 1976). Growth of reading rate and reading comprehension-KB reached statistical significance; $F_{(4.15, 460.468)} = 18.54$, $p < .001$ and $F_{(4.79, 531.747)} = 3.15$, $p < .01$, respectively. No significant growth was found for reading accuracy and reading comprehension-TB. While a linear trend was found for reading rate, $F_{(1, 111)} = 41.73$, $p < .001$, growth in reading comprehension-KB follows a quadratic trend $F_{(1, 111)} = 5.44$, $p < .05$, as can be seen in Figure 1a and 1b.

Feasibility

CBM requires repeated test executions, so computer-based measures are essential to alleviate data collection and data evaluation to permit implementation of CBM in regular education. Since computer rooms were not available at all schools, students finished CBM tests during self-study periods within two weeks. Teachers reported no difficulties with usability and indicated satisfaction with the procedure. No external support was needed for implementation. It can be regarded as an indicator of high acceptance of the process measurement that all schools agreed to participate in a following investigation.

***Figure 1*. Growth on reading rate from Time 2 to Time 7.**

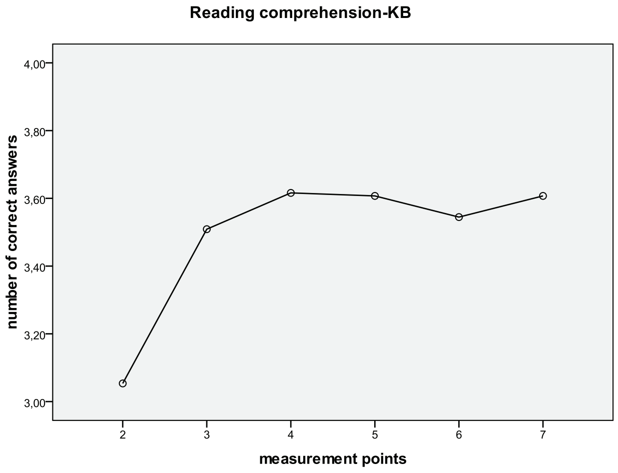
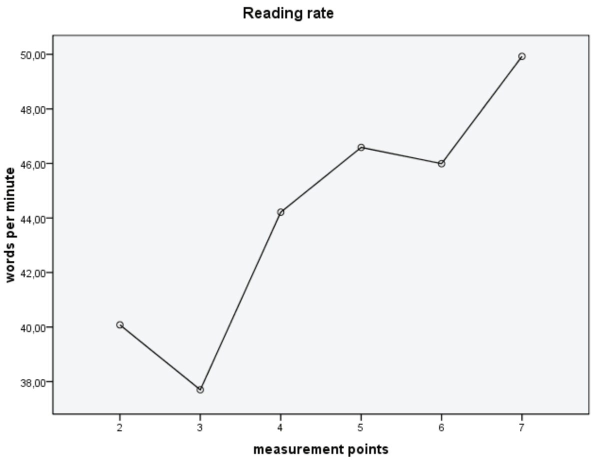


***Figure 2*, Growth on reading comprehension-KB scores from Time 2 to Time 7.**



Discussion

This study examined the reliability, validity, parallelism, and sensitivity of a new test concept based on hierarchical models of reading comprehension to monitor student growth in reading achievement during third grade. Findings support the technical soundness of the concept in the following ways.

First, results provide evidence for the reliability of the tests. All fables demonstrate good internal consistencies. Alternate-form reliabilities are satisfying, but not as high as expected and lower than those found for the maze task by Shin, Deno, and Espin (2000). This finding might be a consequence of the computer-based test setting, where students worked independently on the CBM tests in contrast to the more controlled settings with paper-pencil group tests.

Second, results provide evidence for the validity of the concept. Correlations between the standardized test of reading comprehension (ELFE 1-6) and the CBM total test scores appeared to be rather strong, whereas they turned out to be lower for the DEMAT 3+. In order to examine the appropriateness of the test concept to differentiate between lower-level and higher-level processes of reading comprehension, as proposed by hierarchical models of reading, a more detailed view was taken by investigating the relations between single CBM measures and the ELFE subscales. Correlations between reading rate, reading accuracy and reading comprehension-TB and the corresponding ELFE subtests were moderate to strong.

Interpretation of the correlations must take into account that all ELFE scores contain information of both accuracy and speed, while CBM measures contain either speed (RR) or accuracy (RA, RC-TB, RC-KB). Thus, the maze measure, which combines reading rate and reading accuracy and the total CBM score, show the highest relations to the ELFE scores.

However, correlations between reading comprehension-KB and the ELFE text-comprehension subtest are only moderate. Taking a closer look to the items of the text-comprehension subscale, two main differences compared to the comprehension-KB questions become evident.

First of all, the ELFE questions refer to much shorter texts which consist of at least 7 and at most 56 words, with only one of them having more than five sentences. In contrast, the number of words in the fables varies between 136 to 142, with at least ten sentences.

Second, none of the answers in the text-comprehension subscale requires the construction of a situation model as it is supposed in hierarchical models of text comprehension (Kintsch, 1998; van Dijk & Kintsch, 1983), since the integration of prior knowledge is not needed. In fact, most inferences can be drawn from only one sentence. Thus, the correlation pattern of the strong correlations between comprehension-TB and the text-comprehension subscale on one hand, and only moderate correlations between the comprehension-KB questions and this subscale on the other hand can be interpreted as a proof of two different processing depths for the CBM questions. Additional support for this interpretation comes from the differences in mean item difficulties for the text-based and knowledge-based questions as well as the stronger correlations between the knowledge-based questions and reasoning. However, further investigations of the comprehension-KB questions concerning the depth of reading comprehension and the construction of a situation model are needed. Considering these aspects, the correlation pattern between the single CBM measures and the ELFE subscales provides some evidence that the newly developed CBM-concept differentiates between lower-level and higher-level processes of reading. Compared to former results for the maze measure (Ardoin et al., 2004; Jenkins & Jewell, 1993), our findings for the total test score (CBM_{total}) exceed correlations between maze and reading achievement and reading comprehension (see Table 6).

Third, a precondition to monitor student progress is the availability of alternate tests comparable in difficulty and conceptualization. Because of the design, item difficulties of seven out of eight fables were inspected, revealing similar difficulties for gaps, text-based questions and knowledge-based questions over all points of measurement. Nevertheless, adaptations are indicated for three fables. Though the time needed to finish the tests was not regulated at all, the average time of test execution was similar between fables; this again indicates utility for progress monitoring.

Fourth, with respect to student growth, two out of the four CBM measures showed statistically significant progress over six measurement points. Reading rate and reading comprehension-KB suggested that they are sensitive to improvement over a short period of time. While growth in reading rate follows a linear trend as expected, a quadratic trend was found for reading comprehension-KB. Contrary to the expectation, no growth occurred for reading accuracy and reading comprehension-TB. As shown in table 2, a ceiling effect for these measures occurs, suggesting that for most of the children the selection of the correct answer is easy, if time is not limited. However, analysis of frequencies reveals that some children do have problems even with these basal reading comprehension processes. Hence, further research should investigate the benefits of these two measures for reading progress especially for poor readers.

For reading rate, results can be compared to growth rates for the read aloud measure observed in previous research. The growth of reading rate per week (1.02 words) was within the range of growth rates (0.76 to 1.18) reported at third-grade level (Deno, Fuchs, Marston, & Shin, 2001; Fuchs & Fuchs, 1992; Fuchs, Fuchs, Hamlett, & Walz, 1993; Graney, Missall, Martinez, & Bergstrom, 2009).

Besides, all teachers showed satisfaction with test execution and preparation of the results. The computer-based concept was easily implemented in regular education settings although time needed to finish tests was thrice as long as a usual maze task. Students finished CBM tests independently during self study periods.

Limitations and Implications for Future Research

In summary, results provide evidence for the test concept regarding reliability, validity, and parallelism, and limited evidence for sensitivity. The computer-based instrument can be implemented in regular education settings. However, several limitations of this study should be acknowledged.

First of all, findings should be viewed only as preliminary and must be replicated in a larger sample.

Second, the actual design varied fable and time of measurement, but sequence of fables was only rotated. Thus, the same fables followed each other. Because of small samples in each group, differences in sequence were not analysed. Since some animals occur in more than one fable, conclusions about characters—which was asked for as a knowledge-based question—might be influenced by former fables.

Table 6. ***Comparison of Correlations Between Former Maze Scores and Actual Total Test Scores and Criterion Variables of Reading Achievement and Reading Comprehension at Pre- and Posttest***

	CBM_{maze} Jenkins & Jewell, 1993	CBM_{maze} Ardoin et al., 2004	CBM_{total}
reading achievement	.66	.50	.71/.71
reading comprehension	.64	.31	.70/.66

Third, two out of four measures showed no statistically significant progress between autumn and summer. Results indicate that for reading accuracy and reading comprehension-TB, high scores are already obtained in the first test. Achievement gains appear for reading rate and reading comprehension-KB. Nevertheless, reading accuracy and reading comprehension-TB could be sensitive measures for detecting students at risk for reading disabilities. Future research should explore if these measures provide useful information about reading progress of poor readers and examine their role in predicting reading disabilities.

While this study explored the nature of reading growth on each measure for the whole sample, future research should also investigate slopes for individual students. As a general finding it is known that considerable inter-individual differences in slopes occur (Fuchs et al., 1993; Graney et al., 2009). However, little is known about differences in character of individual reading growth and consequences for academic goal-setting or its use for individualized instructions (Graney et al., 2009; Shin et al., 2000). Finally, although results provide evidence for the test concept and although satisfaction with test execution in regular education was high, the time needed to finish tests is considerably longer than that for the maze task only.

However, especially for low-performing students, the differentiated information provides teachers with information about specific strengths and weaknesses in reading. This information on the basis of component processes of reading may be useful to adapt instruction to the individual needs of the children. As a consequence, students with poor reading fluency might be promoted with interventions to enhance reading fluency (e.g., repeated reading or paired reading; see Topping, 2006), while students with difficulties in reading comprehension would be assigned to programs that focus on reading strategies (Antoniou & Souvignier, 2007; Guthrie, Wigfield, & Perencevich, 2004; Palincsar & Brown, 1984). Thus, differentiated diagnostic information might be a tool that is especially helpful in teaching children with reading disabilities. It seems worthwhile for future research to explore if teachers really "work" with this diagnostic information in terms of planning adaptive instructional approaches or if they need additional coaching (Stecker et al., 2005).

Conclusion

Reading is one of the basic skills that students learn. Regarding the consequences for success in school, it is important to find assessments that help educators to efficiently monitor the progress of students' reading skills. This allows instructions to be adapted to individual needs. The present study investigated the usefulness of a new computer-based instrument for monitoring student reading progress on component processes of reading in general education. Overall, results are promising. Further research is needed to evaluate if the differentiation of component processes of reading helps educators to identify students at risk for reading disabilities and to determine the sub-skills on which to focus instruction.

References

Adam-Schwebe, S., Souvignier, E., & Gold, A. (2009). The Frankfurt Reading Comprehension Test (FLVT 5-6). In W. Lenhard & W. Schneider (Eds.), *Diagnostik und Förderung des Leseverständnisses* (pp. 113-130). Göttingen: Hogrefe.

Antoniou, F., & Souvignier, E. (2007). Strategy instruction in reading comprehension: An intervention study for students with learning disabilities. *Learning Disabilities: A Contemporary Journal, 5*, 41-57.

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 Psychology Review, 33*(2), 218-233.

Begeny, J. C., Eckert, T. L., Montarello, S. A., & Storie, M. S. (2008). Teachers' perceptions of students' reading abilities: An examination of the relationship between teachers' judgments and students' performance across a continuum of rating methods. *School Psychology Quarterly, 23*(1), 43-55.

Daneman, M. (1991). Individual differences in reading skills. In R. Barr, M. L. Kamil, P. B. Mosenthal, & P. D. Pearson (Eds.), *Handbook of reading research, Vol. 2* (pp. 512-538). Hillsdale, NJ, England: Lawrence Erlbaum Associates Inc.

Demaray, M. K., & Elliot, S. N. (1998). Teachers' judgments of students' academic functioning: A comparison of actual and predicted performances. *School Psychology Quarterly, 13*(1), 8-24.

Deno, S. L. (1985). Curriculum-based measurement: The emerging alternative. *Exceptional Children, 52*(3), 219-232.

Deno, S. L. (1990). Individual differences and individual difference: The essential difference of special education. *The Journal of Special Education, 24*(2), 160-173.

Deno, S. L., Fuchs, L. S., Marston, D., & Shin, J. (2001). Using curriculum-based measurement to establish growth standards for students with learning disabilities. *School Psychology Review, 30*(4), 507-524.

Feinberg, A. B., & Shapiro, E. S. (2009). Teacher accuracy: An examination of teacher-based judgments of students' reading with differing achievement levels. *Journal of Educational Research, 102*(6), 453-462.

Flesch, R. (1948). A new readability yardstick. *Journal of Applied Psychology, 32*(3), 221-233.

Foegen, A., Jiban, C., & Deno, S. (2007). Progress monitoring measuring in mathematics: A review of the literature. *The Journal of Special Education, 41*(2), 121-139.

Fuchs, L. S. (2004). The past, present, and future of curriculum-based measurement research. *School Psychology Review, 33*(2), 188-192.

Fuchs, L. S., & Fuchs, D. (1992). Identifying a measure for monitoring student reading progress. *School Psychology Review, 21*(1), 45-58.

Fuchs, L. S., & Fuchs, D. (1998). Treatment validity: A unifying concept for reconceptualizing the identification of learning disabilities. *Learning Disabilities Research & Practice, 13*(4), 204-219.

Fuchs, L. S., Fuchs, D., Hamlett, C. L., & Stecker, P. M. (1990). The role of skills analysis in curriculum-based measurement in math. *School Psychology Review, 19*(1), 6-22.

Fuchs, L. S., Fuchs, D., Hamlett, C. L., & Walz, L. (1993). Formative evaluation of academic progress: How much growth can we expect? *School Psychology Review, 22*(1), 27-48.

Graney, S. B., Missall, K. N., Martinez, R. S., & Bergstrom, M. (2009). A preliminary investigation of within-year growth patterns in reading and mathematics curriculum-based measures. *Journal of School Psychology, 47*(2), 121-142.

Guthrie, J. T., Wigfield, A., & Perencevich, K. C. (2004). *Motivating reading comprehension: Concept-oriented reading instruction*. Mahwah, NJ: Lawrence Erlbaum.

Hoge, R. D., & Coladarci, T. (1989). Teacher-based judgments of academic achievement: A review of literature. *Review of Educational Research, 59*(3), 297-313.

Hosp, M. K., & Fuchs, L. S. (2005). Using CBM as an indicator of decoding, word reading, and comprehension: Do the relations change with grade? *School Psychology Review, 34*(1), 9-26.

Huynh, H., & Feldt, L. S. (1976). Estimation of the Box correction for degrees of freedom from sample data in randomized block and split-plot designs. *Journal of Educational Statistics, 1*(1), 69-82.

Jenkins, J. R., & Jewell, M. (1993). Examining the validity of two measures for formative teaching: Reading aloud and maze. *Exceptional Children, 59*(5), 421-432.

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

Kranzler, J. H., Miller, M. D., & Jordan, L. (1999). An examination of racial/ethnic and gender bias on curriculum-based measurement of reading. *School Psychology Quarterly, 14*(3), 327-342.

Lenhard, W. & Schneider, W. (2006). *Ein Leseverständnistest für Erst- bis Sechstklässler* (ELFE 1-6) [A reading comprehension test for grade 1 to 6; ELFE 1-6]. Göttingen: Hogrefe.

Madelaine, A., & Wheldall, K. (2005). Identifying low-progress readers: Comparing teacher judgment with a curriculum-based measurement procedure. *International Journal of Disability, Development and Education, 52*(1), 33-42.

Mullis, I. V. S., Martin, M. O., Gonzalez, E. J., & Kennedy, A. M. (2003). *PIRLS 2001 international report: IEA's study of reading literacy achievement in primary school.* Chestnut Hill, MA: Boston College.

Palincsar, A. S., & Brown, A. L. (1984). Reciprocal teaching of comprehension-fostering and comprehension-monitoring activities. *Cognition and Instruction, 1*(2), 117-175.

Reschly, A. L., Busch, T. W., Betts, J., Deno, S. L., & Long, J. D. (2009). Curriculum-based measurement oral reading as an indicator of reading achievement: A meta-analysis of the correlational evidence. *Journal of School Psychology, 47*(6), 427-469.

Roick, T., Gölitz, D., & Hasselhorn, M. (2004). *Deutscher Mathematiktest für dritte Klassen (DEMAT 3+)* [German mathematics test for third graders]. Göttingen: Hogrefe.

Shin, J., Deno, S. L., & Espin, C. (2000). Technical adequacy of the maze task for curriculum-based measurement of reading growth. *The Journal of Special Education, 34*(3), 164-172.

Snow, R. (1977). Individual differences and instructional theory. *Educational Researcher, 6,* 11-15.

Stecker, P. M., Fuchs, L. S., & Fuchs, D. (2005). Using curriculum-based measurement to improve student achievement: Review of research. *Psychology in the Schools, 42*(8), 795-819.

Topping, K. (2006). Building reading fluency: Cognitive, behavioral, and socioemotional factors and the role of peer-mediated learning. In S. J. Samuels & A. E. Farstrup (Eds.), *What research has to say about fluency instruction* (pp. 106-129). Newark, DE: International Reading Association.

van Dijk, T. A., & Kintsch, W. (1983). *Strategies of discourse comprehension.* New York: Academic Press.

Wayman, M. M., Wallace, T., Wiley, H. I., Ticha, R., & Espin, C. A. (2007). Literature synthesis on curriculum-based measurement in reading. *The Journal of Special Education, 41*(2), 85-120.

Weiß, R. H. (2006). *Grundintelligenztest Skala 2, CFT 20-Revision* (CFT 20-R) [Culture fair intelligence scale; CFT 20-R]. Göttingen: Hogrefe.

Author Note

Please address correspondence to Natalie Förster, Department of Psychology in Education, University of Muenster, D-48149 Muenster, Germany; E-mail: natalie.foerster@uni-muenster.de.