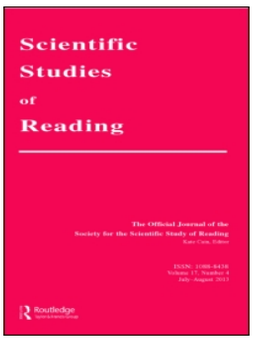

Scientific Studies of Reading

ISSN: 1088-8438 (Print) 1532-799X (Online) Journal homepage: http://www.tandfonline.com/loi/hssr20

What Does the CBM-Maze Test Measure?

Marloes M. L. Muijselaar, Panayiota Kendeou, Peter F. de Jong & Paul W. van den Broek

To cite this article: Marloes M. L. Muijselaar, Panayiota Kendeou, Peter F. de Jong & Paul W. van den Broek (2017) What Does the CBM-Maze Test Measure?, Scientific Studies of Reading, 21:2, 120-132, DOI: 10.1080/10888438.2016.1263994

To link to this article: http://dx.doi.org/10.1080/10888438.2016.1263994

SCIENTIFIC STUDIES OF READING
2017, VOL. 21, NO. 2, 120–132
http://dx.doi.org/10.1080/10888438.2016.1263994

Routledge
Taylor & Francis Group

What Does the CBM-Maze Test Measure?

Marloes M. L. Muijselaar[a], Panayiota Kendeou[b], Peter F. de Jong[a], and Paul W. van den Broek[c]

[a]University of Amsterdam; [b]University of Minnesota; [c]Leiden University

ABSTRACT

In this study, we identified the code-related (decoding, fluency) and language comprehension (vocabulary, listening comprehension) demands of the CBM-Maze test, a formative assessment, and compared them to those of the Gates–MacGinitie test, a standardized summative assessment. The demands of these reading comprehension tests and their developmental patterns were examined with multigroup structural regression models in a sample of 274 children in Grades 4, 7, and 9. The results showed that the CBM-Maze test relied more on code-related than on language comprehension skills when compared to the Gates–MacGinitie test. These demands were relatively stable across grades.

Reading comprehension is generally defined as a complex cognitive process (Kendeou, van den Broek, Helder, & Karlsson, 2014) that involves the construction of a mental representation of what a text is about, termed the *situation model* (Kintsch & van Dijk, 1978). The construction of this situation model is driven by at least two core processes: code-related processes that are responsible for the efficient recognition of words, and language comprehension processes that are responsible for the extraction of meaning from what is read (e.g., Snow, 2002; Verhoeven & Perfetti, 2008). These two sets of processes compose the influential simple view of reading (W. A. Hoover & Gough, 1990). It follows that measures of reading comprehension would tap on *both* code-related and language comprehension processes.

Reading comprehension is measured by both summative and formative assessments, each serving a different purpose (Scriven, 1967). On one hand, summative assessment is used to discern the state of achievement, which summarizes performance at a particular point in time. Various standardized reading comprehension tests are used for this purpose, with users likely assuming that different tests are roughly equivalent or interchangeable (Keenan, Betjemann, & Olson, 2008). However, accumulating evidence in the extant literature challenges this assumption (e.g., Keenan et al., 2014; Keenan & Meenan, 2014; Papadopoulos, Kendeou, & Shiakalli, 2014). For example, Nation and Snowling (1997) were the first to provide evidence that a score on the Suffolk Reading Scale was more strongly associated with code-related skill scores whereas a score on the Neale Analysis of Reading Ability was more strongly associated with language comprehension scores. Since this study, the demands of several other standardized tests have been examined (e.g., Andreassen & Bråten, 2010; Cutting & Scarborough, 2006; Keenan et al., 2008; Kendeou, Papadopoulos, & Spanoudis, 2012; Nation & Snowling, 1997). Specifically, the Woodcock–Johnson Passage Comprehension subtest is most strongly associated with code-related skills and less with language comprehension (Keenan et al., 2008), whereas the Gray Oral Reading Test and the Gates–MacGinitie Reading Comprehension Test associate more comparably with code-related and language comprehension skills (Cutting & Scarborough, 2006; Tilstra, McMaster, van den Broek, Kendeou, & Rapp, 2009). The results of these studies taken together demonstrate that summative, standardized reading comprehension tests differ in the extent to which they tap on code-related and language comprehension abilities.

CONTACT Marloes M. L. Muijselaar m.m.l.muijselaar@uva.nl Research Institute of Child Development and Education, University of Amsterdam, P.O. Box 15776, Amsterdam, NG NL-1001, The Netherlands.

On the other hand, formative assessment is used to discern the needs of a student with respect to instruction in an effort to improve achievement. One class of formative assessments—Curriculum-Based Measures (CBM)—are highly popular in the United States. Various versions of CBM measures have been adopted by schools in 30 U.S. states, including a statewide adoption in Iowa (92%), exceeding 5 million administrations annually (FastBridge Learning, 2015). CBM measures have been shown to be highly sensitive to growth over brief periods (see Espin, McMaster, Rose, & Wayman, 2012, for a review). Specifically, for the assessment of reading comprehension, the CBM-Maze test is established as a reliable and valid measure to assess progress (Espin, Wallace, Lembke, Campbell, & Long, 2010; Fuchs & Fuchs, 2002; Marcotte & Hintze, 2009; Pierce, McMaster, & Deno, 2010; Shin, Deno, & Espin, 2000; Tichá, Espin, & Wayman, 2009). The test has a standardized cloze format (Fuchs & Fuchs, 1992): Every seventh word is deleted and replaced with three multiple-choice alternatives—one correct and two incorrect words. This format is not strictly followed across different available CBM batteries. For example, in some instances the target word and distractors are selected purposely to increase test difficulty (Shinn & Shinn, 2002). The scoring procedures also vary across batteries, such that in some instances a ceiling number of errors is used or a correction for guessing. It is important to note, however, that when these variations were directly compared in different experimental studies, no major differences were observed in reliability or predictive validity of the CBM-Maze test scores (for a review, see Pierce et al., 2010).

Even though there is evidence that scores on the CBM-Maze test rely on code-related abilities (Gellert & Elbro, 2013; Kendeou & Papadopoulos, 2012; Kendeou et al., 2012; Pierce et al., 2010), there are no studies that examined its reliance on language comprehension. In the absence of this information, it is unclear whether the CBM-Maze test serves as a strong predictor of reading comprehension, despite its wide use in the United States. Thus, the purpose of the current study is to identify the relative code-related and language comprehension demands of the CBM-Maze test. Identifying these demands in isolation will make it challenging to validly interpret their magnitude. Thus, these demands were compared to those of the Gates–MacGinitie Reading Comprehension Test, a standardized test that relies on both code-related and language comprehension skills (Cutting & Scarborough, 2006). Further, for the selection of measures to assess code-related and language comprehension skills, we followed the normative paradigm in the extant literature. Specifically, the assessment of code-related skills often includes measures of reading accuracy and fluency (de Jong & van der Leij, 2002; Florit & Cain, 2011; Verhoeven & Perfetti, 2008). Both accuracy of word reading and automaticity and fluency are essential (Kirby & Savage, 2008), because when decoding is accurate but slow, the working memory demands on processing capacity often overcome and interfere with higher level reading comprehension processing (Perfetti & Hart, 2002). Thus, reading should be accurate and fast to be adequate to support higher level comprehension. The assessment of language comprehension skills often includes measures of vocabulary and listening comprehension (de Jong & van der Leij, 2002; Kim, 2016; van den Broek, Kendeou, & White, 2009).

Because we anticipated that the demands of reading comprehension would change developmentally (Adlof, Perfetti, & Catts, 2011; Afflerbach, Cho, & Kim, 2015), we sampled elementary, middle, and high school children. Specifically, a series of studies has shown that the relation between listening comprehension and reading comprehension increases over time, particularly in later elementary school, as the texts that are read and the questions asked about those texts become increasingly complex (e.g., Catts, Adolf, Hogan, & Weismer, 2005; Diakidoy, Stylianou, Karefillidou, & Papageorgiou, 2005; Vellutino, Tunmer, Jaccard, & Chen, 2007). Also, the relation between word reading accuracy and reading comprehension decreases over time (and may disappear by mid–elementary school), as most children have sufficient word reading skill to identify the words in written texts they encounter in reading comprehension tests (García & Cain, 2014; Vellutino et al., 2007). In contrast, the relation between reading fluency and reading comprehension increases, as most children begin to read not only with accuracy but also with speed (e.g., Language and Reading Research Consortium, 2015). Thus, in this study, the developmental patterns of children from elementary, middle, and high school were investigated.

Following previous work, we hypothesized that performance on the CBM-Maze test would depend more on code-related skills than performance on the Gates–MacGinitie test, whereas performance on the Gates–MacGinitie test would depend more on language comprehension skills than performance on the CBM-Maze test. With respect to the developmental patterns, we expected the code-related demands to decrease across grades (e.g., Vellutino et al., 2007), whereas the language comprehension demands we expected to increase (e.g., Catts et al., 2005). To that end, we tested whether the code-related and language comprehension demands were equal or different across grades.

Methods

Participants

Participants were 92 fourth graders, 90 seventh graders, and 92 ninth graders who were part of a larger study on reading comprehension. This study was carried out in suburban schools in a major metropolitan area in the midwestern region of the United States and aimed to create profiles of readers of different age and skill levels by using both online (eye-tracking and think-aloud) and offline (cognitive, linguistic, and achievement) measures. For the goals of this article, a subset of the administered linguistic measures was used to examine the demands of the CBM-Maze and the Gates–MacGinitie Reading Comprehension Tests.

Students who participated in this study were selected from four elementary school classes, two classes from a middle school, and three classes from a high school. Students receiving special or gifted education were excluded from the study. The sample was drawn from an upper middle-class, predominantly White population. There were several missing values on the demographic information variables; thus, descriptive statistics were computed only for the available data. In Grade 4, there were 49 boys and 43 girls, who were on average 9 years 8 months old (SD = 4.30). In Grade 7, there were 36 boys and 53 girls, who were on average 12 years 9 months old (SD = 3.79). In Grade 9, there were 42 boys and 50 girls, who were on average 14 years 8 months old (SD = 4.65). All students spoke English as their native language. The majority of the children were Caucasian (83%), with a few African American (6%), Asian (6%), Hispanic (2%), and other ethnicities (3%). The distribution of race was comparable for the different grade-level groups.

Measures

CBM-Maze test

The CBM-Maze test (Deno, 1985; Espin & Foegen, 1996) requires students to read passages that include incomplete sentences. From these passages, the first sentence is always intact, and after the first sentence, each seventh word is omitted. Students are required to choose the correct word to appropriately complete the sentence out of three options. Three written passages were presented to students, one at a time, in a booklet form. These passages were based on the curriculum for each grade level. Students had 1 min (fourth graders) or 2 min (seventh and ninth graders) to read as much of each passage as possible and circle the appropriate word to accurately complete the sentences. This same pattern was repeated for all three passages. The total time for test administration ranged from 5 to 10 min. A student's score consists of the average number of words selected correctly minus the number of words selected incorrectly across the three passages. The test–retest reliability is .83 (Shin et al., 2000).

Gates–MacGinitie test

The Gates–MacGinitie Reading Comprehension Test (MacGinitie, MacGinitie, Maria, & Dreyer, 2000) includes 11 passages and 48 multiple-choice questions related to these passages for Grades 4, 7, and 9. The questions require constructing an understanding based on information that is either explicitly or implicitly stated in the passage. Students independently read the passages and questions and then record

their answers on a separate answer sheet. Administration of the test takes approximately 45 min (10 min pretest; 35 min for actual test administration). A student's score consists of the total number of questions answered correctly. Test–retest reliability is .88 (MacGinitie et al., 2000).

Decoding

Students completed the Word Identification and the Word Attack subtests from the Woodcock–Johnson–III Achievement Test (WJ-III; McGrew & Woodcock, 2001). Together, they take about 10 min to administer. In the Word Identification task, students read aloud a list of words. The ceiling rule is set at six errors in a row. The procedure is the same for the Word Attack subtest, which includes pseudowords. A student's score is the average number of words and pseudowords read out correctly. The split-half reliability is .97 for the Word Identification subtest and .91 for the Word Attack subtest (Hosp & Fuchs, 2005).

Reading fluency

Students completed a CBM task to assess their oral reading fluency (Deno, 1985). In this task, students read aloud three separate age-appropriate passages for 1 min each. A student's score consists of the average number of words read correctly minus the average number of incorrectly read words (e.g., omissions, insertions, mispronunciations, substitutions, and hesitations of more than 3 s). Reliability ranged between .80 and .91 (Hintze & Silberglitt, 2005).

Listening comprehension

Students completed the Listening Comprehension subtest (H. D. Hoover, Heironymus, Frisbie, & Dunbar, 1996) from the Iowa Test of Basic Skills (ITBS). This test assesses literal meaning, inferential meaning, following directions, visual relations, numerical/spatial/temporal relations, and speaker point of view. There are 33 questions in the fourth-grade test, 38 questions in the seventh-grade test, and 40 questions in the ninth-grade test. Administration of the test takes approximately 45 min. A student's score consists of the total number of questions answered correctly. Reliabilities ranged from .67 to .79 (H. D. Hoover et al., 1996).

Vocabulary

Students completed the vocabulary subtest from the ITBS. The fourth graders completed Level 10, the seventh graders completed Level 13, and the ninth graders completed Level 14. In this test, students read sentences that have a word underlined. Under each sentence are four possible meanings or synonyms for the underlined word, and the student circles the item that has the closest meaning to the meaning of the underlined word. A student's score consists of the total number of questions answered correctly. The reliabilities ranged between .70 and .91 (Malecki & Elliott, 2002).

Procedure

The CBM-Maze test, Gates–MacGinitie Reading Comprehension Test, and the Listening Comprehension subtest from the ITBS were administered to participants at each grade level in two sessions on 2 days. The CBM-Maze and Gates–MacGinitie Reading Comprehension Tests were administered in the first session. The Listening Comprehension Test was administered in the second session. The rest of the assessments (CBM Oral Reading Fluency, WJ-III Word Identification Task, WJ-III Word Attack Task, and ITBS Vocabulary) were administered during a third session that was one-on-one with a research assistant. All test administrations followed a standardized procedure (dictated by each test).

Analyses

A multigroup structural regression model was fitted to the data to evaluate the code-related and language comprehension demands of the CBM-Maze and the Gates-MacGinitie Reading Comprehension Tests. The differences between the demands across tests were tested in each grade by examining whether regression coefficients could be constrained to be equal across the two reading comprehension tests or whether it was better to freely estimate those coefficients. When regression coefficients are constrained to be equal, reliance on each specific skill is hypothesized to be the same for both the CBM-Maze and the Gates–MacGinitie tests. When regression coefficients are freely estimated, then reliance on a specific skill is hypothesized to be different for the CBM-Maze and the Gates–MacGinitie tests. To test the hypothesized developmental patterns, that is, decreasing code-related demands and increasing language comprehension demands, we compared models with these coefficients constrained to be equal across grades, with models in which these coefficients were freely estimated.

However, comparing standardized regression coefficients or correlations is not possible in a regular structural regression model. In such models, analyses are based on the comparison of covariance matrices across groups. These covariances are composites of the correlations between variables and the variance of each variable. Thus, differences between covariances can reflect differences between correlations, between variances, or both. Differences between standardized regression coefficients or correlations can be tested only with a structural regression model in which the differences in variances are taken into account, which is possible with the use of *phantom factors*. Phantom factors are latent variables in which the variance is constrained (de Jong, 1999; Rodríguez, van den Boer, Jiménez, & de Jong, 2015; van den Boer, van Bergen, & de Jong, 2014). See Figure 1 for a graphical display of a structural regression model with phantom factors. In a phantom factor model, the regression coefficients are already standardized and the relations between variables are correlations instead of covariances. For the independent, or exogenous phantom factors (i.e., decoding, reading fluency, vocabulary, and listening comprehension), the observed variables had a loading on their corresponding latent phantom factor. The residual variances of the observed independent variables were fixed to zero, and the variances of the latent phantom factors were fixed to one (Rodríguez et al., 2015). For the dependent, or endogenous

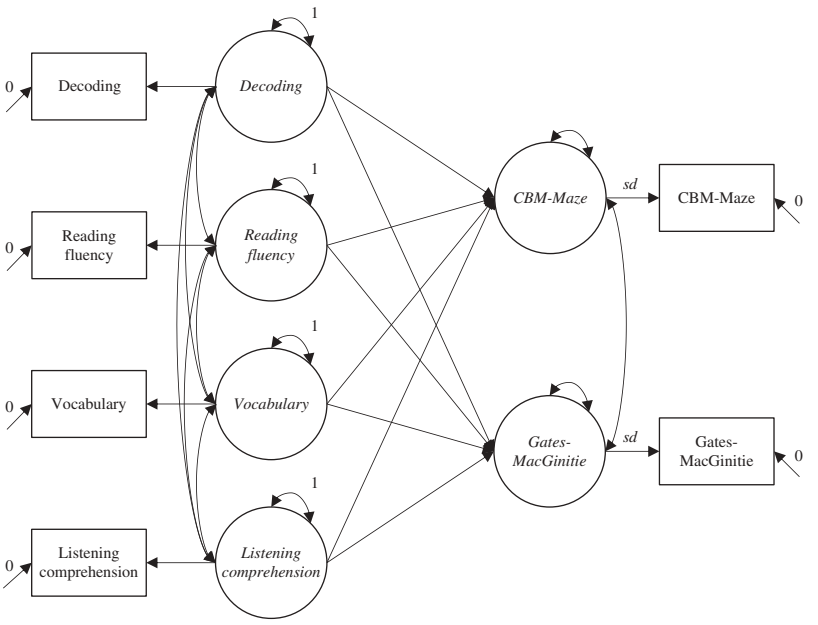

Figure 1. Multigroup structural regression model with phantom factors for the demands of decoding, reading fluency, vocabulary, and linguistic comprehension on the CBM-Maze and the Gates–MacGinitie Reading Comprehension Tests.

phantom, factors (i.e., both reading comprehension tests), the variances were freely estimated and the factor loadings of the observed dependent variables on their latent phantom factors were fixed to their standard deviations (van den Boer et al., 2014). The residual variances of the observed dependent variables were also fixed to zero.

Parameters of the multigroup structural regression model were estimated with Mplus Version 7.11 (Muthén & Muthén, 2012). To obtain parameter estimates, full information maximum likelihood estimation was used. Model fit was evaluated with the chi-square goodness-of-fit test statistic, the root mean square error of approximation, and the comparative fit index (Kline, 2011). A nonsignificant chi-square value ($p > .05$) indicated exact model fit (Hayduck, 1996). Root mean square error of approximation values less than .05 indicated good approximate fit, values between .05 and .08 were taken as satisfactory fit, and values greater than .10 were considered a poor fit (Browne & Cudeck, 1993). Comparative fit index values greater than .90 were considered acceptable, and values greater than .95 were taken as good incremental model fit (Hu & Bentler, 1999). Differences between the fit of any two nested models were tested with the chi-square difference test (Kline, 2011).

Results

Data screening and descriptive statistics

Before conducting the analyses, the data were checked for outliers. Scores that were more than 3 standard deviations above or below the mean were omitted. In total, less than 1% of the data were omitted. Means and standard deviations of reading comprehension, decoding, reading fluency, vocabulary, and listening comprehension are displayed in Table 1. Correlations between all variables are displayed in Table 2. The correlations show that scores on the CBM-Maze and Gates–MacGinitie tests were highly related (correlations ranging from .69 to .79). Also, the CBM-Maze test related to both code-related (correlations ranging from .50 to .88) and language comprehension skills (correlations ranging from .49 to .65). A similar pattern was observed for the Gates–MacGinitie test (correlations with code-related skills ranged from .47 to .80; correlations with language comprehension skills ranged from .55 to .76). A comparison of these relations suggests that the two tests relate to both code-related and language comprehension skills. In addition, the relations of both reading comprehension tests with decoding decreased across grades, whereas the relations with reading fluency remained relatively stable. The relations of the tests with vocabulary and listening comprehension did not show any clear developmental pattern.

Differences between the demands of the CBM-Maze and the Gates-MacGinitie Reading Comprehension tests

A multigroup structural regression model with phantom factors was fitted to the data to evaluate and compare the unique demands of the CBM-Maze and the Gates–MacGinitie tests. In this model (Figure 1), phantom factors were specified for each construct, with the observed variables as single indicators. This model was a just-identified or saturated model, that is, the model had zero degrees

Table 1. Descriptive statistics for reading comprehension, decoding, reading fluency, vocabulary, and listening comprehension.

	Grade 4				Grade 7				Grade 9			
	N	Max	*M*	*SD*	*N*	Max	*M*	*SD*	*N*	Max	*M*	*SD*
CBM-Maze	92	16.33	6.74	4.10	90	36.33	18.32	8.27	92	40.33	21.50	7.88
Gates–MacGinitie	92	48.00	27.47	11.34	90	48.00	33.52	9.24	92	48.00	38.71	6.54
Decoding	91	51.50	36.38	6.86	87	52.00	46.26	3.62	91	53.50	47.82	3.60
Reading fluency	92	227.33	119.59	46.86	89	293.00	187.21	39.80	91	238.00	153.75	33.94
Vocabulary	92	11.00	6.53	2.97	88	14.00	9.59	2.83	92	14.00	10.57	2.78
Listening comprehension	92	28.00	18.79	3.82	90	35.00	25.63	5.09	92	38.00	26.83	5.08

Table 2. Correlations among reading comprehension, decoding, reading fluency, vocabulary, and listening comprehension in Grades 4, 7, and 9.

	1	2	3	4	5
1. CBM-Maze	—				
2. Gates–MacGinitie	.79 /.69 /.71	—			
3. Decoding	.77 /.61 /.50	.74 /.47 /.54	—		
4. Reading fluency	.88 /.78 /.81	.80 /.62 /.66	.82 /.60 /.58	—	
5. Vocabulary	.65 /.63 /.57	.76 /.64 /.72	.70 /.60 /.60	.69 /.56 /.53	—
6. Listening comprehension	.49 /.61 /.52	.65 /.67 /.55	.46 /.33 /.43	.50 /.47 /.45	.57 /.65 /.60

Note. Values are represented as Grade 4/Grade 7/Grade 9. All correlations are significant, $p < .01$.

Table 3. Standardized regression coefficients of the structural regression model with phantom factors of the decoding, reading fluency, vocabulary, and listening comprehension demands of the CBM-Maze (Maze) and the Gates–MacGinitie Reading Comprehension (Gates) tests.

	Grade 4		Grade 7		Grade 9	
	Maze	Gates	Maze	Gates	Maze	Gates
Decoding	.13	.12	.16†	.04	−.06	−.00
Reading fluency	.73**	.39**	.52**	.30**	.72**	.37**
Vocabulary	.03	.27**	.08	.20†	.14†	.46**
Listening comprehension	.04	.25**	.26**	.38**	.12	.10

$^{\dagger}p < .10$. *$p < .05$. **$p < .01$.

of freedom. The regression coefficients of the structural part of this just-identified model are presented in Table 3. Differences between the demands of the two comprehension tests were examined by constraining regression coefficients of each of the four factors of the different demands to be equal across the two tests. Note that in these models the regression coefficients were constrained to be equal in each grade but not across grades. Step-by-step model comparisons were used to test whether the regression coefficients in each grade should be constrained to be equal or freely estimated.

Decoding and fluency demands

With respect to decoding and fluency, two separate models were estimated with each of these demands constrained to be equal across tests. It was hypothesized that the CBM-Maze test would rely more heavily on decoding and reading fluency than the Gates–MacGinitie test. The model in which the regression coefficients of decoding on both reading comprehension tests were constrained to be equal within each grade fitted the data well (see Table 4, Model 1.1). This model could not be further improved by freely estimating regression coefficients of decoding on reading comprehension. With respect to reading fluency, the model in which this demand on both reading comprehension tests was constrained to be equal had a poor fit to the data (see Table 4, Model 2.1). Step-by-step model testing suggested that the equality constraints had to be dropped in all grades (see Table 4, difference between Model 2.1 and 2.2). In sum, these model comparisons showed that the decoding demands of the CBM-Maze and the Gates–MacGinitie tests were equal in all grades, but the CBM-Maze test relied significantly more on reading fluency than the Gates–MacGinitie test.

Vocabulary and listening comprehension demands

The unique demands of the reading comprehension tests with respect to vocabulary and listening comprehension were also examined. It was hypothesized that the Gates–MacGinitie test would rely more heavily on vocabulary and listening comprehension skills than the CBM-Maze test. The model with equality constraints on the vocabulary demands of both reading comprehension tests in all grades had a poor fit to the data (see Table 4, Model 3.1). Step-by-step model comparisons suggested that freely estimating the regression coefficients in Grades 4 and 9 for the vocabulary demands resulted in improvements of the fit of the models (see Table 4, Model 3.1 vs. Model 3.2), whereas removing the

Table 4. Values of the selected fit indices for the models concerning the differences in the demands of the CBM-Maze and the Gates–MacGinitie tests.

Model	Code-Related and Language Comprehension Demands	Constraints Across Tests in Each Grade	χ^2	*df*	RMSEA	90% CI	CFI	Model Comparisons	$\Delta\chi^2$	*df*
1.1	Decoding	4, 7, 9	1.35	3	.000	[.000, .128]	1.00			
2.1	Reading fluency	4, 7, 9	26.31**	3	.292	[.196, .399]	.98			
2.2	Reading fluency	—	0	0	.000	—	1.00	2.1 vs. 2.2	26.31*	3
3.1	Vocabulary	4, 7,9	16.70**	3	.224	[.127, .333]	.99			
3.2	Vocabulary	7	0.93	1	.000	[.000, .273]	1.00	3.1 vs. 3.2	15.77*	2
4.1	Listening comprehension	4, 7, 9	8.92*	3	.147	[.040, .263]	1.00			
4.2	Listening comprehension	7, 9	1.24	2	.000	[.000, .180]	1.00	4.1 vs. 4.2	7.68*	1

Note. RMSEA = root mean square error of approximation; CI = confidence interval; CFI = comparative fit index. *$p < .05$. **$p < .01$.

equality constraint in Grade 7 did not result in improvement in model fit. The model with equality constraints in Grade 7 only had a good fit to the data (see Table 4, Model 3.2). For listening comprehension, the model with equality constraints on this demand of the two reading comprehension tests in all grades had a poor fit to the data (see Table 4, Model 4.1). Step-by-step model comparisons suggested that the fully constrained model could be improved only by removing the equality constraint in Grade 4 (see Table 4, Model 4.1 vs. 4.2). This model with equality constraints in Grades 7 and 9 had a good fit to the data (see Table 4, Model 4.2). In sum, these step-by-step model comparisons showed that the Gates–MacGinitie test relied more heavily on vocabulary in Grades 4 and 9 than the CBM-Maze test. The reliance on vocabulary of both tests was not significantly different in Grade 7; however, the pattern was in the expected direction. With respect to listening comprehension, the Gates–MacGinitie test relied more heavily on this demand than the CBM-Maze test in Grade 4. In Grades 7 and 9, the reliance on listening comprehension of both tests was not significantly different.

Differences in the demands of the CBM-Maze and the Gates-MacGinitie tests across grades

The multigroup structural regression model with phantom factors was also used to investigate the demands of the CBM-Maze and the Gates–MacGinitie tests across grades. Because there is evidence that the demands of reading comprehension change developmentally, we added model constraints to the model in Figure 1, separately for each of the four factors. We hypothesized that the decoding and fluency demands of reading comprehension tests would decrease, whereas the vocabulary and listening comprehension demands would increase across grades. We tested whether equally constraining or freely estimating the regression coefficients resulted in the best-fitting models.

Decoding and fluency demands

A model with an equality constraint on the decoding demands of both tests across grades had a good fit to the data (see Table 5, Model 1.1) and could not be further improved. With respect to fluency, a model with equality constraints of both tests across grades also had a good fit to the data (see Table 5, Model 2.1) and could not be further improved. To sum up, these model comparisons revealed that the decoding and fluency demands of both reading comprehension tests remain equal across grades.

Vocabulary and listening comprehension demands

Changes in the vocabulary and listening comprehension demands of reading comprehension across grades were also examined. A model with equality constraints on the vocabulary demands across grades had a good fit to the data (see Table 5, Model 3.1) and could not be further improved. For listening comprehension, a model with equality constraints had a poor fit to the data (see Table 5, Model 4.1). This model could be improved by removing the equality constraint of the Gates–MacGinitie test in Grade 9 (see Table 5, difference between Model 4.1 and 4.2). This model had

Table 5. Values of the selected fit indices for the models concerning the developmental patterns of the demands of the CBM-Maze (Maze) and the Gates–MacGinitie (Gates) tests.

Model	Code-Related and Language Comprehension Demands	Constraints Across grades in Each Test: Maze	Constraints Across grades in Each Test: Gates	χ^2	*df*	RMSEA	90% CI	CFI	Model Comparisons	$\Delta\chi^2$	*df*
1.1	Decoding	4, 7, 9	4, 7, 9	4.61	4	.041	[.000, .168]	1.00			
2.1	Reading fluency	4, 7, 9	4, 7, 9	3.46	4	.000	[.000, .148]	1.00			
3.1	Vocabulary	4, 7, 9	4, 7, 9	4.26	4	.027	[.000, .162]	1.00			
4.1	Listening comprehension	4, 7, 9	4, 7, 9	10.11*	4	.129	[.027, .231]	1.00			
4.2	Listening comprehension	4, 7, 9	4, 7	5.66	3	.099	[.000, .222]	1.00	4.1 vs. 4.2	4.45*	1

Note. RMSEA = root mean square error of approximation; CI = confidence interval; CFI = comparative fit index. *$p < .05$.

an acceptable fit to the data (see Table 5, Model 4.2) and could not be further improved. In sum, the model comparisons revealed that the vocabulary demands of both reading comprehension tests and the listening comprehension demands of the CBM-Maze test are equal across grades. The listening comprehension demands of the Gates–MacGinitie test, however, decrease in Grade 9.

Discussion

The purpose of the current study was to identify the relative code-related and language comprehension demands of the CBM-Maze test in elementary, middle, and high school children. These demands were compared to those of the Gates–MacGinitie Reading Comprehension Test. Even though the CBM-Maze and the Gates–MacGinitie tests substantially correlated among them and to both code-related and language comprehension skills, the results of the multigroup structural regression modeling, in which unique effects are considered, showed that the CBM-Maze test depends more heavily on code-related skills than does the Gates–MacGinitie test, whereas the Gates–MacGinitie test relies more heavily on language comprehension skills than does the CBM-Maze test. With respect to developmental patterns, the code-related and language comprehension demands of both tests remain relatively stable across grades.

These results highlight that the CBM-Maze test relies more on code-related skills than on language comprehension skills. Indeed, the test comparisons revealed that even though the decoding demands of the CBM-Maze and the Gates–MacGinitie tests are equal in all grades, the CBM-Maze test relies significantly more on reading fluency and less on language comprehension than the Gates–MacGinitie test. These findings are interesting considering the high correlation between the two tests across grades. Further, the findings also raise an important question, namely, whether the wide use of the CBM-Maze test, when used as a formative assessment to influence instructional focus (vs. more simply using as an efficient predictor), can adequately evaluate the strengths and weaknesses of students in reading comprehension. Previous studies also suggested that such *cloze* tests measure primarily lower level comprehension processes (e.g., sentence comprehension; e.g., Gellert & Elbro, 2013).

Further, the comparisons across grades showed that the demands on decoding and fluency remained relatively stable for both tests. Note that at each grade level, students completed grade-level decoding items and read grade-level texts (in CBM-Maze, CBM–Oral Reading, and in Gates–MacGinitie). Thus, "a stable" pattern across grades suggests that both decoding and fluency continue to covary with reading comprehension. Further, the unique contribution of decoding was hardly significant, whereas the role of fluency was substantial, consistent with recent findings showing that the contribution of decoding in reading comprehension is gradually taken over by reading fluency (LARRC, 2015). The small contribution of decoding can be attributed, in part, to suppression effects due to the high correlation of decoding and fluency. Note, however, that the model comparisons produce the same results when decoding and fluency are entered in separate models. A more likely

explanation is that the measure of fluency used in this study (i.e., oral reading fluency) draws on skills beyond those captured by word identification efficiency alone (Eason, Sabatini, Goldberg, Bruce, & Cutting, 2013).

The comparisons across grades also showed that the demands on vocabulary remained stable for both tests, suggesting that vocabulary continues to covary with reading comprehension. The demands of listening comprehension also remained stable across grades even though they were relatively low for both tests (see Table 3), but they decreased in Grade 9 for the Gates-MacGinitie test. One potential explanation is the measurement of listening comprehension itself, as the ITBS Listening Comprehension Test taps on a broad set of skills. Specifically, in addition to comprehension of literal and inferential meaning, which is often the focus of various listening comprehension tests (Hogan, Adlof, & Alonzo, 2014), the test specifications include comprehension of numerical/spatial/temporal relations, following directions, visual relations, and speaker point of view. It is unclear whether the text and questions included in the Gates–MacGinitie test pose consistent demands on this broad set of skills across grades. Another potential explanation is that the presence of both vocabulary and listening comprehension in the model may have resulted in suppression effects. Note, however, that the model comparisons produce the same results when vocabulary and listening comprehension are entered in separate models. A final explanation is that the nature of the demands on comprehension changes across development. Specifically, comprehension demands in higher grades (such as Grade 9 in our study) may not be adequately accounted for by listening comprehension alone; for example, question and text complexity in higher grades (e.g., content, structure) may pose increasing demands on 21st-century higher order skills such as reasoning (Goldman, 2012; Goldman & Pellegrino, 2015; Graesser, 2015; Sabatini, O'Reilly, Halderman, & Bruce, 2014). As previously noted, it is unclear whether the text and questions included in the Gates–MacGinitie test pose such demands, thus further research is needed to examine this issue.

The current set of findings must be considered in the context of certain design and measurement constraints of the present study. One limitation derives from the use of different students in each grade level. Therefore, it is not clear whether the developmental patterns observed are "developmental" or due to cohort differences. Despite the advantages of the cross-sectional design that allowed direct comparisons of elementary, middle, and high school students at a single time point, a longitudinal design would have been more appropriate to establish developmental patterns. Further, the actual interpretation of the comparisons across grades is also limited by the use of a mix of standardized, state, and CBM tests. For example, both the fluency and the Maze tests were CBM measures that used grade-level texts at each grade, and their high correlation likely also reflects common method variance. Thus, the correlation between fluency and CBM-Maze may be difficult to interpret when compared to that of fluency with the Gates–MacGinitie test. Further, this study used one specific variation of the CBM-Maze test. As we discussed earlier, the test has several different variations; even though these variations may have limited, if any, implications for reliability and predictive validity (Pierce et al., 2010), they *do* have implications for the specific demands posed by the test. A final issue pertains to constraints to generalizability as a function of the sample demographics. It would be important to address this issue also in future work, drawing from a more diverse population.

Despite these limitations, the findings from this study taken together have important implications for both researchers and educators. One implication pertains to the diagnosis of students at risk of reading difficulties. Specifically, the type of reading comprehension test that is used determines to a large extent which students are diagnosed as at-risk or struggling readers (Keenan & Meenan, 2014; Papadopoulos et al., 2014). For example, in a sample of 1,500 participants who ranged from 8 to 19 years of age, only half of the individuals who performed poorly on a reading comprehension test that mainly relied on comprehension also performed poorly on a reading comprehension test that mainly relied on decoding (Keenan et al., 2014). It follows that if the CBM-Maze test is used to identify strengths and weaknesses in readers (in reading research or at schools), it will likely result in the identification of struggling readers who experience difficulties with lower level comprehension abilities. A second implication concerns potential revisions that

could further improve the CBM-Maze test so that it fares better when compared to "balanced" standardized reading comprehension measures as a tool for measuring comprehension skills per se. The CBM-Maze test is a useful measure and has many advantages over traditional standardized test measures; it is reliable, fast and easy to administer, and inexpensive (Fuchs & Fuchs, 2002; Pierce et al., 2010). Future work can focus on how to revise the test so that its language comprehension demands are increased. For example, existing test variations that purposely select the target word and distractors are a few ways that can increase comprehension demands (Gellert & Elbro, 2013). Note, however, that several attempts for revising the CBM-Maze test have been made already, but these have not led to significant improvements in capturing *deep* comprehension (Lembke et al., 2016). Thus, more work is needed in this direction.

In conclusion, the current study revealed that the CBM-Maze test relies more on code-related skills than on language comprehension skills. Comparisons between the demands of the CBM-Maze and the Gates–MacGinitie tests also revealed that the CBM-Maze test relies more heavily on reading fluency and less on language comprehension skills than the Gates–MacGinitie test. This study has important implications for the use of the CBM-Maze as a formative assessment measure and suggests that it should be further revised to increase its demands on language comprehension skills.

Acknowledgments

We thank the students, teachers, and administrators who so graciously participated in or facilitated this research. We also thank Christine Espin and John Sabatini for reading and commenting on this article.

Funding

The research reported in this article was partly supported by a grant from the Institute of Education Sciences (R305G040021) to the University at Minnesota (PI: Paul van den Broek). The opinions expressed are those of the authors and do not represent the view of the funding agency.

References

Adlof, S. M., Perfetti, C. A., & Catts, H. W. (2011). Developmental changes in reading comprehension: Implications for assessment and instruction. In S. J. Samuels & A. E. Farstrup (Eds.), *What research has to say about reading instruction* (pp. 186–214). Newark, DE: International Reading Association.

Afflerbach, P., Cho, B. Y., & Kim, J. Y. (2015). Conceptualizing and assessing higher-order thinking in reading. *Theory Into Practice*, *54*(3), 203–212. doi:10.1080/00405841.2015.1044367

Andreassen, R., & Bråten, I. (2010). Examining the prediction of reading comprehension on different multiple-choice tests. *Journal of Research in Reading*, *33*, 263–283. doi:10.1111/j.1467-9817.2009.01413.x

Browne, M. W., & Cudeck, R. (1993). Alternative ways of assessing model fit. In K. A. Bollen & J. S. Long (Eds.), *Testing structural equation models* (pp. 136–162). Newbury Park, CA: Sage.

Catts, H., Adolf, S. M., Hogan, T. P., & Weismer, S. (2005). Are specific language impairment and dyslexia distinct disorders? *Journal of Speech, Language and Hearing Research*, *48*, 1378–1396. doi:10.1044/1092-4388(2005/096)

Cutting, L. E., & 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. *Scientific Studies of Reading*, *10*, 277–299. doi:10.1207/s1532799xssr1003_5

de Jong, P. F. (1999). Hierarchical regression analysis in structural equation modeling. *Structural Equation Modeling*, *6*, 198–211. doi:10.1080/10705519909540128

de Jong, P. F., & van der Leij, A. (2002). Effects of phonological abilities and listening comprehension on the development of reading. *Scientific Studies of Reading*, *6*, 51–77. doi:10.1207/S1532799XSSR0601_03

Deno, S. L. (1985). Curriculum-based measurement: The emerging alternative. *Exceptional Children*, *16*, 99–104. doi:10.1177/001440298505200303

Diakidoy, I. N., Stylianou, P., Karefillidou, C., & Papageorgiou, P. (2005). The relationship between listening and reading comprehension of different types of text at increasing grade levels. *Reading Psychology*, *26*, 55–80. doi:10.1080/02702710590910584

Eason, S. H., Sabatini, J., Goldberg, L., Bruce, K., & Cutting, L. E. (2013). Examining the relationship between word reading efficiency and oral reading rate in predicting comprehension among different types of readers. *Scientific Studies of Reading*, *17*, 199–223. doi:10.1080/10888438.2011.652722

Espin, C. A., & Foegen, A. (1996). Validity of general outcome measures for predicting secondary students' performance on content-area tasks. *Exceptional Children*, *62*, 497–514. doi:10.1177/001440299606200602
Espin, C. A., McMaster, K., Rose, S., & Wayman, M. (Eds.). (2012). *A measure of success: The influence of curriculum-based measurement on education*. Minneapolis: University of Minnesota Press.
Espin, C. A., Wallace, T., Lembke, E., Campbell, H., & Long, J. D. (2010). Creating a progress measurement system in reading for middle-school students: Monitoring progress towards meeting high stakes standards. *Learning Disabilities Research and Practice*, *25*, 60–75. doi:10.1111/j.1540-5826.2010.00304.x
FastBridge Learning. (2015). *Formative assessment system for teachers (FAST): Technical manual*. Minneapolis, MN: Author.
Florit, E., & Cain, K. (2011). The simple view of reading: Is it valid for different types of alphabetic orthographies? *Educational Psychology Review*, *23*, 553–576. doi:10.1007/s10648-011-9175-6
Fuchs, L. S., & Fuchs, D. (1992). Identifying a measure for monitoring student reading progress. *School Psychology Review*, *21*(1), 45–59.
Fuchs, L. S., & Fuchs, D. (2002). Curriculum-based measurement: Describing competence, enhancing outcomes, evaluating treatment effects, and identifying treatment nonresponders. *Peabody Journal of Education*, *77*, 64–84. doi:10.1207/S15327930PJE7702_6
García, J. R., & Cain, K. (2014). Decoding and reading comprehension: A meta-analysis to identify which reader and assessment characteristics influence the strength of the relationship in English. *Review of Educational Research*, *84*, 74–111. doi:10.3102/0034654313499616
Gellert, A. S., & Elbro, C. (2013). Cloze tests may be quick, but are they dirty? Development and preliminary validation of a cloze test of reading comprehension. *Journal of Psychoeducational Assessment*, *31*, 16–28. doi:10.1177/0734282912451971
Goldman, S. R. (2012). Literacy: Learning and understanding content. *The Future of Children*, *22*(2), 89–116. doi:10.1353/foc.2012.0011
Goldman, S. R., & Pellegrino, J. W. (2015). Research on learning and instruction: Implications for curriculum, instruction, and assessment. *Education*, *2*, 33–41. doi:10.1177/2372732215601866
Graesser, A. C. (2015). Deeper learning with advances in discourse science and technology. *Policy Insights from the Behavioral and Brain Sciences*, *2*, 42–50. doi:10.1177/2372732215600888
Hayduck, L. A. (1996). *LISREL issues, debates, and strategies*. Baltimore, MD: Johns Hopkins University Press.
Hintze, J. M., & Silberglitt, B. (2005). A longitudinal examination of the diagnostic accuracy and predictive validity of R-CBM an high-stakes testing. *School Psychology Review*, *34*(3), 372–386.
Hogan, T. P., Adlof, S. M., & Alonzo, C. N. (2014). On the importance of listening comprehension. *International Journal of Speech-Language Pathology*, *16*, 199–207. doi:10.3109/17549507.2014.904441
Hoover, H. D., Heironymus, A. N., Frisbie, D. A., & Dunbar, S. B. (1996). *Iowa test of basic skills*. Itasca, IL: Riverside.
Hoover, W. A., & Gough, P. B. (1990). The simple view of reading. *Reading and Writing*, *2*, 127–160. doi:10.1007/BF00401799
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.
Hu, L., & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling*, *6*, 1–55. doi:10.1080/10705519909540118
Keenan, J. M., Betjemann, R. S., & Olson, R. K. (2008). Reading comprehension tests vary in the skills they assess: Differential dependence on decoding and oral comprehension. *Scientific Studies of Reading*, *12*, 281–300. doi:10.1080/10888430802132279
Keenan, J. M., Hua, A. N., Meenan, C. E., Pennington, B. F., Willcutt, E., & Olson, R. K. (2014). Issues in identifying poor comprehenders. *L'Année Psychologique*, *114*, 753–777. doi:10.4074/S0003503314004072
Keenan, J. M., & Meenan, C. E. (2014). Test differences in diagnosing reading comprehension deficits. *Journal of Learning Disabilities*, *47*, 125–135. doi:10.1177/0022219412439326
Kendeou, P., & Papadopoulos, T. (2012). The use of CBM-Maze in Greek: A closer look at what it measures. In C. Espin, K. McMaster, & S. Rose (Eds.), *A measure of success: The influence of curriculum-based measurement on education*. Minneapolis: University of Minnesota Press.
Kendeou, P., Papadopoulos, T. C., & Spanoudis, G. (2012). Processing demands of reading comprehension tests in young readers. *Learning and Instruction*, *22*, 354–367. doi:10.1016/j.learninstruc.2012.02.001
Kendeou, P., van den Broek, P., Helder, A., & Karlsson, A. K. (2014). A cognitive view of reading comprehension: Implications for reading difficulties. *Learning Disabilities Research and Practice*, *29*, 10–16. doi:10.1111/ldrp.12025
Kim, Y. G. (2016). Direct and mediated effects of language and cognitive skills on comprehension of oral narrative texts (listening comprehension) for children. *Journal of Experimental Child Psychology*, *141*, 101–120. doi:10.1016/j.jecp.2015.08.003
Kintsch, W., & van Dijk, T. A. (1978). Toward a model of text comprehension and production. *Psychological Review*, *85*(1), 363–394. doi:10.1037/0033-295X.85.5.363
Kirby, J. R., & Savage, R. S. (2008). Can the simple view deal with the complexities of reading? *Literacy*, *42*(2), 75–82. doi:10.1111/read.2008.42.issue-2

Kline, R. B. (2011). *Principles and practice of structural equation modeling* (3rd ed.). New York, NY: Guilford Press.
Language and Reading Research Consortium. (2015). Learning to read: Should we keep things simple? *Reading Research Quarterly*, *50*, 151–169. doi:10.1002/rrq.99
Lembke, E., Conoyer, S., Hosp, J., Espin, C. A., Hosp, M., & Poch, A. (2016). Getting more from your maze: Examining differences in distractors. *Reading and Writing Quarterly: Overcoming Learning Difficulties*. Advance online publication. http://dx.doi.org/10.1080/10573569.2016.1142913
MacGinitie, W. H., MacGinitie, R. K., Maria, K., & Dreyer, L. G. (2000). *Gates–MacGinitie reading tests* (4th ed.). Itasca, IL: Riverside.
Malecki, C. K., & Elliott, S. N. (2002). Children's social behaviors as predictors of academic achievement: A longitudinal analysis. *School Psychology Quarterly*, *17*(1), 1–23. doi:10.1521/scpq.17.1.1.19902
Marcotte, A. M., & Hintze, J. M. (2009). Incremental and predictive utility of formative assessment methods of reading comprehension. *Journal of School Psychology*, *27*, 315–335. doi:10.1016/j.jsp.2009.04.003
McGrew, K. S., & Woodcock, R. W. (2001). *Technical manual. Woodcock–Johnson III*. Itasca, IL: Riverside.
Muthén, L. K., & Muthén, B. O. (2012). *Mplus user's guide* (7th ed.). Los Angeles, CA: Muthén & Muthén.
Nation, K., & Snowling, M. J. (1997). Assessing reading difficulties: The validity and utility of current measures of reading skill. *British Journal of Educational Psychology*, *67*, 359–370. doi:10.1111/j.2044-8279.1997.tb01250.x
Papadopoulos, T. C., Kendeou, P., & Shiakalli, M. (2014). Reading comprehension tests and poor readers: How test processing demands result in different profiles. *L'Année Psychologique*, *114*, 726–752. doi:10.4074/S0003503314004060
Perfetti, C. A., & Hart, L. (2002). The lexical quality hypothesis. In L. Verhoeven, C. Elbro, & P. Reitsma (Eds.), *Precursors of functional literacy*. Amsterdam, The Netherlands: John Benjamins.
Pierce, R. L., McMaster, K. L., & Deno, S. L. (2010). The effects of using different procedures to score maze measures. *Learning Disabilities Research and Practice*, *25*, 151–160. doi:10.1111/j.1540-5826.2010.00313.x
Rodríguez, C., van den Boer, M., Jiménez, J. E., & de Jong, P. F. (2015). Developmental changes in the relations between RAN, phonological awareness, and reading in Spanish children. *Scientific Studies of Reading*, *19*, 273–288. doi:10.1080/10888438.2015.1025271
Sabatini, J. P., O'Reilly, T., Halderman, L. K., & Bruce, K. (2014). Integrating scenario-based and component reading skill measures to understand the reading behavior of struggling readers. *Learning Disabilities Research and Practice*, *29*, 36–43. doi:10.1111/ldrp.12028
Scriven, M. (1967). The methodology of evaluation. In R. W. Tyler, R. M. Gagne, & M. Scriven (Eds.), *Perspectives of curriculum evaluation* (Vol. I, pp. 39–83). Chicago, IL: Rand McNally.
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*, 164–172. doi:10.1177/002246690003400305
Shinn, M. R., & Shinn, M. M. (2002). *AIMSweb training workbook: Administration and scoring of reading maze for use in general outcome measurement*. Eden Prairie, MN: Edformation.
Snow, C. E. (2002). *Reading for understanding: Toward an R&D program in reading comprehension*. Santa Monica, CA: RAND.
Tichá, R., Espin, C. A., & Wayman, M. M. (2009). Reading progress monitoring for secondary-school students: Reliability, validity, and sensitivity to growth of reading aloud and maze selection measures. *Learning Disabilities Research and Practice*, *24*, 132–142. doi:10.1111/j.1540-5826.2009.00287.x
Tilstra, J., McMaster, K., van den Broek, P., Kendeou, P., & Rapp, D. (2009). Simple but complex: Components of the simple view of reading across grade levels. *Journal of Research in Reading*, *32*, 383–401. doi:10.1111/j.1467-9817.2009.01401.x
van den Boer, M., van Bergen, E., & de Jong, P. F. (2014). Underlying skills of oral and silent reading. *Journal of Experimental Child Psychology*, *128*, 138–151. doi:10.1016/j.jecp.2014.07.008
van den Broek, P., Kendeou, P., & White, M. J. (2009). Cognitive processes during reading: Implications for the use of multimedia to foster reading comprehension. In A. G. Bus & S. B. Neuman (Eds.), *Multimedia and literacy development: Improving achievement for young learners* (pp. 57–73). New York, NY: Rutledge.
Vellutino, F. R., Tunmer, W. E., Jaccard, J. J., & Chen, R. S. (2007). Components of reading ability: Multivariate evidence for a convergent skills model of reading development. *Scientific Studies of Reading*, *11*, 3–32. doi:10.1080/10888430709336632
Verhoeven, L., & Perfetti, C. (2008). Advances in text comprehension: Model, process and development. *Applied Cognitive Psychology*, *22*, 293–301. doi:10.1002/acp.1417