How to assess and improve children's reading comprehension?

Muijselaar, M.M.L.

Citation for published version (APA):

General rights
It is not permitted to download or to forward/distribute the text or part of it without the consent of the author(s) and/or copyright holder(s), other than for strictly personal, individual use, unless the work is under an open content license (like Creative Commons).

Disclaimer/Complaints regulations
If you believe that digital publication of certain material infringes any of your rights or (privacy) interests, please let the Library know, stating your reasons. In case of a legitimate complaint, the Library will make the material inaccessible and/or remove it from the website. Please Ask the Library: https://uba.uva.nl/en/contact, or a letter to: Library of the University of Amsterdam, Secretariat, Singel 425, 1012 WP Amsterdam, The Netherlands. You will be contacted as soon as possible.

UvA-DARE is a service provided by the library of the University of Amsterdam (http://dare.uva.nl)

Download date: 04 Aug 2019
CHAPTER 3

WHAT DOES THE CBM-MAZE TEST MEASURE?
ABSTRACT

In the current study, we identified the code-related and language comprehension demands of the CBM-Maze test and compared them to those of the Gates-MacGinitie test, a standardized reading comprehension measure. The demands of these reading comprehension tests and their developmental patterns were examined with multigroup structural regression models in a sample of 274 children. The results showed that even though the CBM-Maze test relied on both code-related and language comprehension skills, when compared to the Gates-MacGinitie test, it posed significantly higher code-related and lower language comprehension demands. The developmental patterns were in line with those reported in previous studies.

WHAT DOES THE CBM-MAZE TEST MEASURE?

INTRODUCTION

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, also known as a 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. These two sets of processes, code-related and language comprehension, comprise the influential simple view of reading (Hoover & Gough, 1990). It follows that good measures of reading comprehension should tap on both code-related and on language comprehension processes.

Reading comprehension is often measured by asking students to read texts and to recall what they have read or to answer multiple-choice questions about it (van den Broek et al., 2005). Various standardized reading comprehension tests follow these procedures and are used interchangeably under the assumption that they are equivalent (Keenan, Betjeman, & Olson, 2008). However, accumulating evidence in the extant literature challenges this assumption (e.g., Keenan & Meenan, 2014; Keenan et al., 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 better predicted by code-related skills, whereas a score on the Neale Analysis of Reading Ability was better predicted by language comprehension. Since this study, the demands of several other standardized reading comprehension tests have been examined and reported (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 relies mainly on code-related skills and less on language comprehension abilities (Keenan et al., 2008), whereas the Gray Oral Reading Test and the Gates-MacGinitie Reading Test rely equally on code-related and language comprehension (Cutting & Scarborough, 2006; Tilstra, McMaster, van den Broek, Kendeou, & Rapp, 2009). The results of these studies taken together demonstrate that reading comprehension tests differ in the extent to which they tap on code-related and language comprehension abilities.
Standardized tests, such as those discussed above, have been widely used for research purposes but because of cost and time constraints teachers often choose formative assessments instead. These assessments allow teachers to both assess and monitor the progress of their students’ reading comprehension abilities. For example, the CBM-Maze test is established as a reliable and valid measure to assess reading progress (Espin, Wallace, Lembke, Campbell, & Long, 2010; Pierce, McMaster, & Deno, 2010; Shin, Deno, & Espin, 2000; Tichá, Espin, & Wayman, 2009) and is often used as a formative assessment of reading comprehension in schools to inform instruction (Marcotte & Hintze, 2009). The test has a cloze format: every seventh word is deleted and replaced with three multiple choice alternatives, one correct word and two clearly incorrect words (Shin et al., 2000). 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), 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 could serve as a good formative assessment of reading comprehension.

Thus, the purpose of the current study is to identify the relative code-related and language comprehension demands of the CBM-Maze test. These demands were compared to those of the Gates-MacGinitie reading comprehension test, a standardized test of reading comprehension that has been found to rely on both code-related and language comprehension skills. With respect to 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 (Cutting & Scarborough, 2006; de Jong & van der Leij, 2002; Florit & Cain, 2011; Verhoeven & Perfetti, 2008). The assessment of language comprehension skills often includes measures of vocabulary and listening comprehension (de Jong & van der Leij, 2002). Because we anticipated that the demands of reading comprehension would change developmentally, 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 (e.g., Catts, Adlof, Hogan, & Weismer, 2005; Diakidoy, Stylianou, Karefillidou, & Papageorgiou, 2005). Also, the relation between word reading accuracy and reading comprehension decreases over time (and may disappear by mid-elementary school).
(García & Cain, 2014), whereas the relation between reading fluency and reading comprehension increases, especially from Grade 1 to Grade 3 (e.g., Language and Reading Research Consortium, 2015a). 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 of the demands, we expected the code-related demands to decrease with time, whereas the language comprehension demands we expected to increase. 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 (Tilstra et al., 2009). For the goals of this study, 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. 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, which were on average 9 years and 8 months old ($SD = 4.30$). In Grade 7, there were 36 boys and 53 girls, with an average age of 12 years and 9 months ($SD = 3.79$). In Grade 9, there were 42 boys and 50 girls with a mean age of 14 years and 8 months ($SD = 4.65$). 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

THE 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 (fourth graders) or 2 minutes (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 minutes. 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).

THE GATES-MACGINITIE TEST

The Gates-MacGinitie passage comprehension test (MacGinitie, MacGinitie, Maria, & Dreyer, 2000) includes 11 passages and 48 multiple choice questions related to these passages. 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 record their answers on a separate answer sheet. Administration of the test takes approximately 45 minutes (10 minutes pre-test; 35 minutes 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).
WHAT DOES THE CBM-MAZE TEST MEASURE?

DECODING

Students completed the Word Identification and the Word Attack subtests from the Woodcock Johnson-III Achievement test (McGrew & Woodcock, 2001). Together, they take about ten minutes 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 test and .91 for the Word Attack test (Hosp & Fuchs, 2005).

READING FLUENCY

Students completed a Curriculum Based Measurement task to assess their oral reading fluency (Deno, 1985). In this task, students read aloud three separate age-appropriate passages for one minute 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 three seconds). Reliability ranged between .80 and .91 (Hintze & Silberglitt, 2005).

LINGUISTIC COMPREHENSION

Students completed the Listening Comprehension subtest (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 4th grade test, 38 questions in the 7th grade test, and 40 questions in the 9th grade test. Administration of the test takes approximately 45 minutes. A student’s score consists of the total number of questions answered correctly. Reliabilities ranged from .67 to .79 (Hoover et al., 1996).

VOCABULARY

Students completed the vocabulary subtest from the Iowa Tests of Basic Skills. The 4th-graders completed level 10, the 7th-graders completed level 13, and the 9th-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 different sessions on two different 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, Woodcock-Johnson III Word Identification Task, Woodcock-Johnson III Word Attack Task, and Iowa Tests of Basic Skills Vocabulary) were administered during a third session that was one-to-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.
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

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 only be tested 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; Rodriguez, 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 linguistic 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 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 (FIML) was used. Model fit was evaluated with the chi-square goodness-of-fit test-statistic, the root mean square error of approximation (RMSEA), and the comparative fit index (CFI) (Kline, 2011). A non-significant chi-square value ($p > .05$) indicated exact model fit (Hayduck, 1996). RMSEA values below .05 indicated good approximate fit, values between .05 and .08 were taken as satisfactory fit, and values over .10 were considered as poor fit (Browne & Cudeck, 1993). CFI values larger than .90 were considered acceptable, and values above .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).
WHAT DOES THE CBM-MAZE TEST MEASURE?

Table 1

Descriptive Statistics for Reading Comprehension, Decoding, Reading Fluency, Vocabulary, and Linguistic Comprehension

<table>
<thead>
<tr>
<th></th>
<th>Grade 4</th>
<th>Grade 7</th>
<th>Grade 9</th>
</tr>
</thead>
<tbody>
<tr>
<td>CBM-Maze</td>
<td>M 6.74 SD 4.10</td>
<td>M 18.32 SD 8.27</td>
<td>M 21.50 SD 7.88</td>
</tr>
<tr>
<td>Gates-MacGinitie</td>
<td>M 27.47 SD 11.34</td>
<td>M 33.52 SD 9.24</td>
<td>M 38.71 SD 6.54</td>
</tr>
<tr>
<td>Decoding</td>
<td>M 36.38 SD 6.86</td>
<td>M 46.26 SD 3.62</td>
<td>M 47.82 SD 3.60</td>
</tr>
<tr>
<td>Reading fluency</td>
<td>M 119.59 SD 46.86</td>
<td>M 187.21 SD 39.80</td>
<td>M 153.75 SD 33.94</td>
</tr>
<tr>
<td>Vocabulary</td>
<td>M 6.53 SD 2.97</td>
<td>M 9.59 SD 2.83</td>
<td>M 10.57 SD 2.78</td>
</tr>
<tr>
<td>Linguistic comprehension</td>
<td>M 18.79 SD 3.82</td>
<td>M 25.63 SD 5.09</td>
<td>M 26.83 SD 5.08</td>
</tr>
</tbody>
</table>

Results

Data Screening and Descriptive Statistics

Before conducting the analyses, the data were checked for outliers. Scores that were more than three standard deviations above or below the mean were omitted. In total, less than 1% of the data was omitted. Means and standard deviations of reading comprehension, decoding, reading fluency, vocabulary, and language 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 both tests relate to both code-related and language comprehension skills. In addition, the relations of both reading comprehension tests with decoding seem to decrease across grades, whereas the relations of both reading comprehension tests with reading fluency remained relatively stable. The relations of both tests with vocabulary and linguistic comprehension did not show any clear developmental pattern.
Table 2
Correlations among Reading Comprehension, Decoding, Reading Fluency, Vocabulary, and Linguistic Comprehension in Grades 4, 7, and 9

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

*Note.* All correlations are significant: \( p < .01 \).

**Underlined:** Grade 4  
**Bold:** Grade 7  
**Italic:** Grade 9
Table 3

Standardized Regression Coefficients of the Structural Regression Model with Phantom Factors of the Decoding, Reading Fluency, Vocabulary, and Linguistic Comprehension Demands of the CBM-Maze and the Gates-MacGinitie Reading Comprehension Tests

<table>
<thead>
<tr>
<th></th>
<th>Grade 4</th>
<th></th>
<th>Grade 7</th>
<th></th>
<th>Grade 9</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Maze</td>
<td>Gates</td>
<td>Maze</td>
<td>Gates</td>
<td>Maze</td>
<td>Gates</td>
</tr>
<tr>
<td>Decoding</td>
<td>.13</td>
<td>.12</td>
<td>.16†</td>
<td>.04</td>
<td>-.06</td>
<td>-.00</td>
</tr>
<tr>
<td>Reading fluency</td>
<td>.73**</td>
<td>.39**</td>
<td>.52**</td>
<td>.30**</td>
<td>.72**</td>
<td>.37**</td>
</tr>
<tr>
<td>Vocabulary</td>
<td>.03</td>
<td>.27**</td>
<td>.08</td>
<td>.20†</td>
<td>.14†</td>
<td>.46**</td>
</tr>
<tr>
<td>Linguistic comprehension</td>
<td>.04</td>
<td>.25**</td>
<td>.26**</td>
<td>.38**</td>
<td>.12</td>
<td>.10</td>
</tr>
</tbody>
</table>

Note. † p < .10. * p < .05. ** p < .01.
CHAPTER 3

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 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. Then, it was tested whether the fully constrained model or the model with all coefficients freely estimated in all three grades had a better fit to the data. In addition to the overall model comparisons, 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, $\chi^2 (3) = 1.35, p = .717$, RMSEA = .000, 90% CI [.000 -.128], CFI = 1.00. This model could not be further improved by freely estimating one or all 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, $\chi^2 (3) = 26.31, p < .001$, RMSEA = .292, 90% CI [.196 -.399], CFI = .98. Step-by-step model testing suggested that the equality constraints had to be dropped in all grades. Also, the general model comparisons suggested that the saturated model, with all regression coefficients freely estimated, was the best fitting model. 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 LINGUISTIC COMPREHENSION DEMANDS**

The unique demands of the reading comprehension tests with respect to vocabulary and linguistic comprehension were also examined. It was hypothesized that the Gates-MacGinitie test would rely more heavily on vocabulary and linguistic 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, $\chi^2 (3) = 16.70, p < .001$, RMSEA = .224, 90% CI [.127 -.333], CFI = .99. 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, $\Delta \chi^2 (2) = 15.77, p < .05$, whereas removing the equality constraint in Grade 7 did not result in improvement in model fit. The general model comparisons showed that the model with all regression coefficients of vocabulary freely estimated had a better fit to the data than the model with equality constraints. For linguistic 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, $\chi^2 (3) = 8.92, p = .030$, RMSEA = .147, 90% CI [.040 -.263], CFI = 1.00. Step-by-step model comparisons suggested that the fully constrained model could only be improved by removing the equality constraint in Grade 4, $\Delta \chi^2 (1) = 7.68, p < .05$. The general model comparisons suggested that the model with linguistic comprehension demands freely estimated across tests had a better fit to the data than the model with these demands constrained to be equal across tests. 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, but the reliance on vocabulary of both tests was comparable in Grade 7. With respect to linguistic 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 linguistic comprehension of both tests was comparable. Overall, and based on the general model comparisons, the Gates-MacGinitie test was found to rely more heavily on vocabulary and linguistic comprehension than the CBM-Maze test.
DEVELOPMENTAL PATTERNS OF THE DEMANDS OF THE CBM-MAZE 
AND THE GATES-MACGINITIE TESTS

The multigroup structural regression model with phantom factors was also used to investigate the developmental patterns of the demands of the CBM-Maze and the Gates-MacGinitie tests. Because there is evidence that the demands of reading comprehension change developmentally, we added model constraints to the model in Figure 1, for the four different factors and the two reading comprehension tests separately. We hypothesized that the decoding and fluency demands of reading comprehension tests would decrease, whereas the vocabulary and linguistic 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

For the CBM-Maze test, a model with an equality constraint on the decoding demands across grades had a poor fit to the data, $\chi^2 (2) = 4.18$, $p = .124$, RMSEA = .109, 90% CI [.000 - .259], CFI = 1.00. A model with an equality constraint on the decoding demands in Grades 4 and 7 only had a significantly better fit to the data, $\Delta \chi^2 (1) = 4.14$, $p < .05$. This model fitted the data well, $\chi^2 (1) = 0.04$, $p = .837$, RMSEA = .000, 90% CI [.000 - .162], CFI = 1.00, and could not be further improved. The latter model was thus chosen as the final model. For the Gates-MacGinitie test, a model with an equality constraint on the decoding demands across grades had a good fit, $\chi^2 (2) = 0.89$, $p = .640$, RMSEA = .000, 90% CI [.000 - .164], CFI = 1.00, and could not be further improved. With respect to fluency, a model with equality constraints on the CBM-Maze test across grades had a good fit to the data, $\chi^2 (2) = 3.44$, $p = .179$, RMSEA = .089, 90% CI [.000 - .244], CFI = 1.00, and could not be further improved. A model with equality constraints on the fluency demands of the Gates-MacGinitie test also had a good fit, $\chi^2 (2) = 0.52$, $p = .772$, RMSEA = .000, 90% CI [.000 - .137], CFI = 1.00, and could not be further improved. To sum up, these model comparisons revealed that the fluency demands of both reading comprehension tests remain equal across grades. The decoding demands, however, decrease in Grade 9 for the CBM-Maze test (Grade 4 = Grade 7 > Grade 9), but remain equal across grades for the Gates-MacGinitie test.
WHAT DOES THE CBM-MAZE TEST MEASURE?

VOCABULARY AND LINGUISTIC COMPREHENSION DEMANDS

The developmental changes in the vocabulary and linguistic comprehension demands of reading comprehension were also examined. For the CBM-Maze test, a model with equality constraints on the vocabulary demands across grades had a good fit to the data, $\chi^2 (2) = 1.03, p = .598, \text{RMSEA} = .000, 90\% \text{ CI} [.000 - .171], \text{CFI} = 1.00,$ and could not be further improved. A model with equality constraints on the vocabulary demands of the Gates-MacGinitie test also had a good fit to the data, $\chi^2 (2) = 3.83, p = .148, \text{RMSEA} = .100, 90\% \text{ CI} [.000 - .252], \text{CFI} = 1.00,$ and could not be further improved. For listening comprehension, a model with equality constraints of the CBM-Maze test had a poor fit to the data, $\chi^2 (2) = 5.29, p = .071, \text{RMSEA} = .134, 90\% \text{ CI} [.000 - .279], \text{CFI} = 1.00.$ However, this model could not be further improved by freely estimating coefficients, and was therefore chosen as the final model. The model with equality constraints on the listening comprehension demands of the Gates-MacGinitie test also fitted the data poorly, $\chi^2 (2) = 5.27, p = .072, \text{RMSEA} = .134, 90\% \text{ CI} [.000 - .279], \text{CFI} = 1.00.$ This model, however, could be improved by removing the equality constraint on Grade 9, $\Delta \chi^2 (1) = 3.93, p < .05.$ This model had a good fit to the data, $\chi^2 (1) = 1.34, p = .247, \text{RMSEA} = .061, 90\% \text{ CI} [.000 - .293], \text{CFI} = 1.00,$ 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, however, decrease in Grade 9.

DISCUSSION

The purpose of the current study was to identify the relative unique 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 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 strongly on code-related skills than does the Gates-MacGinitie test, whereas the Gates-MacGinitie test relies more strongly on language
comprehension skills than does the CBM-Maze test. With respect to developmental patterns, the code-related and language comprehension demands of the CBM-Maze and the Gates-MacGinitie tests remained relatively stable across grades, except for the decoding demands of the CBM-Maze test and the listening comprehension demands of the Gates-MacGinitie test that both significantly declined in Grade 9.

The results of the processing demands of the CBM-Maze test highlight that the test relies more on code-related skills than on language comprehension skills. The general model comparisons revealed that the decoding demands of the CBM-Maze and the Gates-MacGinitie test are equal in all grades, but the CBM-Maze test relies significantly more on reading fluency than the Gates-MacGinitie test does. Also, in general the Gates-MacGinitie reading comprehension test relies more heavily on language comprehension skills than does the CBM-Maze test. These findings are interesting considering the high correlation between the two tests themselves across grades, which suggests that the tests are relatively comparable. Despite that, the current findings are in line with previous studies that suggested that cloze tests mainly measure lower-level comprehension processes (e.g., sentence comprehension) (e.g., Gellert & Elbro, 2013).

Further, the developmental comparisons showed that the demands on decoding and fluency remained relatively stable across grades, except for the decoding demands of the CBM-Maze test, which were lower in Grade 9. In addition, the unique contribution of decoding was hardly significant, whereas the role of fluency was substantial. This pattern is in line with recent findings suggesting that with age the role of decoding is gradually taken over by reading fluency in Grades 1 to 3 (Language and Reading Research Consortium, 2015a). In the grade levels targeted by the present study (Grades 4, 7, and 9), students may have already reached a ‘threshold’ level of reading fluency, which may explain the fact that its demands remained stable. This finding is also in line with the idea that Grade 4 is a critical point from where students transition from learning to read to reading to learn (Chall, 1996; McMaster, Espin, & van den Broek, 2014); this transition requires sufficient decoding with speed and accuracy, thus reading fluency.

The developmental comparisons also showed that the linguistic comprehension demands of reading comprehension were lower than expected for both tests (Figure 1 and Table 3). One
potential explanation is the measurement of linguistic comprehension itself. Specifically, in the present study the chosen linguistic comprehension measure taps on a broader set of skills and not only on listening comprehension. The test specifications include comprehension of literal meaning, inferential meaning, numerical/spatial/temporal relations, and also following directions, visual relations, and speaker point of view. Not all of these additional components relate to performance on reading comprehension tests. Another potential explanation is that the nature of the demands on language comprehension changes across development. Specifically, the demands in higher grades (such as Grade 9 in our study) may be more complex than the simple view of reading suggests, and thus cannot be adequately accounted for by linguistic comprehension; for example, they include increasing demands on reasoning skills (Goldman, 2012; Sabatini, O’Reilly, Halderman, & Bruce, 2014) and other 21st century higher-order skills (Goldman & Pellegrino, 2015; Graesser, 2015).

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 poor or good comprehenders (Keenan & Meenan, 2014; Papadopoulos et al., 2014). For example, in a sample of 1500 8 to 19 years old children, only about half of the children were diagnosed as poor comprehenders consistently by two different reading comprehension measures (Keenan et al., 2014). It follows that if the CBM-Maze test is used to identify struggling readers (especially in reading research), 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 it fares better when compared to 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 (Pierce et al., 2010). Future work can focus on how to revise the test so that its language comprehension demands are increased. For example, when cohesion gaps in the text are not placed following a fixed ratio (i.e., each seventh word is removed), one does not have to remove articles, but might choose to always remove verbs or nouns, and thus increase comprehension demands (Gellert & Elbro, 2013). In addition,
the tests’ comprehension demands may increase if the answer possibilities are not clearly different (e.g., She was born on a nice (farm/big/soon) in a valley; Shin et al., 2000), but are likely distractors that need to be evaluated against the greater story context (e.g., She was born on a nice (farm/cave/apartment) in a valley) as is aimed for in standard tests of reading comprehension. 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., in press).

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 if it is to be used as a reading comprehension measure it should be further revised to increase its demands on language comprehension skills.