The technical adequacy of CBM maze as a measure of growth and performance in reading skills of Dutch secondary-school students.

M.L. Donkersloot
s0638811

Leiden University
Master thesis Education and Child Studies
Department: Coach for Learning and Development
Supervisor: Mw. S. Chung, MSc
Second assessor: Mw. Prof.dr. C.A. Espin

August, 2014

Preface

In this thesis I studied the reliability and validity of the Dutch maze reading task as a part of my master studies at Leiden University. The main reason for me to choose this subject was my expectation that it would teach me more about the technical aspects of test construction and test application in education. I supposed this would add to my scientific education as well as my professionalism if I would get a career in education after graduation.

My expectations have been confirmed over the past few months. I studied testing, technical aspects and statistical methods more in depth than I had done before. The things that I learned in this master project were an addition to knowledge from previous courses. Besides, working on this research increased my knowledge of possibilities and limitations of tests. I have gained more insight in types of tests, purposes of testing, which conclusions can be draw from test results and whether it can be justified to take educational decisions based on test results. Having dealt with complications in studying test validity and reliability I became more aware of the contradiction that tests can provide very helpful information, but can as well hardly be constructed, administered, and interpreted as perfect as we would wish. Moreover, information from test outcomes is only a part of the total picture. One of my take-home messages from this project would be that when results either confirm our expectations or when they surprise us we should keep in mind what a test can and cannot do for us.

Finally, the process of writing also taught me much about my competencies and the way I like to work. After some typical struggles to start writing, texts came out easier and faster every time resulting in a -perhaps also typical- ´final thesis writing sprint´. In fact, it was this final sprint including statistical analyses and writing the discussion that I enjoyed most. Although I´m glad to graduate and it was not always easy, I can say without a doubt that I had fun working on this project.

I would like to take this opportunity to thank Prof.dr. C.A. Espin and in particular S. Chung for their supervision. I have appreciated our conversations and your feedback on my work a lot. And the combination of your enthusiasm, commitment, patience and flexibility was a great support as our meetings always got me motivated and inspired to continue. Therefore I conclude with a sincere ‘thank you' for the great time that I had in the department of Coach for Learning and Development.

Myriël Donkersloot

August, 2014

Abstract

This study examines the technical adequacy of the CBM maze task as an indicator of growth and performance in a sample of 578 Dutch 7th grade students. Maze data was collected during 16 weeks in the second semester. A strong alternate-form reliability was found for the first and final three passages ($.74 < r < .82$). Maze growth-rates could not be predicted from VAS reading comprehension scores ($R^2 = .02$). As well, only a small effect was found for the difference in growth rates between education levels. Maze performance in spring and the end of the year were predicted from VAS reading comprehension scores. Explained variances were respectively 27.6% ($\beta = .53$, $p < .05$) and 29.7% ($\beta = .55$, $p < .05$). Also significant differences in maze performance were found between education levels on both time points, respectively $F(2, 556) = 42.29$, $p < .01$, $\omega^2 = .13$ in spring and $F(2, 251) = 49.14$, $p < .01$, $\omega^2 = .27$ at the end of the year. Differences were significant for all groups, being lower, average and higher education levels ($p < .01$). Summarizing no empirical support was found for the validity of maze as an instrument to monitor growth in 7th grade. Results indicate a moderate validity of maze as an indicator of reading performance.

Keywords: Curriculum Based Measurement, reading, maze, growth, performance, secondary school, technical adequacy, validity, alternate-form reliability

The technical adequacy of CBM maze as a measure of growth and performance in reading skills of Dutch secondary-school students.

A new educational law took effect in the Netherlands in August 2010, by which reference levels for Dutch language and mathematics were introduced ('Wet referentieniveau's Nederlandse taal en rekenen'; Education Inspectorate, 2012). According to this law students in 6th grade, the final year of elementary school, are expected to perform at a certain level. If this is not achieved a remedial teaching program must be offered by the secondary school. Secondly, schools are requested to monitor language and mathematics development of all students in all education levels.

The range of educational levels in Dutch secondary education include pre-university education (VWO), senior general secondary education (havo), pre-vocational secondary education (VMBO) and practical training (PRO) (Government of the Netherlands, 2014a). These types of education are entered after primary education at age of 12 and provide curriculums of respectively six, five, four and six years (Netherlands Youth Institute, 2014). Pre-vocational secondary education (VMBO) is designed with four different learning tracks: theoretical programme, combined programme, middle-management vocational programme and basic vocational programme (Government of the Netherlands, 2014b). These programmes prepare students for either more theoretical or more practical further education. Practical training (PRO) prepares students for a specific job by teaching mainly the practical skills needed and is accessible for students who struggle in pre-vocational secondary education (Government of the Netherlands, 2014c). Pre-vocational secondary education provides a combined practical and theoretical curriculum that prepares students for vocational training or secondary vocational education (Government of the Netherlands, 2014b). The remaining levels senior general secondary education and pre-university education are theoretical tracks

that prepare for further education in college or university (Government of the Netherlands, 2013). Students are placed in an education level based on academic capabilities or expected capabilities (Government of the Netherlands, 2013). Still, not all students perform at the level of the instruction they were placed in (Education Inspectorate, 2012). Furthermore the Education Inspectorate (2012) discovered that most secondary schools did not seem to know which students entered with a delay in learning and neither were additional programs available to support these students. In the context of the 'Wet referentieniveau's Nederlandse taal en rekenen' and the variety of education levels within secondary schools, schools need instruments which are suitable to measure a wide range of performance and progress in basic academic skills of their students.

Measuring and monitoring

VAS. A widely used assessment in grade 7 to 9 in Dutch schools is the annual VAS test, developed and published by assessment company Cito (2010). The test consists of eight parts measuring proficiency in reading comprehension in Dutch, mathematics, reading comprehension in English, and general study skills (Cito, 2014). Dutch language is assessed in two reading ability tests of 45 minutes each (Cito, 2010). Several difficulty levels are available that match the education levels. Individual results can be compared to a large sample of the Dutch student population (Cito, 2009). VAS can be used as an evaluation tool and reference levels are included in student reports.

CBM. Another progress monitoring tool is Curriculum-Based Measurement (CBM). CBM was designed by Stanley Deno and his colleagues. Their intention was to create a reliable, valid, simple, efficient, easily understood, and inexpensive alternative for existing assessment methods (Deno, 1985). CBM assesses a general construct rather than sub skills (Fuchs & Deno, 1991). A large amount of research has focused on the academic areas of reading (Wayman, Wallace, Wiley, Tichá, & Espin, 2007), spelling (Marston, 1989),

mathematics (Foegen, Jiban, & Deno, 2007), and written expression (McMaster & Espin, 2007). The general purpose of CBM is to measure a student's proficiency repeatedly in any of these academic domains using alternate forms of similar difficulty level (Deno, Fuchs, Marston, & Shin, 2001). Tests are administered in only a couple of minutes and available material from the curriculum can be used for testing (Deno, 1985). Collected CBM scores are considered to be indicators of performance of both individual students, groups and the relative standing of students compared to their group (Deno et al., 2001), and informs teachers about the effectiveness of their instruction (Stecker, Fuchs, & Fuchs, 2005). As well, the data contain information about the past development and change over time, and enable professionals to predict future performance (Deno, 1985). The absence of change in student's scores over time can indicate that adaptation of instruction is needed (Tichá, Espin, & Wayman, 2009). Other examples of applications of CBM are evaluation of educational programs, identification and monitoring of low performing students, referral to special education services, and reintegration in regular classrooms after interventions (Stecker et al., 2005). Collected CBM data is typically plotted in a graph that facilitates easy communication between professionals and parents about performance and progress (Deno, 1985; Madeleine & Wheldall, 2004).

An important theme in studies about CBM has been its technical adequacy. CBM procedures can only be applied reliably for progress monitoring if technical features like validity, stability, inter-scorer reliability (Deno et al., 2001), and alternate-form reliability (Wayman et al., 2007) are satisfactory. Marston (1989) summarized the results of initial research on the technical adequacy of CBM in the areas of reading, spelling, written expression, and mathematics. Criterion validity was high ($r > .73$) for reading, spelling, and written expression. Also discriminative validity was strong, distinguishing reliably between for example mildly handicapped students and students in regular education programs. Validity

of math appeared to be lower with only few correlations above $r = .60$ and median coefficients varying between $r = .43$ and $r = .54$. Test-retest and alternate-form reliability coefficients of reading ($.82 < r < .97$) and spelling ($.72 < r < .97$) were highest, while moderate to high coefficients were reported for written expression ($.42 < r < .96$). CBM math was found reliable, but only for single administrations. Inter-scorer agreement was high for all measures. Similar results were found in more recent research (reviewed by e.g., Espin & Campbell, 2012; Foegen et al., 2007; McMaster & Espin, 2007; Wayman et al., 2007). CBM procedures were indicated to be suitable as a screening instrument and to distinguish between performance in special and regular education (Madeleine & Wheldall, 2004).

Other than assessing proficiency on one particular moment, CBM is also suitable for measuring growth. The use of equivalent forms allows changes in scores to be interpreted as changes in student performance (Stecker et al., 2005). Deno et al. (2001) describe in more detail that CBM procedures contain sufficient alternate forms of equal difficulty level while assessing the same construct without reaching floor or ceiling effects. Within one school year growth lines can be expected to be approximately linear (Deno et al., 2001). This linear model enables educators to predict future results and to set goals for individual students.

In the late 90s researchers began to investigate the technical adequacy of CBM in secondary school (Madeleine & Wheldall, 2004). Research proved CBM adequate for measuring performance and progress of content-area learning and reading (Espin & Campbell, 2012). Studies on CBM writing in secondary education found lower reliability and validity coefficients and were mainly focused on performance, not on growth.

CBM reading. A vast amount of research was conducted in the area of CBM reading. The technical adequacy of CBM reading seems strongest among CBM domains (Espin & Campbell, 2012). The two most common reading measures for older students are (a) reading aloud from a text for 1 minute and (b) the maze task in which the reader chooses words that fit

in a text from multiple choice options during 1 to 4 minutes (Deno et al., 2004; Tichá et al., 2009; Wayman et al., 2007).

Reading aloud, also referred to as oral reading fluency (ORF), is often administered with passages from the curriculum which, in difficulty, are comparable to the instructional level that is to be achieved at the end of the current academic year (Stecker et al., 2005). The number of words read correctly is the student's score (Madeleine & Wheldall, 2004). Validity of reading aloud is generally high ($r > .70$) when compared to performance on reading comprehension tests (Deno, 1985; Madeleine & Wheldall, 2004). Scores on reading aloud differentiate between students receiving special education services and students in regular classrooms, which supports the criterion validity of the test (Deno, 1985). Also strong reliability coefficients (> .80) are found for reading aloud (Petscher, Cummings, Biancarosa & Fien, 2012). Repeated measurement of reading aloud gives an indication of growth in reading performance in elementary school (Deno, 1985). Slopes are normally steepest in the beginning of the school year and in the first years of reading instruction (Baker et al., 2008; Deno et al., 2001; Petscher et al., 2012). Differences in growth rates are measurable between special education (SE) students and students in regular education (Deno et al., 2001). Slopes of SE students are more flat than their peers' in regular education during first grade, resulting in lower performance levels when entering Grade 2. Growth rates can be comparable between both groups when effective instruction is provided to disabled students, but nevertheless average performance levels of SE students stayed lower than their nondisabled peers' until the end of elementary school (Deno et al., 2001). Limitations of reading aloud are the time needed for testing since each student has to be seen individually and its face validity (Madeleine & Wheldall, 2004). Despite high correlations with other reading comprehension measures, teachers have difficulty accepting reading aloud as an indicator of general reading ability. Moreover, reading aloud does not seem to be suitable for monitoring growth in

secondary school as opposed to the second CBM reading instrument maze (Espin & Campbell, 2012; Espin et al., 2010).

The main alternative for reading aloud in CBM is maze. When completing a maze task students read a text in which from the second sentence every seventh word is changed into a multiple choice set of three words (Espin, Wallace, Lembke, Campbell, & Long, 2010; Shin, Deno, & Espin, 2000). Each set of three words contains one correct option, while the other two words do not fit in the sentence. Students are supposed to select the correct answers whilst read the text. The data collected with maze can be used to identify weak readers, detect stagnation in reading development, and to evaluate effectiveness of instruction and intervention over relatively short periods of time (Deno et al., 1985; Shin et al., 2000). Empirical research shows maze to be a reliable and valid instrument to monitor general reading development (Tichá et al., 2009; Espin et al., 2010; Wayman et al., 2007). Maze can be group administered in one or two grade levels below or above the instruction level, meaning that texts from Grade 4 to 8 can be taken by sixth graders (Wayman et al., 2007). Consequently, it is possible to administer passages of equal difficulty during several years and compare individual results over years. Other benefits are the short time frame of 2 to 4 minutes (Espin et al., 2010), sensitivity for growth (Espin et al. 2010; Tichá et al., 2009) of both group and individual performance, suitability for repeated measures (Shin et al., 2000), and high teacher satisfaction (Wayman et al., 2007).

Both Espin et al. (2010) and Tichá et al. (2009) investigated the technical adequacy of maze in secondary school and found similar results. Alternate-form reliability coefficients were found to be generally high ($r > .80$) and a positive correlations were found with high stake measures ($.80 < r < .88$), indicating strong criterion validity (Tichá et al., 2009). Tichá et al. (2009) found that scores in 8th grade were significantly different for lower versus higher performing students (respectively in average 11.74 and 23.32 correct answers), contributing to

maze's validity as a measure of reading performance. Face validity of maze is stronger than was found for reading aloud (Shin et al., 2000).

As opposed to reading aloud, maze is sensitive for change in reading performance of students in secondary education (Espin et al., 2010; Tichá et al., 2009). Change in scores can be interpreted as change in performance because alternate forms are of comparable difficulty (Espin et al., 2010). The standard error when measuring growth with maze is relatively small (Deno et al., 2001). However, researchers recently pointed out that slopes can be computed with little error only after a longer period of eight to ten measurements (Christ, Zopluoglu, Monaghen & Van Norman, 2013). Different growth rates can be expected for students with and without learning disability. Growth rates for higher performing students were found to be 1.31 words per week, while lower performing students showed a significantly slower improvement of 0.41 words per week (Tichá et al., 2009). Performance and growth on standard state reading tests correlates positively with growth on maze, with higher scores on standard tests relating to higher growth rates on maze tasks (Espin et al., 2010; Tichá et al., 2009).

In summary, maze seems to have several advantages over read aloud in the context of monitoring reading progress in secondary school. Also for practical reasons such as administration in groups, maze seems the most efficient CBM reading probe available. However, few studies on technical adequacy of growth on maze in secondary school are published and in earlier studies data was collected in relatively small samples (Espin & Campbell, 2012).

This study intends to contribute to the investigation of the Dutch maze's reliability and validity. The main research question of the current study is: What is the technical adequacy of the maze task as a measure for growth and performance of reading skills of 7th grade secondary school students? To find an answer to this question a) alternate-form reliability is

investigated, b) growth and performance on the maze task is predicted from performance on a widely used reading test called VAS, and c) differences in growth rates and performance between students in different education levels are examined.

Espin et al. (2010) and Tichá et al. (2009) described that higher performing readers usually show more growth over time than lower performing students. Based on these former results higher growth rates are expected to be found in better readers and therefore higher growth rates are expected to be found in students who perform better on the VAS reading sub test. Besides, better performance on maze is expected for students with higher scores on the VAS reading sub test, which is a reading comprehension test. Since comprehension is the purpose of reading, CBM reading tests should correlate with outcomes of comprehension measures (Deno, 1985). Finally, maze scores and education level are expected to relate to each other. The Dutch education system provides practical education to students in lower tracks and more theoretical instruction to students in higher tracks. Consequently students with difficulty in theoretical courses receive more practical education in the middle-management vocational programme and basic vocational programme (Government of the Netherlands, 2013). The more theoretical tracks demand better reading skills from students and students will practice their reading more often because more books are used for instruction. Being better readers and spending more time reading, higher performance and more growth in reading skills can be expected in higher education levels. If maze is a valid instrument for measuring performance and monitoring change in reading skills, higher performance and more improvement should be reflected in both higher scores and growth rates for students in higher school levels than in lower school levels.

Method

Setting and Participants

Data were collected in three secondary schools in mid-/west Holland. The final dataset included 578 students (294 boys) in the age of 12 to 16 years old (M = 12.72, SD = 0.70). The majority of participants were born in the Netherlands (91.5 %), some were born in countries where Dutch was one of the recognized languages (1.2 %). Few were born in other western countries including USA or Australia, (0.3 %), East European countries, Russia or non-western Mediterranean countries (5.3 %), or other countries (1.4 %). All types of secondary education were represented with 271 students in the practical and basic prevocational tracks (46.9 %), 175 students in the advanced prevocational or combined track (30.3 %), 37 students in the theoretical prevocational track (6.4 %) and 95 students in senior general secondary or pre-university education (16.4 %). In total 59 students were diagnosed with dyslexia (10.2 %).

Measures

Maze. Maze consisted of a set of texts of equal difficulty level and a length of about 400 words. In the texts every seventh word was replace by a multiple choice set of three words (Espin et al., 2010; Shin et al., 2000). One of the three words fit in the sentence, the others were distracters. In the current research the testing time per text was set on two minutes. Students were instructed by their teachers to read as much as possible in two minutes and select correct answers while reading. The number of correct answers was the final score. Writers of the Dutch maze passages were instructed to write an informative text on 6th grade level while attempting to avoid bias due to gender, culture, or pre-existing knowledge. The reliability and validity of the maze passages were examined in the current study.

VAS reading. VAS reading was available in three levels meant for basic prevocational education, advanced to theoretical prevocational education and senior general secondary or pre-university education. Each of these three forms consisted of two reading tests of 45 minutes including questions about the content of the passages. The Dutch Committee on Tests and Testing (COTAN) considered VAS reading to be a good test in terms of test construction, quality of the test and teacher manual, norms and reliability and to have satisfactory construct validity. Not enough research was done to draw conclusions about criterion validity (COTAN, 2009a). Reliability coefficients of all levels of the 7th grade reading tests were between .77 and .82 (COTAN, 2009b). Standardized proficiency scores were computed from raw results on VAS reading, enabling professionals to compare results of students from different education levels (Cito, 2009).

Procedures

Data used in this study were part of a larger study. Maze tasks were taken weekly from March until June under supervision of classroom teachers. The total set existed of 18 passages, including one example text. The texts were accessible through a secured website where students could log in with personal codes. Teachers received instructions to guide their students while taking the test. Trained master students visited schools to observe the data collection moments and to provide feedback to teachers. The maze website was designed to export data in a format suitable for further data analysis.

VAS reading was taken once in autumn with the regular protocol from Cito. The scores were obtained from the schools.

Data analysis

After data inspection and investigation of the alternate-form reliability of relevant passages with Pearson's correlation, growth on the maze task was computed as the difference in performance between initial and final passages. To enhance reliability, performance on initial passages was derived by calculating the mean value of the first three passages. Similarly, the mean value of the final three passages formed the average value of the end of the data collection period. In a simple regression analysis, performance on VAS reading was tested as a predictor of maze growth.

Finally, groups classified by education level were compared with growth rates on maze using an AVOVA. Three groups were formed based on either a practical or more theoretical focus in educational tracks and the numbers of participants in groups. Group A included both students in the practical track, and students in basic prevocational track who were receiving more practical education. Group B included students from the advanced prevocational track, combined track, and theoretical prevocational track who were receiving a combination of practical and basic theoretical education. Finally group C included students who received senior general secondary education or pre-university education. Groups are referred to as group A, B, or C or respectively lower, average and higher education levels.

Results

VAS reading-scores were collected from 82% (N = 474) of the sample and ranged from 165 to 267 (M = 204.82; SD = 17.79). Means of separate maze passages varied from 24.63 to 31.34, with minimum scores between 1 and 8 and maximum scores between 49 and 56. Group sizes differed for each passage ($222 < N < 441$). New variables were computed to calculate maze growth. The means of the first three ($maze_{t1}$) and passages 13, 14, and 15 ($maze_{t2}$) were calculated, allowing maximally one missing value in each calculation. The final

passage 16 was not included in analyses because of the large number of missing data (74%). An estimate of growth of student's performance on the maze task ($maze_{t2-t1}$) was computed from the difference between $maze_{t1}$ and $maze_{t2}$. Descriptive statistics of maze passages, $maze_{t1,}$ $maze_{t2,}$ and $maze_{t2-t1}$ are summarized in Table 1.

Table 1
Descriptive statistics of individual passages, average maze values, and growth

	N	*M*	*SD*	Minimum	Maximum	Missing (%)
Passage 0	336	24.63	8.30	1	49	41.9
Passage 1	441	24.77	6.83	8	53	23.7
Passage 2	291	26.03	7.77	1	54	49.7
Passage 13	320	26.05	9.18	1	53	44.6
Passage 14	224	28.35	9.25	1	51	61.2
Passage 15	222	31.34	9.23	4	56	61.6
$Maze_{t1}$	559	24.70	7.01	3.00	50.00	3.3
$Maze_{t2}$	254	29.23	8.70	3.50	53.50	56.1
$Maze_{t2-t1}$	245	2.60	5.46	-13.83	19.00	57.6

Distribution, normality, missing values, and outliers of variables were investigated before computing other analyses. The distribution of the VAS scores was approximately normal and no outliers were found. Also results of separate maze passages and newly computed maze variables were distributed approximately normally and no outliers were detected. The percentages of missing values was high for most variables, yet remaining group sizes were large enough to execute ANOVA and regression analysis (Table 1). Regarding ANOVA, inequality of variances between groups of education levels was found for $maze_{t1}$ with Levene's test ($F(2, 556) = 13.20$, $p < .01$). This violation of the assumption of homogeneity of variances could have been caused by larger differences in group sizes. Therefore ANOVA was computed including Welch's test, which is robust to unequal group sizes (Field, 2013). Subsequently, Gabriel and Games-Howell *post hoc* tests were executed to interpret possible results more reliably in case of respective differences in group size or in case of unequal variances (Field, 2013). As a final step in the data inspection procedure,

histograms and scatterplots were inspected to verify whether assumptions of linearity and homoscedasticity were met before performing regression analysis on VAS and maze scores. No violation of either of the assumptions was found.

Alternate form reliability

Three maze passages from both the beginning and end of the data collection period were selected to calculate growth on the maze task. The alternate form reliability among the first three passages was between $r = .77$ and $r = .82$, with N ranging from 125 to 242. The final three passages also correlated strongly with each other, $r = .74$ and $r = .80$, with N between 151 and 174. An overview of all correlations is presented in Table 2.

Table 2
Alternate form reliability: Pearson's correlation coefficients of maze passages 0-2 and 13-15

Passage	0	1	2	Passage	13	14	15
0	-			13	-		
1	.82**	-		14	.77**	-	
2	.80**	.77**	-	15	.80**	.74**	-

** $p < .01$

Validity

Validity of maze as a measure of growth in reading was investigated in two different ways. First, the relation between maze growth and VAS reading scores from the beginning of the school year was examined using regression analysis in SPSS. Second, an ANOVA was conducted to answer the question of whether there was a relation between education level and maze growth.

A simple regression analysis showed that VAS-scores explained 1.6% of variance in maze growth and appeared not to be related to growth in maze, although the model was significant ($F(1, 229) = 4.81$, $p < .05$). Further analysis of the relation between VAS reading

scores on maze performance revealed that VAS reading explained 27.6% of variance in $maze_{t1}$, $F(1, 458) = 174.75$, $p < .01$. Also VAS reading scores accounted for 29.7% of variance in $maze_{t2}$, $F(1, 236) = 99.65$, $p < .01$. A slightly higher beta value was found for $maze_{t2}$ ($\beta = .55$, $p < .05$) than for $maze_{t1}$ ($\beta = .53$, $p < .05$), being both medium to large effect sizes. From the beta values it can be concluded that VAS reading predicted $maze_{t2}$ better than $maze_{t1}$, although this difference was minimal.

In an ANOVA differences in maze growth and performance between education levels were tested. The commonly used significance level of $\alpha = .05$ was divided by 3 and set on $\alpha = .016$ to prevent Type I-error when computing ANOVA three times.

Significant differences in maze growth with a small effect size were found when comparing education levels, $F(2, 242) = 7.50$, $p < .01$, $\omega^2 = .05$. *Post hoc* tests demonstrated a significant difference between mean maze growth of group A ($M = 1.24$, $SD = 5.56$) and group C ($M = 4.32$, $SD = 4.74$), $p < .01$. It can be concluded that students in higher education levels showed a significantly stronger average growth on the maze tasks than students in lower education levels. Differences between group B and groups A and C were not significant.

Significant differences in performance on $maze_{t1}$ were found between education levels with a medium to large effect size, $F(2, 556) = 42.29$, $p < .01$, $\omega^2 = .13$. Differences between groups were significant also when controlled for unequal group sizes, *Welch's adjusted* $F(2, 169.15) = 35.03$, $p < .01$. Both Gabriel and Games-Howell *post hoc* tests showed significant differences between all groups ($p < .01$), meaning that group C ($M = 29.65$, $SD = 8.30$) scored significantly higher on $maze_{t1}$than group B ($M = 26.51$, $SD = 4.83$) and group A ($M = 23.06$, $SD = 6.37$), as well as group B scored significantly higher than group A.

Also for $maze_{t2}$ differences between education levels were found, $F(2, 251) = 49.14$, $p < .01$, $\omega^2 = .27$. This effect size was large. Differences were significant at .016 level for all groups ($p < .01$), which means that similarly to $maze_{t1}$ the average performance of students in group C on maze ($M = 35.93$, $SD = 7.80$) was significantly higher than performance of students in group B ($M = 29.13$, $SD = 7.00$) and group A ($M = 25.01$, $SD = 7.36$). Likewise, students in group B performed better on the maze than students in group A.

Discussion

CBM is a measurement system that intends to assess both performance and growth in academic areas such as reading (Deno, 1985). Various instruments can be used for assessment in a CBM setting. The aim of this study was to contribute to the research on the technical adequacy of the CBM maze task for assessing growth and performance in reading. A high alternate-form reliability was found in the current sample ($.74 < r < .82$), although values in earlier research by Tichá et al. (2009) ranged up to $r = .90$. The validity of maze as a monitoring instrument could not be supported by predicting maze growth rates from VAS reading scores. Although the results were significant, VAS reading scores explained only 1.6% of the variance in growth rates. Further analysis demonstrated that VAS reading could predict a substantial part of the maze performance scores of both the beginning of spring and the end of 7^{th} grade, explaining respectively 27.6% and 29.7% of the variance in maze performance. Surprisingly, VAS predicted maze performance at the end of the year better than earlier in spring while VAS was administered at the beginning of the year, although differences in beta values were small. Espin et al. (2010) found higher validity coefficients (> .70) when predicting MBST state standard test results from maze performance. When comparing current results to the Espin et al. (2010) study it could be concluded that for the

Dutch maze a moderate validity was found when predicting performance from criterion variable VAS reading.

Second, the relation between education level and maze growth and performance was examined. Although maze growth could not be predicted from VAS reading results, some differences between education levels were found in growth on the maze task. Students in the higher levels presented higher growth rates than students in lower education levels, but the effect size of this relation was small. Differences between lower and average, and between average and higher levels were not significantly reflected in maze growth rates. Meantime, these differences between the average education level and other groups were found in maze performance in early spring and the end of 7th grade with medium to large effect sizes.

The expectation to find differences in maze scores is partially supported with these results. A larger effect was expected for growth rates and the results did not support the hypothesis that maze would be a valid instrument for monitoring growth in reading. The finding that maze scores did distinguish between education levels with a moderate to large effect, and that maze performance could to some extent be predicted from VAS-scores, suggests a moderate validity of maze as indicator of reading performance in 7th grade.

Even though some support is found for maze's validity as a reading performance indicator, stronger relations with the VAS reading test and education were expected. Several explanations can be found for the moderate relation with criterion variable VAS reading. First, a diversity in development patterns could have existed for which the relative ranking of students' performance levels changed in the time between VAS reading and maze administration. Second, it is possible that VAS reading and maze do not measure the exact same construct. Maze intends to measure general reading proficiency while VAS reading is a reading comprehension test. Also the types of texts might have been different. In maze solely informative texts were used, while in VAS reading tests of various types of texts are

implemented (Cito, 2012). Third, although students seemed focused during observations, motivation to complete the test with good results could have affected outcomes in some way. Cito's VAS is known as an important test and students know that educational decisions are based on their performance on the test. Perhaps students' motivation to complete VAS reading was different from their motivation to perform at their best on the relatively simple, short and repeated maze tests.

The relation between education level and maze performance was somewhat more convincing, although effects were still moderate to large. It is possible that variety within education groups caused overlap in reading performance of the groups used in this study. For example, in 2009 the Dutch Inspectorate of Education has investigated performance of students who entered secondary schools on several academic areas and reported that substantial groups did not meet the level of instruction (Education Inspectorate, 2012). The normative study of VAS reported that mean VAS reading-scores in six education levels ranged between 191 and 234, with *SD*s between 11 and 14 (Cito, 2009). These findings illustrate that even though students in the Dutch system are classified in groups with comparable academic capacities, a variety of performance levels can be found within the education levels. Moreover, the distribution of VAS reading performance of separate education levels seem to overlap strongly, for which differences between education levels are difficult to distinguish. If in reality differences in reading performance are not clearly present between educational levels, it is evident that neither maze scores will differ strongly between groups.

Several explanations can be brought up for the lack of relation between VAS and maze growth, as well as for the small relation between education level and growth rates. From earlier research is known that growth rates for CBM reading are strongest in the first years of elementary school and more flat until grades 5 and 6 (Baker et al., 2008; Deno et al., 2001;

Petscher et al., 2012). Secondly, the same research showed that growth rates tend to be more flat in the second semester. It is possible that this pattern applies to maze growth in 7th grade as well and that therefore the difference in growth is too small in all students to distinguish between good and weak readers and education levels. Since this study was conducted in the second semester of 7th grade, growth and differences in growth were possibly small and hard to detect. Besides, much data of the final maze passages from students in the lowest education levels was missing, which can have influenced the range and distribution of maze scores. Also this can have affected the growth rates, which were computed from the difference between final and initial maze results. As well the generalizability of results is restricted due to these missing values.

Another issue regarding measuring growth is the stability of the used passages. The intention in CBM is to work with different passages that are equal in difficulty, because equal difficulty of alternate forms increases the stability of growth lines and therefore the precision when estimating growth (Shin, Espin, Deno & McConnell, 2004). Preferably, the alternate-form reliability would be as close to $r = 1$ as possible, for differences in scores can be interpreted as growth only with equality of the passages. But in practice equivalence of reading passages is very hard to ensure (Albano & Rodriguez, 2012, Christ et al., 2013). In this research a strong alternate form reliability was found, however, the correlation was not as high as desired. The reliability might have been higher if missing values of students could have been prevented. For example, students in the lowest education level missed many scores in the final passages. Therefore not only the group size, but also the group composition was different at the beginning and end of the study. It can be assumed that this instability can have caused variability in maze scores too. Ideally data would be collected more continuously from all students.

After data collection, statistical methods to analyze data are an important issue when examining validity. Various methods are applicable, but all have their limitations. Repeated measures ANOVA for example is complicated to use because missing values are not allowed and data collection of each passage has to happen at the exact same moment (Field, 2013). These circumstances are difficult to realize when working with several schools for one study. Regression analysis is less suitable for categorical variables such as education level. Moreover, regression analysis controls for shared variance when independent variables are entered together (Field, 2013). In validity research we would like to look at relations separately. This study for example did not intend to investigate the additional explained variance or interaction effect of education level in relation to VAS reading performance, but the relation of these two factors separately with maze results.

On the other hand, effects can be overestimated when the interaction between independent variables is ignored. This dilemma can occur easily when research focusses on criterion variables, since criterion variables usually are supposed to measure the same construct. A similarity will exist and therefore also overlap in explained variance in a regression model when several criterion variables are entered in one model. Also other types of variables can interact, like education level and measures of reading correlated in this study ($-.04 < r < .69$). A relation between these two constructs was expected and needed to demonstrate validity of maze. In the current study it was decided to analyze the independent variables separately to see their individual relations to maze values. Therefore a part of the effects that were found can overlap and results should be interpreted carefully.

A more optimal analysis than ANOVA and regression analysis when studying growth is Multilevel Analysis, also known as Hierarchical Linear Modeling (HLM), which was not possible to use in this study due to practical reasons. The advantage of Multilevel Analysis is that it can analyze multiple data points, it is robust to missing data, and it increases the

reliability of results when analyzing growth estimates (Shin et al., 2004). Even so, also working with Multilevel Analysis the question applies whether variables should be put together in one model or not.

The results and discussion so far gave rise to ideas for further research. This study focuses only on the technical aspect of the maze tests. Another important area to investigate is how teachers can and like to implement a test and which benefits maze could have in the practice of reading education. Considering the technical features and the prediction that usage of CBM would improve teaching and learning, Stanley Deno expected CBM to be "a promising alternative approach for continuously measuring student achievement of proficiency in basic school skills" (Deno, 1985, p.230). Nowadays this expectation has been confirmed by various empirical studies. Deno et al. (2001) summarize for example how implementation of CBM in primary schools was found to increase teacher's expectations, the frequency of changing instruction and learning outcomes. Madeleine and Wheldall (2004) reviewed literature on CBM in reading and found that teachers rated CBM higher than norm referenced measures and that the data was mainly used for communication and to check the correctness of their observations. It must be noted that data collection itself does not improve student achievement (Stecker, Fuchs & Fuchs, 2005). Although CBM is simple to use, it is most effective when teachers are trained in the procedures of data collection and evaluation (Deno, 1985; Stecker et al., 2005). Further research could for example focus on the application of CBM maze in classrooms. The way how teachers use CBM and evaluate data, as well as their satisfaction about the method, and effectivity reflected in student performance are interesting and essential research themes. More research on technical adequacy is needed too. Further research could add to the results of this study by attempting to decrease the amount of missing values, especially in lower education levels. It is worth considering carefully which criterion variables to select and how data of these variables can be analyzed.

It could be interesting to collect data throughout the whole school year instead of only in the second semester. A final idea is to conduct a study in elementary school to be able to compare Dutch results better to the body of American research. In elementary school growth is steeper and results are therefore more explicit. Moreover, the majority of research in the USA was conducted in elementary schools, so more is known about technical aspects of CBM in that setting than in secondary school. For these reasons research in elementary school could give more information about the quality, validity and reliability of the Dutch instrument.

It can be concluded that moderate empirical support is found for the Dutch maze as a measure of reading performance in 7th grade. No support was found to apply maze as a monitoring instrument, which would have been useful in the context of the recently introduced reference levels in secondary education. For now maze seems to be a fairly valid method to estimate reading skills on one moment in time. It seems promising as a screening instrument in 7th grade and can give teachers information about reading ability in a shorter time frame than the administration of other tests like VAS reading. On the other hand, maze provides only a general score and no detailed information about the cause or type of reading problem, nor about the type of change that could be needed in instruction (Madeleine & Wheldall, 2004). This diagnostic information can be collected with other instruments. Future research should be done to draw conclusions about the technical adequacy and implementation in classrooms with more certainty.

References

Albano, A.D., & Rodriguez, M.C. (2012). Statistical equating with measures of oral reading fluency. *Journal of School Psychology, 50,* 43-59.

Baker, S., Smolkowski, K., Katz, R., Fien, H., Seeley, J., Kame'enui, E.J., & Thomas Beck, C. (2008). Reading fluency as a predictor of reading proficiency in low-performing high poverty schools. *School Psychology Review, 37,* 18-37.

Cito (2009). *Volg- en Adviessysteem (VAS): wetenschappelijke verantwoording.* Netherlands, Arnhem: Cito B.V.

Cito (2010). *Meting taal en rekenen: Tweede meting: een indicatie van leerprestaties in termen van referentiekader.* Retrieved November 14, 2013 from http://www.rijksoverheid.nl/documenten-en-publicaties/rapporten/2011/06/07/meting-taal-en-rekenen-2010.html.

Cito (2014). *Monitoring and Evaluation System for Students.* Retrieved May 27, 2014 from http://www.cito.com/products_and_services/education/monitoring_and_evaluation_system/monitoring_and_evaluation_system_for_students.

COTAN (2009a). *Volg- en Adviessysteem Cito, VAS, 2009*. Retrieved March 7, 2014 from http://www.cotandocumentatie.nl/test_details.php?id=733.

COTAN (2009b). *Volg- en Adviessysteem (VAS): Researchbeschrijving*. Retrieved March 7, 2014 from http://www.cotandocumentatie.nl/test_details.php?id=733.

Christ, T.J., Zopluoglu, C., Monaghen, B.D., & Van Norman, E.R. (2013). Curriculum-Based Measurement of Oral Reading: Multi-study evaluation of schedule, duration, and dataset quality on progress monitoring outcomes. *Journal of School Psychology, 51,* 19-57.

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

Deno, S.L., Fuchs, L.S., Marston, D., & Shin, J. (2001). Using Curriculum-based Measurement to Establish Growth Standards for Students with Learning Disabilities. *School Psychology Review, 30*(4), 507-524.

Education Inspectorate (2012). *Achterstandbestrijding en referentieniveaus voor taal en rekenen in het vo*. Retrieved October 21, 2013 from http://www.onderwijsinspectie.nl/binaries/content/assets/Actueel_publicaties/2012/achterstandbestrijding-en-referentieniveaus-voor-taal-rekenen-in-het-vo-printversie.pdf.

Espin, C.A., & Campbell, H.M. (2012). They're Getting Older…but Are They Getting Better? The Influence of Curriculum-Based Measurement on Programming for Secondary School Students with Learning Disabilities. In Espin, C.A., McMaster, K.L., Rose, S., & Wayman, M.M. (Eds.), *A Measure of Success: The Influence of Curriculum-Based Measurement on Education* (p. 149-162). Minneapolis, MN: The University of Minnesota Press.

Espin, C., Wallace, T., Lembke, E., Campbell, H., & Long, J.D. (2010). Creating a Progress-Monitoring System in reading for Middle-School Students: Tracking Progress Toward Meeting High-Stakes Standards. *Learning Disabilities Research & Practice, 25*(2), 60-75.

Field, A. (2013). *Discovering Statistics Using IBM SPSS Statistics.* London: SAGE Publications Ltd.

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

Fuchs, L.S., & Deno, S.L. (1991). Paradigmatic distinctions between instructionally relevant measurement models. *Exceptional Children, 57*, 488-500.

Government of the Netherlands, 2013. *Voortgezet Onderwijs 2013-2014: Gids voor ouders, verzorgers en leerlingen.* Retrieved August 10, 2014 from

http://www.rijksoverheid.nl/onderwerpen/voortgezet-onderwijs/documenten-en-publicaties/brochures/2014/03/21/voortgezet-onderwijs-2013-2014.html.

Government of the Netherlands, 2014a. *Secondary education.* Retrieved March 24, 2014 from http://www.government.nl/issues/education/secondary-education.

Government of the Netherlands, 2014b. *Pre-vocational education (VMBO).* Retrieved March 24, 2014 from http://www.government.nl/issues/education/pre-vocational-education-vmbo.

Government of the Netherlands (2014c). *Wanneer krijgt mijn kind praktijkonderwijs?* Retrieved March 24, 2014 from http://www.rijksoverheid.nl/documenten-en-publicaties/vragen-en-antwoorden/wanneer-krijgt-mijn-kind-praktijkonderwijs.html.

Madelaine, A., & Wheldall, K. (2004). Curriculum-based measurement of reading: Recent advances. *International Journal of Disability, Development and Education, 51*, 57-82.

Marston, D.B. (1989). Curriculum-Based Measurement Approach to Assessing Academic Performance: What It Is and Why Do It. In Shinn, M.R. (Ed.): *Curriculum-Based Measurement Assessing Special Children.* New York: The Guilford Press.

McMaster, K., & Espin, C.A. (2007). Technical features of Curriculum-Based Measurement in writing: A literature review. *The Journal of Special Education, 41*, 68-84.

Netherlands Youth Institute. *The Dutch Education System*. Retrieved March 24, 2014 from http://www.youthpolicy.nl/yp/Youth-Policy/Youth-Policy-subjects/Education-and-Youth-Unemployment/The-Dutch-Education-System.

Petscher, Y., Cummings, K.D., Biancarosa, G., & Fien, H. (2012). Advanced (Measurement) Applications of Curriculum-Based Measurement in Reading. *Assessment for Effective Education, 38*(2), 71-75.

Shin, J., Deno, S.L., & Espin, C. (2000). Technical Adequacy of the Maze Task for Curriculum-Based Measurement of Reading Growth. *The Journal of Special Education, 34*(3), 164-172.

Shin, J., Espin, C.A., Deno, S.L. & McConnell, S. (2004). Use of Hierarchical Linear Modeling and Curriculum-Based Measurement for Assessing Academic Growth and Instructional Factors for Students with Learning Difficulties. *Asia Pacific Education Review, 5*(2), 136-148.

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

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 & Practice, 24*(3), 132-142.

Wayman, M.M., Wallace, T., Wiley, H.I., Tichá, R., & Espin, C.A. (2007). Literature Synthesis on Curriculum-Based Measurement in Reading. *The Journal of Special Education, 41*(2), 85-120.