

Universiteit
Leiden
The Netherlands

CBM progress monitoring in reading and foreign-language learning for secondary-school students

Chung, S.

Citation
Chung, S. (2018, June 26). *CBM progress monitoring in reading and foreign-language learning for secondary-school students*. Retrieved from https://hdl.handle.net/1887/63990

Version:	Not Applicable (or Unknown)
License:	Licence agreement concerning inclusion of doctoral thesis in the Institutional Repository of the University of Leiden
Downloaded from:	https://hdl.handle.net/1887/63990

Note: To cite this publication please use the final published version (if applicable).

Cover Page



Universiteit Leiden



The handle http://hdl.handle.net/1887/63990 holds various files of this Leiden University dissertation.

Author: Chung, S.
Title: CBM progress monitoring in reading and foreign-language learning for secondary-school students
Issue Date: 2018-06-26

CBM PROGRESS MONITORING IN READING AND FOREIGN-LANGUAGE LEARNING FOR SECONDARY-SCHOOL STUDENTS

SIUMAN CHUNG

CBM Progress Monitoring in Reading and Foreign-Language Learning for Secondary-School Students

Siuman Chung

CBM progress monitoring in reading and foreign-language learning for secondary-school students

Thesis, Leiden University, The Netherlands

Author: Siuman Chung
Print: ProefschriftMaken || www.proefschriftmaken.nl
Cover: Mocking Grafisch Ontwerp, Zaltbommel

© 2018, Siuman Chung
All rights reserved. No parts of this publication may be reproduced, stored in a retrieval system, or transmitted in any form or by any means, mechanically, by photocopy, by recording, or otherwise, without the prior permission from the author.

CBM Progress Monitoring in Reading and Foreign-Language Learning for Secondary-School Students

Proefschrift

ter verkrijging van

de graad van Doctor aan de Universiteit Leiden,

op gezag van Rector Magnificus prof. mr. C. J. J. M. Stolker,

volgens besluit van het College voor Promoties

te verdedigen op dinsdag 26 juni 2018

klokke 13.45 uur

door

Siuman Chung

geboren te Haarlem

in 1984

Promotiecommissie

Promotor:
Prof. dr. C. A. Espin

Co-promotor:
Dr. C. E. Stevenson (Universiteit van Amsterdam)

Overige leden:
Prof. dr. W. C. M. Resing
Prof. dr. P. C. J. Segers (Radboud Universiteit Nijmegen)
Prof. dr. A. E. Souvignier (University of Münster)

Contents

Chapter 1	General introduction	7
Chapter 2	Curriculum-based measurement for secondary-school students	15
Chapter 3	CBM progress monitoring in foreign-language learning for secondary-school students: Technical adequacy of different measures and scoring procedures	27
Chapter 4	CBM maze-scores as indicators of reading level and growth for seventh-grade students	49
Chapter 5	Stability of maze reading slopes: Implications for use of CBM maze scores for tracking the progress of secondary-school students	69
Chapter 6	Summary and general discussion	85
	Nederlandse samenvatting (Dutch summary)	95
	References	103
	Dankwoord (Acknowledgements in Dutch)	117
	About the author	121
	Curriculum Vitae	123
	List of publications	125

Chapter 1

General Introduction

Introduction

Educators have the important task of providing students with optimal learning conditions so that students can reach their full potential with respect to academic performance. A key element for promoting optimal student performance is implementation of evidence-based interventions; yet, implementation of evidence-based interventions does not guarantee success for *all* students because students respond differentially to interventions.

Deno (1990) addressed the issue of individual differences in responses to interventions, where he contrasted nomothetic and idiographic approaches to special education. A nomothetic approach was a search for generalized laws about human behavior and was characterized by educational researchers' efforts to test the effectiveness of interventions via well-designed randomized controlled trials. An idiographic approach, in contrast, was the search for what worked for individuals, and was characterized by educators' efforts to search for interventions that work for individual children. Deno (1990) argued that both nomothetic and idiographic approaches were important to special education: Thus, although it was important to implement interventions with scientific evidence for effectiveness, it was equally important to recognize that such interventions would not be effective (or equally effective) for all students. Deno (1990) further argued that it was not possible to identify *a priori* how an individual student would respond to a given intervention. He thus rejected the then popular *diagnostic-prescriptive approach* in which a diagnosis was first made in order to identify an appropriate intervention for a given student. He proposed an alternative approach in which interventions were viewed as instructional hypotheses whose effectiveness for individuals were tested via implementation of reliable and valid measures of an individual's learning growth. This approach has come to be known as a *problem-solving approach*.

In the problem-solving approach, it is not the diagnosis, but the student's response to interventions, that determines whether the interventions are appropriate for that student. When the student does not respond to an intervention – that is, when the students is not making sufficient academic improvement – the instructional program is changed. An essential element of a problem-solving approach is the use of data to evaluate the effectiveness of the instructional programs on students' academic growth and to make informed decisions about the need to continue or change those instructional programs. The use of data to inform instructional decisions is also reflected in the educational system in the Netherlands, and referred to as *opbrengstgericht werken* (Dutch Inspectorate of Education, 2015) or data-based decision making (DBDM, Mandinach, 2012). DBDM has been seen by the Dutch Ministry of Education, Culture, and Science, as one of the key elements in improving the quality of students' academic achievement (Dutch Ministry of Education, Culture and Science, 2011). One data-based decision making system often used within a problem-solving approach is Curriculum-Based Measurement (CBM; see Deno, 1985, 2003).

Curriculum-Based Measurement (CBM)

CBM is a procedure designed for use by educators to monitor the growth of students with learning difficulties, to evaluate the effectiveness of instruction on that growth, and ultimately, to make decisions regarding the need to continue or change the instruction. CBM involves frequent (for example, weekly) measurement via alternate-forms on measures in an academic area. For example, a CBM measure typically used in the area of reading is maze selection. A CBM maze-selection is a text in which every 7th word is deleted and replaced with three choices: the correct word and two distractors. Students read silently for 2 minutes, selecting the correct word as they read. The number of correct selections made in 2 minutes serves as a general indicator of the student's level of performance and growth in reading proficiency (see Wayman, Wallace, Wiley, Tichá, & Espin, 2007, and Chapter 2 of this dissertation).

CBM scores are graphically depicted in a progress graph (see Figure 1.1) in which a goal line and long-range goal is drawn. Student's growth is represented by a linear growth line (slope). The educator compares the slope to the goal line to determine whether student's scores from CBM measures produce sufficient growth or improvement over time. If growth is sufficient – if the slope is steeper than the goal line – the goal is raised. If growth is insufficient – if the slope is below or less steep compared to the goal line – the instruction is modified or changed, and the effects of the change are then evaluated via continued data collection. This recursive process of data collection, evaluation, data collection, evaluation is designed to provide educators the means to determine the effectiveness of instruction and to provide them with a tool to build powerful, effective interventions for students who struggle.

Essential Characteristics of CBM and Vital Signs

Deno (1985) specified a set of essential characteristics for CBM measures. First and foremost, scores from the measures had to be reliable and valid *indicators* of general proficiency in an academic area such as reading. Second, the measures had to allow for repeated, frequent measurement; that is, it had to be possible to produce multiple, equivalent forms of the measures so that students could be monitored on a frequent basis, and so that growth could be attributed to learning rather than to form differences. Third, the measures had to be simple, efficient, and inexpensive if they were to be used by educators to frequently monitor student growth. Finally, scores from the measures had to be easy to understand if the results were to be clearly and correctly communicated to parents, teachers and students.

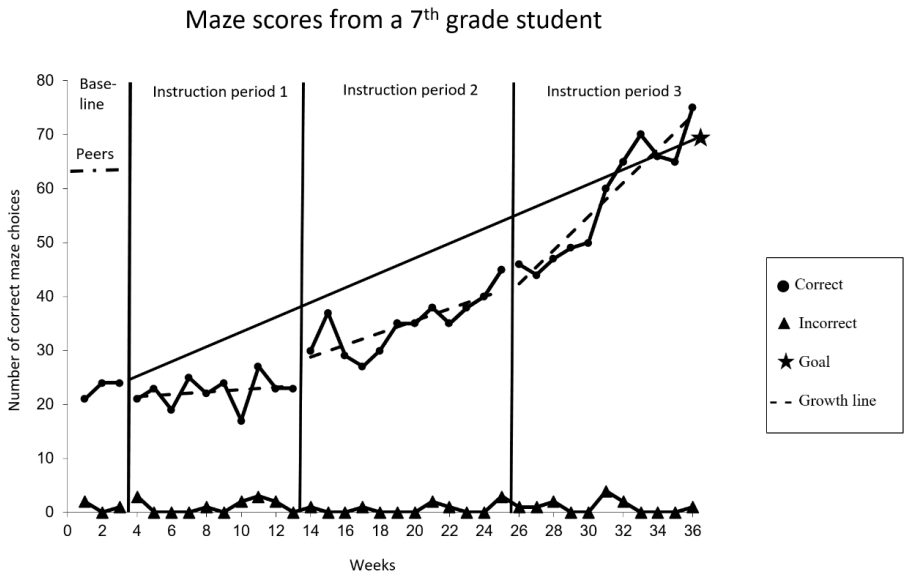


Figure 1.1. Sample of a CBM graph

An important theoretical concept underlying the development of CBM measures is the concept of *indicators* or *vital signs* of academic performance (Deno, 1985; L. S. Fuchs & Deno, 1994). Just as measures such as blood pressure or cholesterol levels are used as indicators or vital sign of general health, scores on CBM measures are designed to be used as indicators or vital signs of a student's academic health. As such, a key question related to the validity of scores from CBM measures is to what extent the scores serve as *indicators* of *general proficiency* in an academic area. If we consider the CBM maze-task described earlier, the key research question with regard to validity is whether CBM maze-scores reflect students' level of performance and growth in general reading proficiency, not whether the scores reflect level of performance and growth in reading fluency or reading comprehension.

There is often confusion surrounding the concept of *indicator*. Many practitioners, and even researchers, incorrectly view CBM measures as tasks designed to measure subskills of reading such as fluency or comprehension, rather than as indicators of general reading proficiency (see discussion in Muijselaar, Kendeou, de Jong, & van den Broek, 2017). However, if scores from the CBM measures are to be used to reflect improvement in an academic area (such as reading) over time, then the evidence must support their validity as indicators of level of performance and growth in that academic area in general. Validity is supported when the level of performance and growth on CBM scores show an expected pattern of relations with performance levels and growth on scores from other measures in the academic area (Cronbach & Meehl, 1955; Espin & Deno, 2016; Messick, 1989).

Gaps in the CBM Research

Over the past 40 years, a large body of research has been conducted on the development of CBM measures in areas such as reading, writing, mathematics, and content-area learning (see reviews by Foegen, Jiban, & Deno, 2007; McMaster, & Espin, 2007; Wayman, et al., 2007). Despite the large research base, there have been noticeable gaps in the research. First, the majority of research on CBM has been conducted at the elementary-school level, with far less research at the secondary-school level (see Espin, Chung, Foegen, & Campbell, in press, or see also Chapter 2; Wayman et al., 2007). Second, the majority of research has focused on what L. S. Fuchs (2004) refers to as the first stage in CBM research, namely the technical adequacy of scores as 'static' measures, where measures are administered at one point in time in order to obtain a rank ordering of student performance level. Much less research has been conducted at Stages 2 or 3. Stage 2 focuses on the technical features of CBM slope/growth trajectories, and Stage 3, on the instructional utility of scores from CBM measures (L. S. Fuchs, 2004). There is a need for additional research in CBM at the secondary-school level, and for research at Stages 2 and 3 (L. S. Fuchs, 2004).

Outline of the Dissertation

The research presented in this dissertation focuses on the development of CBM at the secondary-school level in reading and foreign-language learning, and focuses on both Stages 1 and 2 in CBM research. The focus of the research is the technical adequacy of the scores from CBM measures as indicators of both performance level and growth at the secondary-school level. In Chapter 2, the current research on CBM at the secondary-school level is reviewed. In Chapters 3 through 5, three empirical studies are presented addressing the technical adequacy of scores from CBM measures in reading and foreign-language learning at the secondary-school level.

In **Chapter 2**, a review of the CBM research at the secondary-school level in reading and content-area learning is presented. A description of the CBM measures typically used in reading and the content-area is provided, and research on the technical adequacy of the scores from these measures is presented. The chapter highlights the fact that CBM research at secondary-school level is sparse, especially with regard to studies that included frequent, repeated measurement over the course of several months, perhaps because such studies are time and resource intensive.

Chapter 3 focuses on the development of CBM measures in a content area not previously addressed in the CBM research, foreign-language learning. The technical adequacy from scores on two particular CBM measures are examined: maze selection and word translation. The study examines the extent to which scores from these measures serve as indicators of performance level in foreign-language learning. Various scoring procedures and administration times are compared to determine the most efficient procedure for future research. This study represents Stage 1 in CBM research, which is appropriate given the fact

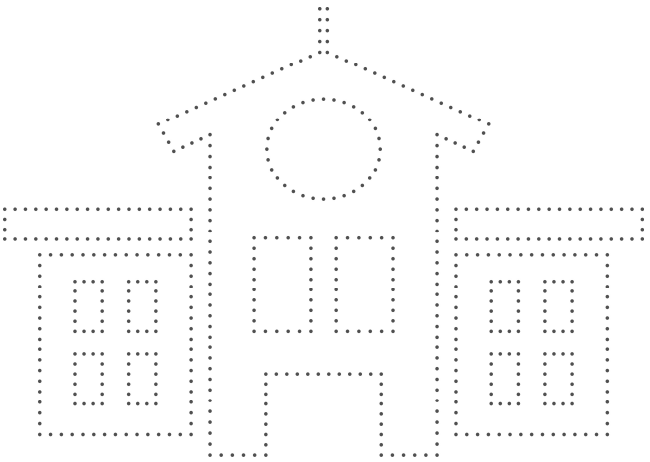
that the study is one of the first to be done in the area of foreign-language learning. Via the results of such an initial study, promising measures are identified that can then be examined in more time and resource intensive growth studies.

In **Chapters 4 and 5**, our attention turns to reading, specifically, to using CBM to monitor growth in reading for secondary-school students. The studies in these chapters represent Stage 2 in CBM research. The focus of the studies is on scores from 2-min maze measures administered to students on a weekly basis for a period of 23 weeks. In **Chapter 4**, alternate-form reliability, sensitivity, and validity of the scores as indicators of reading performance level and growth are examined. Both linear and nonlinear growth trajectories are considered. In **Chapter 5**, the stability of slopes generated from CBM maze scores, and factors related to the stability, are examined. The factors examined include duration, data collection schedule, and variation in maze scores. The results of Chapter 5 are especially important related to the use of CBM for instructional decision-making.

Finally, in **Chapter 6**, the results of the studies are summarized and discussed within the broader context of CBM research, and future directions are described.

Chapter 2

Curriculum-based measurement for secondary-school students



Partly based on:

Christine Espin

Siuman Chung

Anne Foegen

Heather Campbell

In: P. C. Pullen, & M. J. Kennedy (Eds.), Handbook of Response to Intervention and Multi-Tiered Systems of Support (in press). London: Taylor & Francis/Routledge.

Introduction

Curriculum-Based Measurement (CBM; Deno, 1985) is an ongoing progress monitoring system designed to provide educators with data that can be used to evaluate the effectiveness of instructional programs for individual students. Originally, CBM was designed to monitor the progress of students with severe and persistent learning difficulties (see Jenkins & Fuchs, 2012). Interventions for these students were seen as "instructional hypotheses" whose effectiveness was tested empirically via collection and inspection of progress data (Deno & Mirkin, 1977). This approach was referred to as a *problem-solving approach* (Deno, 2013; Deno & Fuchs, 1987).

The majority of work to date on CBM has focused on elementary-school students (see for example Wayman et al., 2007), perhaps because the needs of these students are less complex, and the goals of instruction are more clear, than for secondary-school students. However, in recent years, attention has turned to the development of CBM measures for screening and progress monitoring for secondary-school students.

Development of CBM at the Secondary-school Level: Unique Challenges

The development of CBM measures at the secondary-school level has presented unique challenges (Espin & Campbell, 2012; Espin & Tindal, 1998). One of these has been to decide *what to measure*. Decisions about what to measure depend on how "curriculum" is defined for secondary-school students with learning disabilities (LD). For example, if curriculum is defined as the continued development of basic reading and writing skills, then CBM measures must be designed to measure progress in reading and writing. If curriculum is defined as the acquisition of content-area knowledge, then CBM measures must be designed to measure progress in content areas such as history and science. In this chapter, we outline research and development of CBM progress measures in both basic skills and content areas.

A second and related challenge has been to determine *what the long-range goals* for secondary-school students with LD should be. That is, what level of reading or writing proficiency should students with LD be expected to achieve, or how much content-area knowledge should they be expected to acquire? The answers to these questions guide and direct the development and implementation of CBM measures, and the use of the scores for instructional decision-making at the secondary-school level. More research is needed on determining appropriate long-range learning goals for secondary-school students with severe and persistent learning difficulties.

Overview of the Chapter

In the following chapter, we outline research and development of CBM progress monitoring at the secondary-school level in reading and content-area learning (social studies and

science). For each area, we first describe what and how to measure, then briefly summarize the research done in the area, and finally outline future directions for research. Although the measures and procedures for administering and scoring CBM measures often differ from those used at the elementary-school level, the essence of CBM remains the same at both elementary- and secondary-school level – that is, the use of progress data to empirically test the effectiveness of interventions for students who struggle.

Reading

At the secondary-school level, the reading demands for students increase significantly, with an increase in reading complexity (Seifert & Espin, 2012), and a corresponding decrease in formal reading instruction (see NAEP data, U. S. Department of Education, 2016a), creating significant challenges for students who struggle with reading. Students with severe and persistent reading difficulties might benefit from continued, systematic, intensive reading instruction throughout their middle- and high-school years (see, for example, Edmonds et al., 2009; Solis, Miciak, Vaughn, & Fletcher, 2014), and from implementation of CBM progress monitoring to evaluate the effectiveness of such instruction.

What and How to Measure in Reading

Two types of reading CBM measures have been examined at the secondary-school level: reading aloud and maze selection (Wayman et al., 2007). For reading aloud, students read aloud from a passage, and the number of words read correctly is scored. The examiner's copy of the passage includes a cumulative count of the number of words per line to aide in scoring. The time given is usually 1 min, although time varies across studies. Reading aloud is administered 1:1; thus, it is time consuming for monitoring large number of students.

For maze selection, students read silently from a passage in which every seventh word is deleted and replaced with a multiple-choice item consisting of the correct answer and two distractors. Students select words that restore meaning to the story in the text. The time given is usually 2 to 3 min, although, again, time varies across studies. Either correct or correct-minus-incorrect choices is scored. To control for guessing, scoring stops after three consecutive incorrect choices.[1]

Rules for selecting the distractor items for the maze vary somewhat, but usually consist of selecting words that are within one letter in length of the correct word and are clearly wrong choices – that is, are semantically different from, do not rhyme with, and have a different sound and letter configuration than the correct word (Espin, Deno, Maruyama, & Cohen, 1989; L. S. Fuchs & Fuchs, 1992). Because maze passages are read silently, the

[1] See Hale et al. (2011) and McCane-Bowling, Strait, Guess, Wiedo, & Muncie (2014) for examples of alternative methods for creating and scoring the maze.

measure can be administered to students in groups, and can be administered and scored electronically.

Both narrative and informative texts have been used to develop CBM reading passages. If informative texts are used, it is important that they do not require detailed background information to read and understand the text. Text length varies between 300 to 800 words (depending on the time given to complete the task). Several commercial sites have reading-aloud and/or maze passages available for up to grades 8 or 9 (*AIMSweb*, http://www.aimsweb.com; *EasyCBM*, https://www.easycbm.com; *DIBELS*, https://dibels.uoregon.edu). In addition, via *Intervention Central,* one can create maze passages from any text (http://www.interventioncentral.org).

Research on CBM in Reading

Studies have examined the technical adequacy of scores from reading aloud and maze selection as both performance level (screening) and growth (progress) measures, although the majority of studies have focused on performance level. Results have varied substantially across studies.

Performance level

To use CBM measures as performance level or screening measures in reading, the scores must reliably and validly rank-order students at a single point in time. Thus scores must have good alternate-form reliability, must positively relate to scores from criterion reading measures, and must accurately classify students as proficient or non-proficient (at-risk) in reading. In recent years, a number of studies have examined the technical adequacy of scores from reading-aloud and maze-selection measures as indicators of performance level in reading (Baker et al., 2015; Barth et al., 2012; Barth et al., 2014; Codding, Petscher, & Truckenmiller, 2015; Decker, Hixson, Shaw, & Johnson, 2014; Denton, et al., 2011; Espin, Wallace, Lembke, Campbell, & Long, 2010; Kim, Petscher, & Foorman, 2015; McMaster, Wayman, & Cao, 2006; Silberglitt, Burns, Madyun, & Lail, 2006; Tichá, Espin, & Wayman, 2009; Tolar et al., 2012; Yeo, Fearrington, & Christ, 2012; Yovanoff, Duesbery, Alonzo, & Tindal, 2005). We summarize the results of this research below.

Alternate-form reliability

Alternate-form reliability coefficients for reading-aloud scores have ranged from r = .75 to .97 and for maze-selection scores have ranged from r = .52 to .96, with a majority of coefficients above r = .70. Not surprising, obtained reliability coefficients have been lower when the interval between repeated testing is longer (2 to 6 months; r = .52 to .79, Tolar et al., 2012; Yeo et al., 2012) than when it is shorter (same day or week; r = .69 to .96, Espin et al., 2010; Tichá et al., 2009). Related to probe duration, obtained reliability coefficients for reading-aloud scores have been similar across time frames (1- vs. 2- vs. 3-min vs. untimed;

Barth et al., 2014; Espin et al., 2010; Tichá et al., 2009); however, for maze-selection scores, a small increase in obtained coefficients has been seen with increased time (2- vs. 3- vs. 4-min), although the significance of the differences has not been tested (Espin et al., 2010; Tichá et al., 2009).

Validity

To examine evidence for validity, correlations between CBM scores and scores on criterion measures – including a variety of standardized and state-standards tests in reading – have been examined. Correlation coefficients between CBM and criterion-measure scores have varied widely across studies, ranging from $r = .32$ to .89 for reading aloud and from $r = .37$ to .88 for maze selection. Coefficients have been found to be similar across administration time (for example 1- vs. 2- vs. 3-min) and scoring procedure (for example, correct vs. correct-minus-incorrect choices, Barth et al., 2014; Espin et al., 2010; Tichá et al., 2009).

A small number of studies have examined the use of the CBM scores for predicting which students are "at-risk" and in need of intensive reading instruction. In many of these studies, a measure of predictive power called Area Under the Curve (AUC) has been calculated (but see, for example, Stevenson, 2015, for a different approach). AUC refers to the chance that a student is correctly classified as a proficient or non-proficient reader. AUC values range from .50 to 1.00, with .90 to 1.00 considered to be excellent, .80 to .89 considered to be good, and .70 to .79 considered to be poor (Christ, Zopluoglu, Monaghen, & van Norman, 2013). Most studies have used scores on state-standards tests as the criterion. For reading aloud, AUC values have ranged from .58 to .87 and for maze selection from .59 to .80 (Barth et al., 2014; Decker et al., 2014; Denton et al., 2011).

Summary of research on performance level

In sum, scores from both reading-aloud and maze-selection measures have reasonable alternate-form reliability. There is some evidence for validity, however results are variable. This variability might be due to differences across studies related to factors such as the passages used, the participants included, the setting (school/district), the number of passages administered per assessment occasion, scoring procedures, etc. The influence of such factors on the validity of CBM reading measures has been investigated within studies (see Baker et al., 2015; Barth et al., 2012; Codding et al., 2015; Decker et al., 2014; Espin et al., 2010; Kim et al., 2015; Silberglitt et al., 2006; Tichá et al., 2009; Tolar et al., 2012), but there is a need to examine whether such factors account for differences in results between studies.

Growth

For measuring growth or progress in reading, it is important that scores from CBM measures accurately reflect change in reading. Thus, the scores must be sensitive to growth at a group

level, sensitive to interindividual differences in growth, and there must be a positive relation between change in CBM scores and change in scores on criterion measures.

For many progress studies, growth has been based on a small number of data points. For example, Codding et al. (2015) found that the number of correct maze choices on one 3-min maze passage administered in the fall, winter, and spring of the school year to 247 7th-graders reflected significant growth over time and interindividual differences in growth trajectories. Yeo et al. (2012) found that three 1-min reading-aloud and one 3-min maze-selection probe administered to 261 7th- and 225 8th-grade students in the fall, winter, and spring of the school year did not contribute to the prediction of performance on a statewide reading assessment after controlling for initial status. In addition, Yeo et al (2012) found that growth on the two measures was not correlated.

Tolar, Barth, Fletcher, Francis, and Vaughn (2014) and Tolar et al. (2012) measured students five times across the school year. They administered three to five 1-min reading-aloud passages and one 3-min maze-selection passage per measurement occasion to 1,343 students in grades 6-8. Reading-aloud scores were the mean number of words read correct across the three to five passages per occasion, and maze-selection scores were the number of correct-minus-incorrect choices. In general, results revealed that scores for both reading aloud and maze selection reflected growth over time, but for neither was change in scores related to growth on the criterion measures (Tolar et al., 2014; Tolar et al., 2012), although change on maze-selection scores was related to performance on a reading measure given at the start of the study (Tolar et al., 2012). McMaster et al. (2006) also measured students five times, but over a period of 13 weeks. Reading-aloud and maze-selection measures were administered to 25 English Learners (ELs) once every three weeks. Both reading-aloud and maze-selection scores reflected statistically significant growth over the 13 weeks, but standard error of estimates for both measures were large relatively to the slopes.

Only two studies involved repeated, weekly data collection, and these studies were of short duration. In Espin et al. (2010) and Tichá et al. (2009), 236 and 35 8th-grade students (respectively) completed weekly reading-aloud and maze-selection probes over a period of 10 weeks. The reading-aloud and maze-selection probes were created from the same materials, controlling for potential differences in outcomes due to passage content. Results revealed that for reading aloud, there was minimal or no improvement in scores across the 10 weeks, regardless of time frame (1-, 2-, or 3-min) or scoring procedure (total words read vs. words read correctly). For maze selection, improvement in scores was significantly different from 0 at all time frames (2-, 3-, and 4-min), and improvement in maze-selection scores was related to scores on a state standards test in reading (Espin et al., 2010), to placement in reading groups, and to improvement in scores on a standardized test of reading (Tichá et al., 2009). The pattern of results did not differ across maze scoring procedures (correct vs. correct-minus-incorrect choices).

Summary of research on growth

In sum, there is some evidence that scores on CBM measures reflect change over time, but it is unclear whether such changes are valid reflections of growth in reading. One factor that may affect the validity of the growth trajectories produced by scores on the CBM measures is the error associated with scores on "parallel" forms. That is, there is a great deal of variability or error associated with scores across different passages, despite the fact the passages are designed to be equivalent. The problem with passage effect has led to recent recommendations to use equated rather than raw scores in CBM progress monitoring (Betts, Pickart, Heistad, 2009; McMaster et al., 2006; Tolar et al., 2014; Tolar et al., 2012). Other factors that may affect the validity of the growth trajectories produced by CBM scores include the frequency and schedule of data collection (see Jenkins, Graff, & Miclioretti, 2009), the effects of instruction (Tolar et al., 2014; Tolar et al., 2012), the application of linear vs. nonlinear growth models (Nese et al., 2013; Tolar et al., 2014), and the participants' level of reading and/or English-language ability (see McMaster et al., 2006; Tolar et al., 2014; Tolar et al., 2012).

Summary and Future Directions in Reading

Perhaps the most striking outcome of both the performance level and progress research in CBM reading is the inconsistency in results across studies. There is a need for additional research and/or systematic reviews and meta-analyses to examine factors that account for the inconsistency in results. Several potential factors have been mentioned in the previous section. These factors can be categorized as factors related to participant characteristics, measure characteristics, scoring approach, data-collection schedule, effects of instruction on growth, criterion measures selected, and the growth model applied to the data (linear vs. nonlinear).

There is also a pressing need for more research on the technical adequacy of the scores as growth indicators. Specifically, there is a need for research on the technical adequacy of the growth rates produced by scores from repeated, frequent (weekly) administration of the CBM measures. Such research is time-intensive and expensive because it requires a large number of students that need to be monitored over an extended period of time, but the research is critical for the use of CBM measures within a problem-solving approach. In such an approach, CBM measures are administered frequently (weekly or bi-weekly) so that the growth data can be used to evaluate the effectiveness of instruction and make instructional decisions in a timely fashion. Research is needed to address questions related to the effects of schedules and frequency of data collection, the error due to form effects, and the use of linear vs. nonlinear growth models, on the technical adequacy of growth rates produced by CBM measures. There is also a need for more research focused on older students. To date, the majority of research at the secondary-school level has

focused on students in grades 6 to 8. More research is needed focused on students in grades 9 to 12.

Finally, there is a need to reflect upon the issue of what is being measured with CBM measures at the secondary-school level, and upon the nature of the research questions addressed at the secondary-school level. With regard to the issue of "what" is being measured, at the secondary-school level, maze often is referred to as a measure of "reading comprehension", however, within a CBM approach, both maze selection and reading aloud are designed to produce scores that serve as *indicators of general reading proficiency*, not as measures of fluency and/or comprehension (see Conoyer et al., 2017, for a more in-depth discussion of this point). Such a distinction is important in terms of both formulating research questions and using and interpreting CBM scores for decision-making.

With regard to the nature of the research questions, we wonder how important screening is at the secondary-school level. A proportionally large number of CBM studies in reading focus on use of the scores as performance or screening measures – but to what extent is it necessary to screen secondary-school students for reading difficulties? By the time students reach 7th or 8th grade, is it not clear, based on existing information, which students have reading difficulties? Do scores from CBM measures add anything to the accuracy of identification of students with reading difficulties? As an example, Denton et al. (2011) found that the best predictor of whether students would pass or fail the state standards test in reading was the previous score on the state standards test; scores from CBM measures did not improve prediction accuracy. Future research should consider whether the cost of testing a large number of students (especially with reading aloud) three times a year outweighs the benefits associated with use of the scores to identify students at-risk.

Content-Area Learning

Students with learning difficulties often receive their content-area instruction in the general education classrooms, and are held to the same standards as peers without disabilities. Yet learning in the content areas presents significant challenges for many students with learning difficulties, not the least of which is reading large amounts of text, often from textbooks that are "inconsiderate" – that is, not well-structured, replete with specific vocabulary terms, and dense with new ideas and concepts (Armbruster & Anderson, 1988; Groves, 1995; Yager, 1983). Students with disabilities may require additional supports/interventions to succeed in their content-area classes. CBM progress-monitoring can be used to help determine whether such supports/interventions are effective in helping students to learn content-area material, to achieve content-area standards, and to attain passing grades on state standards tests.

What and How to Measure in the Content Areas

Various types of CBM content-area measures have been examined over the years, including reading aloud (see Espin & Deno, 1993a, 1993b; Espin & Foegen, 1996; Fewster & MacMillan, 2002), story writing (see Fewster & MacMillan, 2002), and maze selection (see Espin & Foegen, 1996; Johnson, Semmelroth, Allison, & Fritsch, 2013; Ketterlin-Geller, McCoy, Twyman, & Tindal, 2006); however, most research has employed a vocabulary-based type of measure. The focus on vocabulary is perhaps not surprising given the importance of *academic* or *key vocabulary* in the content-areas. Academic or key vocabulary refers to words or terms that represent specific concepts and processes within an academic area (Antonacci, O'Callaghan, & Berkowitz, 2015; Vannest, Parker, & Dyer, 2011). These concepts and terms form the basis for learning in the content areas (Vannest et al., 2011). The vocabulary-based tasks used in CBM typically involve matching key vocabulary terms with definitions; however, the specific form of the task varies from study to study. For simplicity's sake, we outline one commonly-used approach in this section, namely vocabulary matching (see Busch & Espin, 2003, for a detailed description of how to create and use vocabulary-matching measures).

The first step in the development of the vocabulary-matching measure is to create a pool of key terms and definitions from the glossaries of the textbooks and with teacher input. From this pool, 20 terms and definitions are randomly selected to appear on each probe. The 20 terms are listed in alphabetical order vertically on the left side of the page, and 20 definitions plus 2 distractor definitions are listed in random order on the right. Students have 5 min to match terms with definition. Scores are the number of correct matches.

Probes must be long enough so that students do *not* finish within the allotted time. The growth on the CBM measures is reflected in both accuracy and automaticity; thus, as students increase content-area knowledge, they more accurately and efficiently read and define key vocabulary terms. If students finish the probes before the allotted time, there is little room for growth on subsequent probes. The procedures outlined in the previous paragraph should thus be viewed as guidelines: It may be necessary to place more than 20 terms on each probe or to provide less time for measurement in order for the measures to be sensitive to growth across an entire school year.

Research on CBM in the Content Areas

Research on CBM in the content areas has been conducted in various areas, including social studies, science, and foreign-language learning. Espin and Tindal (1998) reviewed the early work on the development of CBM measures in the content areas, thus we focus on research conducted after the time of that review, and on research conducted in the areas of social studies (including history) and science because the majority of work has been conducted in those areas. Finally, we focus on research conducted with students in grades 6 and above

(see Mooney, McCarter, Russo, & Blackwood, 2013, and Vannest et al., 2011, for studies with participants in grade 5).

In general the research has supported the reliability and validity of scores from vocabulary-based CBM measures as indicators of both performance level and growth in the content areas. In both social studies and science, alternate-form (AF) reliabilities have varied from .58 to .87, with averages in the .70s. Combining scores across two probes has resulted in higher AF reliabilities than for single probes (r = .80 and above, Beyers, Lembke, & Curs, 2013; Espin et al., 2013; Espin, Busch, Shin, & Kruschwitz, 2001). Espin et al. (2013) found that reliabilities of adjacent scores on weekly science probes increased across time, perhaps because the scores became more stable as students accumulated science knowledge over time. Espin et al. (2001) found a similar increase in AF reliabilities in social studies from weeks 1 to 8, but reliabilities then decreased from weeks 9 to 11.

With regards to the validity of CBM vocabulary scores as indicators of content-area performance, correlations with scores on state standards tests and commercial standardized tests in social studies and science typically have ranged from .53 to .76, with the majority of correlations between .60 and .65 (Beyers et al., 2013; Espin et al., 2013; Espin, et al., 2001; Mooney, McCarter, Schraven, & Callicoatte, 2013; Mooney, McCarter, Schraven, & Haydel, 2010). Coefficients have been found to be similar across race, gender, SES, and exceptionality (Mooney, McCarter, Schraven, et al., 2013; Mooney et al., 2010). Scores on the CBM vocabulary measures also have been found to relate to scores on researcher-made knowledge tests, with concurrent validity coefficients ranging from .59 to .84, and predictive validity coefficients from .66 to .67 (Espin et al., 2013; Espin et al., 2001). Finally, correlations with course grades have been found to be significant, but have been somewhat lower than correlations with the standardized or research-made instruments, ranging from .57 to .65 in science (Espin et al., 2013) and .27 to .51 in social studies (Espin et al., 2001).

With regard to the sensitivity of the measures to progress or learning in content-areas and the validity of the growth rates produced by the measures, growth on CBM vocabulary measures has been shown to be significantly different from zero, to reflect interindividual differences in growth rates, and to be significantly related to disability status, SES, scores on standardized/state tests, course grades, and improvement in scores on research-made knowledge tests (Beyers et al., 2013; Borsuk, 2010; Espin et al., 2013; Espin, Shin, & Busch, 2005; Mooney, McCarter, Schraven, et al., 2013). The amount of observed growth has varied from study to study. Espin et al. (2013) and Espin et al. (2005) found growth rates of .63 matches per week in science and .65 in social studies respectively, whereas Beyers et al. (2013) found growth rates of only .15 matches per week in social studies.

Growth rates may vary with the type of probe used. For example, Mooney, McCarter, Schraven, et al. (2013) found that growth was higher for probes in which terms were randomly selected from the entire pool of terms (.23 high SES and .10 low SES) than when probes included equal number of terms from the first and second half of the year (.11 high

SES, .02 low SES). Espin et al. (2005) found lower growth rates when probes were read to the students than when the students read the probes themselves (.22 vs .65 matches per week). Borsuk (2010) found growth rates of .26 for probes that were read to the students, and for which students were presented a definition and asked to select the correct term from 6 options. Reasons for differences in growth rates across studies need to be systematically addressed in future research.

Summary and Future Directions

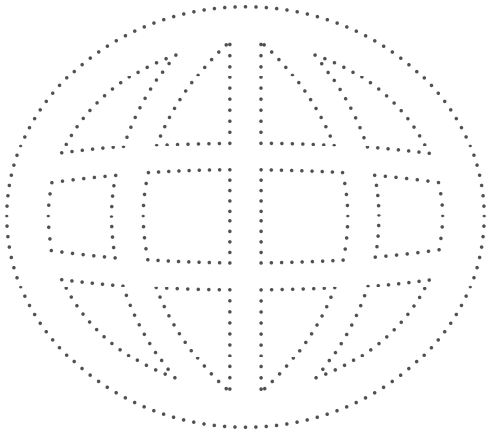
In sum, research in the content areas has supported the technical adequacy of CBM vocabulary-based measures as indicators of performance level and growth in social studies and science; however, several questions must be addressed in future research. As stated earlier, reasons for the differences in growth rates among studies should be systematically examined. Relatedly, different approaches for creating probes should be systematically compared, as was done in Espin et al. (2005) and Mooney, McCarter, Schraven, et al. (2013). Such comparisons could examine the effects of different time frames and number of items on the technical adequacy of the measures. This research also could examine administration of the measures across an entire school year to determine whether there are floor or ceiling effects for the scores.

Conclusion

In recent years, much research has been done on the development of CBM measures for secondary-school students in reading and content-area learning. Within both areas, we have provided a summary of the research and suggestions for future directions. In this final section, we comment briefly on two themes that cut across the areas. The first is the need for more research on the use of the measures as progress measures. Such research is critical if the measures are to be used within a problem-solving approach. The second theme is the need for more research related to teachers' use of the measures for instructional decision-making. At the elementary-school level, research has been done on teachers' use of CBM progress data for instructional decision-making (see Stecker, Fuchs, & Fuchs, 2005, for a review), but little research has been done on this topic at the secondary-school level. Such research is essential because the answers to data-use questions goes to the heart of CBM.

Chapter 3

CBM progress monitoring in foreign-language learning for secondary-school students: Technical adequacy of different measures and scoring procedures



Siuman Chung

Christine Espin

Assesment for Effective Intervention (2013), *38*, 236-248.

doi: 10.1177/1534508413489723

Abstract

The reliability and validity of three curriculum-based measures as indicators of learning English as a foreign-language were examined. Participants were 260 Dutch students in grades 8 and 9 who were receiving English-language instruction. Predictor measures were maze selection, Dutch-to-English word-translation, and English-to-Dutch word-translation. Criterion variables were years of English instruction, school level, course grades, and scores on a standardized reading test. Different scoring procedures and time frames were compared. Alternate-form reliabilities ranged from .44 to .88. Significant differences in maze scores were found between school level, but not between years of English-language instruction. Correlations between predictor and criterion variables ranged from .19 to .79. A regression analysis revealed that a combination of maze and English-to-Dutch translation predicted English course grades better than a single measure alone.

Introduction

In foreign-language learning, a new or different language is being learned in a context where the mother tongue is regularly spoken (Verspoor, de Bot, & van Rein, 2010). Learning a foreign language is obligatory in many European countries (Eurydice, 2001), where nearly all students, including those with learning disabilities (LD), complete at least one language course prior to graduation from high school. Although in American schools students with LD often can substitute a foreign-language course with a non-language course (Sparks, Javorsky, & Philips, 2005), these students may once again face difficulties in college where foreign-language learning often is compulsory (Skinner & Smith, 2011).

Learning a foreign language can present significant challenges for students with LD because their difficulties typically are language-based (Hallahan, Lloyd, Kauffman, Weiss, & Martinez, 2005), affecting performance in reading, writing, listening, and spelling (Hallahan et al., 2005). Given adequate help, however, students with LD often can succeed in foreign-language learning (Sparks, 2006). The design of specialized foreign-language programs for students with LD would be enhanced if they were to include methods for evaluating the effects of the program on individual student learning. One such method is Curriculum-Based Measurement (CBM).

CBM is a progress monitoring system designed to be used by educators for monitoring student progress and judging the effectiveness of instructional programs (Deno, 1985). To date, there has been little to no work conducted on the development of CBM progress measures in foreign-language learning. A small number of studies have examined the technical adequacy of CBM for students who are learning a second language (for an overview, see Sandberg & Reschly, 2011), but these studies have focused mainly on bilingual students who are learning English as a second language at the elementary-school level and who are living in an English-speaking country. The results of these studies show promising possibilities for bilingual students (Sandberg & Reschly, 2011); however, these positive results cannot be generalized to students learning a foreign language. Students learning a foreign language do not necessarily have the advantage of living in an environment where the foreign language is regularly spoken and they usually learn the language in secondary school or higher education.

The goal of the present study was to develop CBM progress measures for foreign-language learning for secondary-school students. Because little previous research exists to guide the selection and development of measures for foreign-language progress-monitoring, we draw from research on the assessment of foreign-language skills and research on CBM progress-monitoring in native-language reading.

Research on Assessment of Foreign-Language Learning

Research on the assessment of foreign-language learning has focused on various language domains such as vocabulary/word knowledge, reading, listening, speaking, writing, and grammar (Alderson & Banerjee, 2002). Most research on foreign-language assessment has focused on the domains of vocabulary/word knowledge and reading. Other areas of language proficiency have been researched less often, perhaps because it is more difficult to determine relevant features of those domains (Alderson & Banerjee, 2002).

One of the initial steps in learning a foreign language is the development of word knowledge (Wang, 2011). Word-knowledge proficiency has been assessed in several different ways. One type of measure is a word-translation measure, where students translate words from the native to the foreign language or the foreign to the native language. Translation tasks have been found to be more meaningful than other types of vocabulary tasks in the initial period of foreign-language learning because the learner understands foreign words better as translations than as synonyms or descriptions in the foreign language (Nation, 1982).

Once students have a core of vocabulary words available in the foreign language, they can begin reading text (Wallace, 2008). Opinions differ as to precisely how students learn to read a foreign language (for an overview, see Alderson & Banerjee, 2002) and several different measures have been used to assess foreign-language reading skill. One of these measures is a cloze test (Alderson & Banerjee, 2002). In a cloze test, a fixed number of words are deleted and replaced by a blank, and students fill in the blanks as they read (Taylor, 1956). In a modified cloze test, the deleted word is replaced with a multiple-choice item instead of a blank, and students select the correct word as they read. Each multiple-choice item consists of the correct word and a number of distracters. The distracters are semantically or lexically comparable to the target word (Cranney, 1972-73). The modified cloze test has been found to have good reliability and validity as a measure of reading comprehension in English as a foreign language across nine different language groups (Hale et al., 1989). Reliability coefficients ranged from .79 to .89 (although the authors did not specify what type of reliability was calculated). Correlations between the modified cloze score and scores on the reading comprehension subtest of the TOEFL (Test of English as a Foreign Language) ranged from $r = .67$ to .78.

Research on CBM Progress-Monitoring in Reading

In the CBM research, a measure similar to a modified cloze has been used to measure reading progress in students' native language. This measure typically is referred to as a *maze* or *maze-selection* task (see L. S. Fuchs & Fuchs, 1992; Wayman et al., 2007). Although similar to the modified-cloze used in foreign-language assessment, there are a few differences between the two measures. First, the CBM maze-selection task is timed. Students work for a fixed amount of time (e.g., 2 min.) and the number of correct answers selected in that time

is scored. Second, the distractors in the CBM maze-selection task are designed to be clearly incorrect, and are semantically and syntactically different from (rather than similar to as in the modified-cloze) the correct choice (L. S. Fuchs & Fuchs, 1992).

Research has supported the use of a maze-selection measure as an indicator of general reading proficiency in the native language (Wayman et al., 2007). At the secondary-school level, alternate-form reliabilities for maze-selection have ranged from $r = .75$ to .96, and correlations with other measures of reading performance have ranged from $r = .76$ to .88 (Espin et al., 2010; Tichá et al., 2009). In addition, maze-selection has been found to produce reliable and valid growth trajectories over time (Espin et al., 2010; Tichá et al., 2009; Tolar et al., 2012).

Selection of Potential Foreign-Language Progress-Monitoring Measures

Based on the types of measures typically used in the assessment of foreign-language proficiency and in CBM reading progress measurement, we decided to examine maze-selection and word-translation tasks as potential progress measures for foreign-language learning. Because we were interested in developing ongoing progress-monitoring measures, we adopted the maze-selection task used in CBM research rather than the modified-cloze approach used in foreign-language assessment research.

For the word-translation tasks, we included both foreign-to-native and native-to-foreign language translation tasks. A foreign-to-native language translation task requires students to recognize the foreign-language term, and then produce it in their native language. One might expect this task to be easier than the native-to-foreign language task, which requires students to recall and produce the foreign-language terms. The two might function differently at different levels of language proficiency. Given that little research has been done on the development of foreign-language progress measures, we examined differences in reliability and validity for various time frames and scoring procedures.

Purpose and Research Questions

The purpose of this study was to examine the reliability and validity of three potential CBM measures as indicators of foreign-language proficiency: maze-selection, Dutch-to-English word-translation, and English-to-Dutch word-translation. Because this study was one of the first to examine CBM measures for foreign-language learning, we cast our net wide in our first set of analyses, and examined the reliability of different time frames and scoring procedures for each measure. To reduce the overall number of analyses for the subsequent validity analysis, we selected the time frames and scoring procedures that best met the requirements of reliability and efficiency. In our final analyses, we examined whether a combination of maze and word-translation measures predicted English foreign-language proficiency better than a single measure. The following research questions were addressed in the study:

1. What is the alternate-form reliability of maze-selection and word-translation tasks, and does alternate-form reliability differ for different scoring procedures and time frames?
2. What is the validity of the maze-selection and word-translation tasks as indicators of general foreign-language proficiency in English?
3. Does a combination of the maze-selection and word-translation tasks improve the prediction of general foreign-language proficiency in English over the use of either measure alone?

Method

Participants

Participants were 260 (112 male) students in grade 8 (36.9%) and 9 (63.1%) from a secondary school in an urban city in the Netherlands. The mean age of the students was 15 (*SD* = 1.21; range 12 to 19). The birthplace of participants was: The Netherlands (61.8 %), Morocco (7.7%), Turkey (7.3%), Suriname (6.9%), and other countries (16.3%). Place of birth for the parents of the participants was: The Netherlands (9.1%), Morocco (44.7%), Turkey (20.3%), Suriname (8.6%), Netherlands Antilles (5.1%) and other countries or a different country for each parent (12.2%).

Participants were recruited from English foreign-language courses, with a total of 20 classrooms and 20 teachers. Students in 8th grade were in their second year and students in 9th grade in their third year of English language instruction. Secondary schools in the Netherlands are organized into different educational levels, from practical (lowest level) to pre-university (highest level). Different school levels often are housed in different buildings, although it is possible for two to three levels to be combined in one building. Assignment of students to school level is based primarily on students' grades during their elementary school years and on scores on a test given to students at the end of 6th grade. Participants in our study were from four school levels: very low (27.7%), low (32.3%), intermediate (22.7%) and high (17.3%). The level of instruction in English differed per grade and school level.

Predictor Variables

Predictor variables were the three potential foreign-language progress-monitoring measures: maze-selection, Dutch-to-English word-translation, and English-to-Dutch word-translation.

Maze task

The maze task was a reading passage in which the first sentence was left intact and thereafter every seventh word was replaced by a multiple-choice item with one correct answer and two

distractors (L. S. Fuchs & Fuchs, 1992). The maze tasks used in this study were selected from EdCheckup (http://www.edcheckup.com), a system designed to monitor progress in reading for English-speaking elementary-school children. The maze passages were developed from narrative texts and were approximately 265 words in length. Two texts were selected for the study, with topics chosen that were not country specific and the content was appropriate for secondary-school students. The passages were selected from 4th-grade level material because it was thought to be easy enough for beginning language learners, but difficult enough to be sensitive to growth. Each passage contained 28 multiple-choice items. Distractors for the passages were modified if necessary so that they were approximately the same length (plus or minus one letter) as the correct word. Median alternate-form reliabilities for the 4th-grade maze passages of EdCheckup (http://www.edcheckup.com) were reported to be above .70. Correlations between the maze-selection tasks and scores on the Measures of Academic Performance (MAP, Northwest Evaluation Association [NWEA], 2003) were .57 (construct validity) and .53 (predictive validity).

Maze scoring

We examined the effects of different time frames and scoring procedures for the maze task. With regard to time frames, scores for both one and two minutes were examined. With regard to scoring procedures, four different procedures were examined, each combining two methods used to control for error due to guessing in scoring. The first method was a rule vs. no-rule comparison. In the rule condition, a counting rule employed in previous maze-selection research (e.g., Espin et al. 2010; Tichá et al., 2009) was used; after three consecutive incorrect choices, scoring was stopped. In the no-rule condition, this rule was not applied and scores included all correct choices that the student made. The second method was a correct vs. correct-minus-incorrect comparison. In the correct condition, only the number of correct selections was counted. In correct-minus-incorrect condition, the number of incorrect answers was subtracted from the number of correct answers to derive the score. These two approaches were crossed to create four scoring procedures a) rule and correct choices, b) rule and correct-minus-incorrect choices, c) no rule and correct choices, and d) no rule and correct-minus-incorrect choices.

Word-translation tasks

The word-translation tasks required students to translate words from English to Dutch or from Dutch to English. The words represented all parts of speech, where nouns were overrepresented. Fifty words were presented on the left side of a page. Next to each word was a blank for students to write the translated word. Two parallel probes were created for each type of task (English-Dutch or Dutch-English translation). Words for the probes were randomly selected without replacement from the English language curriculum used in the school. Two levels of the word-translation tasks were created: intermediate and high. Levels

were assigned to classes based on teacher recommendation. The 8th-grade very-low school-level students did not complete the word-translation tasks. An overview of task assignments is presented in Table 3.1

Table 3.1
Overview of tasks and English reading test levels administered to students broken down per grade and school level

			Word translation		Reading test level					
Grade level	School level	Maze	Intermediate difficulty	High difficulty	1	2	3	4	5	6
Grade 8	Very low	x			x					
	Low	x	x			x				
	Intermediate	x	x			x				
	High	x		x			x			
Grade 9	Very low	x	x							
	Low	x		x					x	
	Intermediate	x		x					x	
	High	x		x						x

'x' indicates which measures the students received. Empty cells indicate that students did not complete the measure.

Word-translation scoring

As with the maze, we examined different time frames and scoring procedures for the word-translation tasks. With regard to time, scores for both one and two minutes were examined. With regard to scoring procedures, four different procedures were examined, each combining two scoring methods. The first method was spelling vs. no spelling. For the spelling method, the translated word had to be spelled correctly to be counted as correct. For the no-spelling method, the word had to approximate the correct spelling of the word to be counted as correct. If the word was read aloud and it sounded like the correct word, it was marked as correct. The second method was a correct vs. correct-minus-incorrect comparison. In the correct condition, only the number of correct translations was counted. In the correct-minus-incorrect condition, the number of incorrect translations was subtracted from the number correct to derive the score. These two methods were crossed to create four scoring procedures: a) spelling and correct translations, b) spelling and correct-minus-incorrect translations, c) no spelling and correct translations, and d) no spelling and correct-minus-incorrect translations.

Criterion Variables

Years of English instruction and school level

Students were either in their second (grade 8) or third (grade 9) year of English-as-a-foreign-language instruction, and were in one of four different school levels. School level was on an

ordinal scale ranging from very low to high. Students with more years of English instruction or in a higher school level were expected to be, on average, more proficient in English than students with fewer years of English instruction or in a lower school level.

English course grades

Course grades were assigned by the English teacher at the end of the school year. Grades was based on performance on various elements of English language learning: reading, vocabulary, grammar, writing, listening, and pronunciation. Grades ranged from 1 to 10, with 1 being the lowest grade, 6 representing a passing mark, and 10 being the highest grade. Decimal points were possible for grades (e.g., 6.7). Analyses with course grades were conducted within school year (8th and 9th grade) and school level.

Scores on standardized English reading test

At the end of the school year, the English reading subtest of a standardized achievement test, *Cito Volgsysteem Voortgezet Onderwijs* (Cito Progress Monitoring System for Secondary Schools [CITO-VVO]; Cito, 2010), was administered by the school. The subtest consisted of short expository passages and multiple-choice questions with three, four or five possible answers. For each passage, one or two questions were asked. Administration time was approximately 90 minutes with 40 multiple-choice questions in total.

The technical adequacy of the English reading subtest was available only for an earlier version of the test. (Research on the technical adequacy of the newer version was underway at the time of the study.) The newer version was similar to the older version, differing primarily in content. The internal-consistency reliability of the earlier version was reported as high with Cronbach's alpha ranging from .87 to .95. Validity was established by examining differences across grades and school levels. Mean scores were found to increase both by grade level within school and by school level within grade (Nederlands Instituut van Psychologen, 2011).

The English reading subtest consisted of three different levels for each grade. Scores on the English reading subtest were provided to schools as both standard and percentile scores. Only the percentile scores were available from the school; thus, comparisons were done within English reading test level and grade. In Table 3.1 an overview is provided of the English reading test levels for the students who completed the standardized English reading test.

Procedure

The maze and word-translation tasks were administered in March in a group setting by the classroom teachers, who received training in using the correct administration procedures. The training was approximately 1.5 hours, and included a description of the theoretical background of CBM and the progress measures, and practice in implementation. Some

teachers administered both the maze and word-translation tasks in one session; others administered the maze in one session and the word translation in another; however all tasks were administered within the same week. The duration of the administration was approximately 15 to 20 minutes per classroom.

All students first completed the two maze tasks. For each task, they silently read the text, circled choices during reading, made a slash at one minute to mark their progress, and stopped working at two minutes. In the same or second session, the students completed four word-translation tasks. For the first two probes, (English-to-Dutch), students translated words from English to Dutch. For the following two probes (Dutch-to-English) students translated words from Dutch to English. For all four word-translation probes students circled the last word they had read after one minute, and stopped working after two minutes. All tasks were preceded by an example exercise. All students received the same maze tasks, but received the word-translation task at either the intermediate or high difficulty level (see Table 3.1). The order for parallel forms of each task was counterbalanced across classrooms.

To check fidelity in administration, teachers' first administration of all tasks was observed by trained graduate students. Training of the observers took approximately an hour and consisted of a description and demonstration of the administration procedures and a practice observation with the trainer. The students then observed each other administering the task.

In all observed classes, the instructions were read clearly by the teacher and students were judged by the observers to understand the instructions. Five classrooms (20%) were removed from the sample because of incorrect timing. Demographic information for the students and scores on criterion variables were obtained from the school at the end of the school year.

Scoring Agreement

The maze and word-translation tasks were scored by graduate students who were provided approximately 1.5 hours of training. For each type of measure, one probe was scored together, then two probes were scored individually and discussed. Scoring accuracy for the individually scored probes had to be above 90% for the maze and above 80% for the word-translation tasks before scorers could continue scoring. All scorers reached this level of agreement on their first attempt.

During scoring, every 20th maze task for each scorer was double scored. Scoring agreement for the maze was 99.93% (94.1%-100%). For the word-translation task, every 10th task for each scorer was double scored. Scoring agreement for the English-to-Dutch translation was 97.71 (range 89-100%) for spelling and 97.43 (range 85-100%) for no spelling. For Dutch-to-English translation, the scoring agreement was 96.79 (range 89-100%) for spelling and 96.13 (range 82-100%) for no spelling.

Data Analyses

To address alternate-form reliability, Pearson correlations were calculated between the two parallel forms for each measure, time frame, and scoring procedure. For each measure, the time frame and scoring procedure with the highest reliability were used in subsequent validity analysis.

To determine validity, two types of analyses were conducted. First, for the maze, where the same maze passages were used for all participants, a two-way analysis of variance was conducted to examine differences in grade and school level on the maze task. Differences in mean score by grade and school level would support the validity of the measure as an indicator of performance. Second, for all measures, Pearson correlations between progress measures and English course grades and the standardized English reading test were calculated within subgroups. Given that sample sizes were small for this analysis, we viewed the analysis as exploratory, and focused on the general pattern of results and the magnitude of the correlations. Finally, two forward stepwise regression analyses were executed on two subsamples to determine whether the combination of the maze and word-translation tasks would improve the prediction of English reading and language proficiency over the use of a single measure alone.

Results

For the maze task, means and standard deviations were similar across Forms A and B (see Table 3.2). Students made approximately 9 to 9.5 choices per passage in the first minute, and 7 to 8 choices in the second minute. The mean scores were higher for correct than for correct-minus-incorrect scores, with an average of 2.5 to 3 incorrect choices per passage in two minutes in the rule condition and 5 to 5.5 in the no-rule condition. Within the correct and correct-minus-incorrect scoring methods, mean scores for the rule vs. no rule scoring approaches were similar, suggesting that students did not do much guessing while completing the maze task. Ceiling effects were found for 9th-grade students in the high-level school.

Table 3.2
Means and SDs for the Maze scores: Form A, B, and combined score (sum of form A and B)

			Form A		Form B		Combined scores (A+B)	
			1 min	2 min	1 min	2 min	1 min	2 min
Scoring		*n*	*M (SD)*	*M (SD)*	*M (SD)*	*M (SD)*	*M (SD)*	*M (SD)*
Rule	C	260	9.60 (4.23)	16.12 (7.88)	9.39 (4.98)	16.97 (8.63)	18.99 (8.13)	33.09 (15.17)
Rule	C-I	260	7.85 (4.86)	13.07 (8.60)	7.87 (5.50)	14.48 (9.04)	15.71 (9.33)	27.55 (16.63)
No rule	C	259	9.92 (3.92)	17.78 (6.35)	9.94 (4.45)	18.91 (6.41)	19.86 (7.44)	36.68 (11.91)
No rule	C-I	259	7.32 (5.63)	12.32 (9.89)	7.50 (6.08)	14.36 (9.59)	14.82 (10.60)	26.68 (18.38)

'C' is the number of correct choices. 'C-I' is the number of correct-minus-incorrect choices.

Results for the word-translation tasks are reported in Table 3.3 (intermediate difficulty level) and Table 3.4 (high difficulty level). For the intermediate level, scores were somewhat different for Forms A and B, with consistently lower scores seen for Form B. Examination of the combined scores across Forms A and B (last column) reveals that students tended to make more correct translations in the first (approximately 11 to 14) than in the second (approximately 7.5 to 9) minute. Mean scores for the spelling and no-spelling rules were similar in the English-to-Dutch version, revealing that students made few spelling mistakes in their native language; however, in the Dutch-to-English version, observed scores for the spelling and no-spelling methods were different with students making on average 4.65 spelling errors across the two probes in two minutes (mean scores of 18.18 for spelling vs. 22.83 for no spelling). The number of correct translations in two minutes was larger for the English-to-Dutch version than for the Dutch-to-English version (approximately 22 vs. 18), but only when spelling was taken into account. Scores for correct-incorrect were substantially lower than for correct only.

For the high-difficulty level probe, scores tended to be similar across Forms A and B (see Table 3.4). As with the intermediate level, examination of the combined scores across Forms A and B (last column) reveals that students tended to make more correct translations in the first (approximately 7.5 to 13) than in the second (approximately 6 to 7.5) minute. Mean scores tended to be somewhat lower for the spelling than the no-spelling rule, especially for the Dutch-to-English version, where, similar to the intermediate level, students made on average 4.68 spelling errors across the two probes in two minutes (mean scores of 16.12 for spelling vs. 20.80 no spelling). The number of correct translations was lower for the English-to-Dutch version (approximately 13.5 for spelling vs. 16 no spelling) than for the Dutch-to-English version (approximately 16 for spelling vs. 21 no spelling). As with the intermediate level, scores for correct-incorrect were substantially lower than for correct only.

Alternate-Form Reliability

Alternate-form reliability coefficients are reported in Table 3.5. Correlations ranged from r = .44 to .88. Across all measures and scoring approaches, reliability increased with probe duration. Thus in discussing the table, we focus on the coefficients for the 2-min probes. For the maze tasks, reliability coefficients tended to be higher for the correct-minus-incorrect than correct-only scoring procedures, but were similar for the rule vs. no-rule scoring procedures. Coefficients for correct-minus-incorrect scoring procedure were .78 for both rule and no-rule methods.

For the word-translation tasks, similar patterns were found across the two difficulty levels. For both levels, scoring the number of correct translations resulted in stronger reliabilities than scoring correct minus incorrect. Within the correct scoring method, few differences were seen for spelling vs. no-spelling rules. Finally, reliability coefficients tended to be higher for the Dutch-to-English (r = .65 to .88) than the English-to-Dutch versions (r =

Table 3.3
Means and SDs for the word-translation scores, intermediate difficulty level: Form A, B, and combined scores (sum of form A and B)

				Form A		Form B		Combined scores (A+B)	
				1 min	2 min	1 min	2 min	1 min	2 min
Measure	Scoring		*n*	*M (SD)*	*M (SD)*	*M (SD)*	*M (SD)*	*M (SD)*	*M (SD)*
Translation Dutch to English	Spelling	C	92	6.46 (3.80)	9.61 (6.36)	4.35 (3.23)	8.58 (6.01)	10.80 (6.56)	18.18 (11.96)
	Spelling	C-I	92	3.04 (4.88)	3.32 (7.24)	0.12 (4.25)	1.43 (6.85)	3.16 (8.04)	4.75 (12.94)
	No spelling	C	92	7.62 (3.96)	11.58 (6.90)	6.11 (3.57)	11.25 (6.72)	13.73 (7.06)	22.83 (13.21)
	No spelling	C-I	92	5.36 (4.93)	7.21 (7.86)	3.64 (4.14)	6.78 (7.30)	9.00 (8.29)	13.99 (14.29)
Translation English to Dutch	Spelling	C	94	8.32 (3.05)	12.35 (4.90)	5.43 (2.99)	9.59 (5.07)	13.74 (5.48)	21.94 (9.36)
	Spelling	C-I	94	3.53 (3.94)	3.21 (5.84)	0.79 (4.17)	0.63 (6.57)	4.32 (6.89)	3.84 (11.25)
	No spelling	C	94	8.49 (3.05)	12.74 (5.08)	5.64 (3.03)	10.21 (5.39)	14.13 (5.57)	22.96 (9.89)
	No spelling	C-I	94	3.87 (3.92)	4.11 (6.06)	1.21 (4.17)	1.88 (6.91)	5.09 (7.05)	5.99 (11.98)

'C' is the number of correct translations. 'C-I' is the number of correct-minus-incorrect translations.

Table 3.4
Means and SD for the Word-translation scores, high difficulty level: Form A, B, and combined scores (sum of form A and B)

				Form A		Form B		Combined scores (A+B)	
				1 min	2 min	1 min	2 min	1 min	2 min
Measure	Scoring		*n*	*M (SD)*	*M (SD)*	*M (SD)*	*M (SD)*	*M (SD)*	*M (SD)*
Translation Dutch to English	Spelling	C	143	4.58 (2.80)	7.80 (4.86)	5.61 (3.97)	8.32 (6.26)	10.19 (6.38)	16.12 (10.64)
	Spelling	C-I	143	0.38 (3.78)	1.17 (5.36)	1.36 (4.80)	0.90 (7.10)	1.73 (7.72)	2.07 (11.37)
	No spelling	C	143	6.11 (3.02)	10.03 (5.53)	7.31 (4.16)	10.78 (6.67)	13.43 (6.68)	20.80 (11.60)
	No spelling	C-I	143	3.44 (3.67)	5.63 (5.81)	4.77 (4.58)	5.78 (6.87)	8.21 (7.35)	11.41 (11.52)
Translation English to Dutch	Spelling	C	136	3.76 (2.77)	6.33 (4.63)	3.70 (2.64)	7.19 (4.30)	7.46 (4.98)	13.52 (8.46)
	Spelling	C-I	136	-1.42 (3.92)	-2.72 (6.02)	-1.43 (3.81)	-1.86 (5.61)	-2.85 (6.76)	-4.58 (10.38)
	No spelling	C	136	4.79 (3.12)	7.70 (5.11)	3.99 (2.70)	8.24 (4.50)	8.78 (5.33)	15.93 (9.12)
	No spelling	C-I	136	0.64 (4.37)	0.01 (6.66)	-0.86 (3.75)	0.23 (5.66)	-0.22 (7.04)	0.24 (11.00)

'C' is the number of correct translations. 'C-I' is the number of correct-minus-incorrect translations.

.59 to .80), but differences were more evident for the intermediate- than for the high-difficulty level tasks.

To reduce the overall number of statistical tests for the subsequent validity analysis, the scoring procedures and duration with the highest reliabilities for each measure were selected to carry forward. If reliabilities were similar across scoring procedures and duration, then ease, efficiency, and previous research were taken into consideration in making the selection.

For the maze task, the number of correct-minus-incorrect choices made in two minutes was selected. Reliability coefficients were the same across the rule and no-rule conditions. Although the maze is easier to score without the use of a three-in-a-row rule, in this case, we elected to use the scoring rule because in most previous maze research, this rule has been implemented. Use of the rule allowed us to compare our results with those of previous research.

For the word-translation task, the number of correctly spelled translations in two minutes was selected. Although reliabilities tended to be somewhat higher for the Dutch-to-English than the English-to-Dutch version, we maintained both versions for the validity analysis because these measures have not been considered in previous research as CBM progress measures. Reliabilities for spelling and no-spelling scoring methods were similar, but scoring correctly spelled translations involves less judgment and is thus less time consuming; thus, we selected scoring correctly spelled translations.

Table 3.5
Alternate-form reliability of form A and B: Maze, Dutch-to-English, and English-to-Dutch translation

		n	*r*	*r*	*n*	*r*	*r*
Scoring			1 min	2 min		1 min	2 min
Measure			Maze				
Rule	C	260	.55	.69			
Rule	C-I	260	.62	.78			
No rule	C	259	.58	.74			
No rule	C-I	259	.64	.78			
Measure			Word translation intermediate difficulty level				
			Dutch to English			English to Dutch	
Spelling	C	92	.74	.87	94	.65	.76
Spelling	C-I	92	.55	.69	94	.44	.64
No spelling	C	92	.76	.88	94	.68	.78
No spelling	C-I	92	.67	.78	94	.52	.71
Measure			Word translation high difficulty level				
			Dutch to English			English to Dutch	
Spelling	C	143	.77	.83	136	.69	.80
Spelling	C-I	143	.62	.66	136	.53	.59
No spelling	C	143	.72	.81	136	.68	.80
No spelling	C-I	143	.58	.65	136	.50	.59

All correlations were significant at $p < .001$ level. 'C' is the number of correct choices on the maze or translations on the word-translation task. 'C-I' is the number of correct-minus-incorrect choices on the maze or translations on the word-translation task.

Validity

To examine the validity of the measures, correlations between the potential progress measures and English course grades and percentile scores on the standardized English reading test were examined. In addition, for the maze-selection task (which was the same across all participants) differences in mean scores for students in their second vs. third year of English and across very low to high school levels were examined. Combined scores across Forms A and B were used for all analyses.

Mean differences in maze-selection

Mean score differences in maze performance between more and less proficient groups (in our case, students with more or fewer years of English-language instruction or students in higher and lower school levels) would support the validity of the maze measure as an indicator of general foreign-language proficiency. We conducted a two-way analysis of variance to determine whether there were significant differences between grade- and school level. Recall that students in 8th-grade were in their second year and students in 9th-grade in their third year of English foreign-language instruction. Means and standard deviations broken down by grade and school level are presented on the left side of Table 3.6. A main effect was found for school level (F3, 252) = 24.38, $p < .001$) but not grade (F1, 252) = 1.67, $p = .20$). There were no interaction effects (F3, 252) = .15, $p = .93$).

Post-hoc analyses of school-level differences revealed that there were significant differences between each adjacent school level, except for between intermediate and high levels. Regarding grade, although differences were not statistically significant, mean scores tended to be higher for 9th- than for 8th-grade students.

Table 3.6
Means and SDs 2-min Maze, Dutch-to-English and English-to-Dutch translation scores broken down per grade and school level

		Maze			Word translation Dutch to English			Word translation English to Dutch	
		Correct minus incorrect			Correct			Correct	
Grade level	School level	*n*	*M*	*SD*	*n*	*M*	*SD*	*M*	*SD*
Grade 8	Very low	21	16.24	14.71	21	x	x	x	x
	Low[1]	30	23.67	14.32	21	18.24	11.69	21.67	9.64
	Intermediate[1]	18	34	13.72	18	22.17	10.33	25.89	8.59
	High[2]	27	36.7	14.55	25	13.80	7.62	14.44	5.46
Grade 9	Very low[1]	51	17.78	16.18	43	17.02	12.94	21.84	9.40
	Low[2]	54	25.78	14.64	49	11.61	8.35	9.76	5.89
	Intermediate[2]	41	35.56	13.53	30	17.20	10.23	14.60	8.75
	High[2]	18	41.72	14.15	14	27.29	12.65	25	11.73

[1]word translation: intermediate difficulty level
[2]word translation: high difficulty level
'x' indicates that students did not complete the measure.

Correlations with English course grades

In the second set of analyses, scores on the maze-selection and word-translation tasks were correlated with end-of-the-year English course grades. This analysis was conducted within grade and school level because the relative meaning of grades differs across grade and school levels. Splitting the sample by grade and school level resulted in small samples, ranging from 14 to 48 participants per group.

Means and standard deviations for the maze and word-translation tasks broken down by grade and school level are presented in Table 3.6. Means and standard deviations for the criterion variables broken down by grade and school level or reading test level are reported in Table 3.7. English course grades ranged from 3.0 to 8.8 with a mean score of 6.29. Percentile scores on the reading test ranged from 0 to 100, with mean scores per test level ranging from approximately the 19.5 to 66.5 percentile.

Pearson correlations between each progress measure and English course grades are reported in Table 3.8. Correlations between the maze tasks and course grades ranged from r = .20 to .79, with all but one correlation above .40. Correlations between the word-translation tasks and course grades ranged from r = .44 to .77. Correlation coefficients tended to be similar across the three measures, although patterns varied somewhat by school level and grade.

Table 3.7
Means and SD English course grades and percentile scores on English reading test, broken down in grade- and school level or in English reading test level

English course grades	School level	*n*	*M*	*SD*	Min	Max
8th grade	very low	21	7.08	1.30	3.8	8.8
	low	21	6.28	7.85	4.9	7.9
	intermediate	17	6.77	.98	5.0	8.5
	high	25	6.14	1.06	4.2	7.9
9th grade	very low	43	5.74	1.55	3.0	8.6
	low	48	6.38	1.07	4.0	8.7
	intermediate	23	6.62	.84	4.6	8.0
	high	14	5.69	1.10	3.0	7.2
Reading test level		*n*	*M*	*SD*	Min	Max
1		20	32.60	23.90	0	99
2		27	19.59	25.72	1	92
3		20	47.10	23.52	8	95
4		x	x	x	x	x
5		21	66.48	25.67	27	100
6		20	40.30	34.67	1	97

'x' indicates that no information was available for reading test level.

Percentile scores on the standardized English reading test

In the third set of analyses, correlations were calculated between scores on the progress measures and percentile scores on a standardized English reading test. The English reading test consisted of six possible reading test levels. At 8th grade, very-low schools administered

level 1, low and intermediate schools administered level 2, and high schools administered level 3.At 9th grade, very-low schools administered level 1, low and intermediate schools administered level 2, and high schools administered level 3. The scores for 9th-grade very-low schools were not available. Because of the difference in test levels, and because only percentile scores (and not standard scores) were available from the school, comparisons could only be made within reading test level, resulting in small samples ranging from 20 to 27 participants per group. The results are presented in Table 3.9.

Correlations between the scores from the English reading test and maze, Dutch-to-English, and English-to-Dutch translation ranged respectively from $r = .40$ to .78, $r = .37$ to .65, and $r = .19$ to .75. Correlations tended to be higher for the maze task than for either of the word-translation tasks.

Table 3.8
Correlations between English course grades and scores from maze and word-translation

Measure				Maze		Word translation		Word translation	
				Correct minus		Dutch to English		English to Dutch	
				incorrect		Correct		Correct	
	Grade	School level	*n*	*r*	*p*	*r*	*p*	*r*	*p*
Course grades	8	Very low	21	.43	ns	x	x	x	x
		Low[1]	21	.79	<.001	.61	<.01	.72	<.001
		Intermediate[1]	17	.56	<.05	.59	<.05	.62	<.01
		High[2]	25	.42	<.05	.71	<.001	.44	<.05
	9	Very low[1]	43	.78	<.001	.77	<.001	.77	<.001
		Low[2]	48	.63	<.001	.62	<.001	.64	<.001
		Intermediate[2]	23	.20	.ns	.46	<.05	.44	<.05
		High[2]	14	.70	<.01	.60	<.05	.65	<.05

[1]word translation: intermediate difficulty level
[2]word translation: high difficulty level
'x' indicates that the progress measure was not administered for the grade and school level.

Table 3.9
Correlations between percentile score on English reading test, per reading test level, and scores from maze and word-translation

Measure				Maze		Word translation		Word translation	
				Correct minus		Dutch to English		English to Dutch	
				Incorrect		Correct		Correct	
	Grade	Reading test level	*n*	*r*	*p*	*r*	*p*	*r*	*p*
Reading test percentile scores	8	1	20	.40	ns	x	x	x	x
		2[1]	27	.53	<.01	.47	<.05	.42	<.05
		3[2]	20	.44	ns	.37	ns	.19	ns
	9	4	x	x	x	x	x	x	x
		5[2]	21	.65	<.01	.59	<.01	.54	<.05
		6[2]	20	.78	<.001	.65	<.01	.75	<.001

[1]word translation: intermediate difficulty level
[2]word translation: high difficulty level
'x' indicates that no information was available for either reading test level or progress measure.

Regression analysis

In the final set of analyses, a regression analysis was conducted to examine whether a combination of maze and word-translation tasks would improve the prediction of English course grades over using either measure alone. If measures were to be given at the beginning of the school year to identify students who might experience difficulties in language learning, then it would be important to know whether two measures would improve prediction over one measure alone. Students from 9th grade very-low (n = 43) and low (n = 48) school levels were selected for the regression analyses. These groups had the largest sample size, and represented both levels of the word-translation task (intermediate level for the very low group and high level for the low group).

A separate forward stepwise regression was conducted for each group in which the three predictor measures were entered into the equation (maze, English-to-Dutch translation and Dutch-to-English translation). Given the necessity of doing the analysis with small samples, we consider the analysis exploratory. Results can be used to guide future analyses with larger sample sizes.

Results of the regression analyses are presented in Table 3.10. For the very-low group, maze was entered first into the equation and was a significant predictor of English course grades, explaining 60% of the variance in course grades. The scores from English-to-Dutch word translation significantly added to the prediction of course grades over the maze task alone, with the explained variance increasing to 66%. The Dutch-to-English word-translation task did not improve the prediction.

Table 3.10
Relative contributions of maze and word-translation tasks in stepwise regression analyses on English course grades, n = 43 (9th grade, very-low school-level students) and n = 48 (9th grade, low school-level students)

Predictor variables	Cumulative R^2	F Change	p
Very-low level students			
Maze	.60	62.16	<.001
WT En-Du[1]	.66	6.95	<.05
WT Du-En[1]	ns	ns	ns
Low-level students			
WT En-Du[2]	.41	32.49	<.001
Maze	.49	6.38	<.05
WT Du-En[2]	ns	ns	ns

[1]word translation: intermediate difficulty level
[2]word translation: high difficulty level

For the low group, the English-to-Dutch translation task entered first into the equation, and was a significant predictor of English course grades, explaining 41% of the variance. Adding the maze task significantly improved the prediction of course grades over the word-translation task alone, with the explained variance increasing to 49%. The Dutch-to-English word-translation task did not improve the prediction.

In sum, the results of the regression revealed that for both the very-low and low groups, the use of two measures improved prediction over the use of one measure alone. The pattern of results was similar for the two groups, with maze and English-to-Dutch translation contributing to the prediction of course grades, but the order differed for the two groups with maze entering first for the very-low group, and English-to-Dutch word translation entering first for the low group. The overall predictive power of the two measures was less for the low group than for the very-low group.

Discussion

The purpose of this study was to examine the technical adequacy of potential progress monitoring measures for foreign-language learning. Maze-selection and word-translation tasks were examined. Both emerged as potential measures for foreign-language progress-monitoring.

The first research question addressed the alternate-form reliability of the measures. If CBM measures are to be administered on a repeated basis to reflect growth, a necessary (but not sufficient) requirement is that alternate-forms be reliable. For each measure, differences in reliabilities for various scoring procedures and time frames were investigated. For maze selection, four scoring procedures were examined, each consisting of two approaches: a rule vs. no-rule approach and correct vs. correct-minus-incorrect approach. In addition, one- and 2-min were compared. Results of the alternate-form reliability analysis supported the use of a 2-min probe scored for the number of correct-minus-incorrect choices. Few differences in reliability coefficients were seen between the rule and no-rule approaches.

The cessation of scoring following three incorrect choices (rule condition) is done to help to control for guessing; however, as evidenced by the similarities in mean correct scores between the rule and no-rule scoring conditions (see Table 3.2), it would seem that students in our sample did not do much guessing. Given that scoring is easier without the use of a three-in-a-row rule, we *could* suggest scoring without use of a scoring rule; however, there are two points to consider before abandoning the rule completely. First, our study is one of the first to examine CBM progress measures in foreign-language learning. It would be wise to replicate these findings before abandoning the scoring rule altogether. Second, it may be that the scoring rule is important for a small number of students who struggle with foreign-language learning or who do not always complete the probes carefully. Such small numbers may not affect group mean scores or alternate-form reliability results, but would affect the individual growth rates produced via repeated administration of the measures. Thus, at this moment, we suggest that the scoring rule be maintained until further research is conducted.

Similar to previous research on the maze in native language at the secondary-school level, reliabilities in our study increased with administration time (Espin et al., 2010; Tichá et al., 2009); however, different from this previous research, our reliability coefficients were consistently below $r = .80$ (range of $r = .69$ to .78 compared to $r = .79$ to .90), and reliabilities for correct-minus-incorrect scoring were larger than for correct only. These differences are most likely related to the fact that previous research was conducted with students reading the maze tasks in their native language rather than in a foreign language. The average number of errors found in maze research in the native language is approximately one error in two minutes (Espin et al., 2010), whereas in our study, students made on average 2.5 to 3 errors (per passage) in 2 minutes. It would seem that not only fluency but also accuracy contributes to the stability of measurement in foreign-language learning.

The reliability coefficients are lower than those typically found for CBM measures, where coefficients typically exceed $r = .80$. Given that a consistent finding in previous research on maze is that reliabilities increase with probe duration, it would be prudent in future research to examine reliabilities for a three-minute probe. It is especially important that reliability be higher if the measures are used as screening measures to identify students who might experience difficulties with foreign-language learning. If administering repeated measures over time, the current reliabilities might be acceptable, especially if slopes are drawn after collection of 10 to 12 data points. In addition, in other content areas such as science (see Espin et al., 2013), there is evidence that reliability of word knowledge measures increases as knowledge in the content area increases. Thus, it may be that if measures were to be administered over time, alternate-form reliabilities would increase.

For the word-translation tasks, four scoring procedures were also examined, each consisting of two scoring approaches: spelling vs. no spelling and correct vs. incorrect. In addition, 1 and 2-min scoring procedures were compared. Finally, two versions of the word-translation measure were compared: Dutch-to-English and English-to-Dutch. Results of the alternate-form reliability analysis supported a two-minute Dutch-to-English translation probe scored for the number of correctly spelled translations.

The second research questions addressed the validity of the measures. To reduce the overall number of analyses, only a select number of measures were considered for the validity analysis. For the maze selection, the number of correct-minus-incorrect choices made in two minutes was examined. Based on the fact that previous research typically has made use of a three-in-a-row scoring rule in maze, scores calculated with the scoring rule were used for the validity analysis to allow for comparison with previous research. For the word-translation task, the number of correctly spelled translations in two minutes was examined. Both the Dutch-to-English and English-to-Dutch versions were considered in the validity analysis.

The patterns that emerged from the validity analysis provide *tentative* support for both the maze and word-translation tasks as indicators of performance, especially for

students in their third year of English (that is, 9th-grade students). The majority of correlations for students in their third year of English ranged from r = .60 to .78. For students in their second year of English (that is, 8th-grade students), correlations tended to be lower, ranging in general from r = .42 to .79. This pattern was evident for both the criterion variables of course grades and standardized test scores. There was no consistent pattern across groups with regard to the relative strength of the correlations for maze vs. the two word-translation tasks or for Dutch-to-English vs. English-to-Dutch translation. Patterns differed by group, and given the fact that groups were small, it is difficult to draw conclusions about the relative validity of one measure vs. the other.

For the maze-selection task, it was possible to compare mean scores across years of English language instruction and across school levels. Mean scores on the maze task were higher for students in their third year of English (9th grade) than in their second year of English (8th grade) but these differences were not significant. However, significant differences were found between school levels in the expected direction, with students in higher school levels scoring higher on the maze task than students in lower school levels.

In sum, results of mean difference scores among groups and of correlations between the CBM and criterion measures provide tentative support for the use of CBM measures as general indicators of foreign-language performance. Data were more positive for students in their third year of foreign-language learning than in their second year. It may be that performance becomes more stable in the third year of learning a foreign language, and is therefore easier to sample with a brief probe of performance. Also, the criterion variables used in this study, course grades and the standardized reading test, had unknown technical adequacy. In any case, before recommending use of such measures for beginners in language learning, it is necessary to replicate the results with larger samples and with criterion variables with good technical adequacy for the validity analysis.

In our final set of analyses, we examined whether combining two measures predicted performance in English foreign-language learning better than a single measure. The sample sizes were small, so we view the results as suggestive only, and recommend that they be followed up in future research. In general, results revealed that a combination of measures predicted better than a single measure, and that both a reading-related task (maze) and a word-knowledge task (word translation) contributed to the prediction of course grades, with the addition of a second variable accounting for an additional 6% to 8% of the variance. Further, for both groups, the English-to-Dutch word-translation task contributed to the prediction of course grades, but not the Dutch-to-English task, although in a different order with maze entering the equation first for the very-low group and second for the low group.

Both the maze and English-to-Dutch word-translation tasks require recognition rather than production of English words. It may be that at lower levels of learning a foreign language (such as was the case in our two samples), recognition tasks differentiate students

better than production tasks. If we had had been able to conduct a regression with higher performing students, we may have obtained a different pattern of results. It is also important to keep in mind that the English-to-Dutch version of the word-translation task had lower alternate-form reliabilities than did the Dutch-to-English version, especially for the intermediate level task. Thus, at this point, we cannot recommend that one translation task is better than the other.

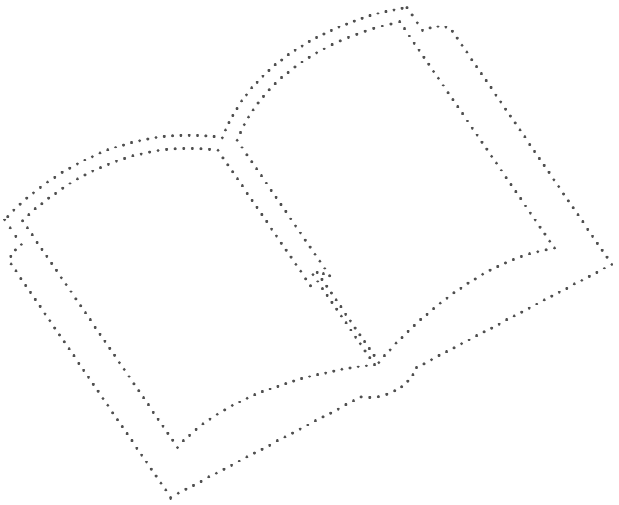
Conclusion

In conclusion, our results provide tentative support for the use of three potential measures for progress monitoring in foreign-language learning: the number of correct-minus-incorrect choices in two minutes on a maze-selection task and the number of correctly spelled translations in two minutes on a word-translation task (English to Dutch, Dutch to English). For screening purposes, using both a reading and word-translation measure resulted in better prediction than either measure alone. We found no clear pattern of differences between Dutch-to-English or English-to-Dutch translation tasks, and recommend that this be explored further.

This study is one of the first to explore the development of CBM progress measures for foreign-language learning. Many questions remain unanswered. First, it is important that the results be replicated. Second, it might be interesting to investigate the technical adequacy of the progress measures with criterion variables in which different aspects of language proficiency (e.g., vocabulary/word knowledge, reading, listening, speaking, writing and grammar) are represented. Third, once reliable and valid measures are identified, it is important that their technical characteristics as growth measures be examined. Most important is to examine whether teachers use progress data to inform instructional decision-making, and whether data use leads to improved student achievement in foreign-language learning.

Chapter 4

CBM maze-scores as indicators of reading level and growth for seventh-grade students



Siuman Chung

Christine Espin

Claire Stevenson

Reading and Writing: An Interdisciplinary Journal (2018), *3*, 627-648.

doi: 10.1007/s11145-017-9803-8

Abstract

The technical adequacy of CBM maze-scores as indicators of reading level and growth for 7th-grade secondary-school students was examined. Participants were 452 Dutch students who completed weekly maze measures over a period of 23 weeks. Criterion measures were school level, dyslexia status, scores and growth on a standardized reading test. Results supported the technical adequacy of maze scores as indicators of reading level and growth. Alternate-form reliability coefficients were significant and intermediate to high. Mean maze scores showed significant increase over time, students' growth trajectories differed, and students' initial performance levels (intercepts) and growth rates (slopes) were not correlated. Maze reading level and growth were related to reading level and/or growth on criterion measures. A nonlinear model provided a better fit for the data than a linear model. Implications for use of CBM maze-scores for data-based decision-making are discussed.

Introduction

To be effective and productive in life, students must be able to read proficiently (Organisation for Economic Co-operation and Development [OECD], 2016), yet many students struggle with reading, even into adolescence. In 2015, 20% of 15-year old students in OECD countries (OECD, 2016) and 24% of 8th-grade students from the United States (U.S. Department of Education, 2016b) could not read at basic levels of proficiency, meaning that these students had not mastered the reading skills expected of students their age. Such students face higher risks of poor educational and employment outcomes at age 19 and 21 (OECD, 2016).

Given the importance of reading, it would seem sensible to provide specialized reading instruction to struggling readers throughout their school careers, even into their secondary-school years. Specialized reading instruction for secondary-school students has been shown to be effective (Edmonds et al., 2009; Wexler, Vaughn, Edmonds, & Reutebuch, 2008), although effects sizes often have been small (Scammacca et al., 2016). Such small effects might be due in part to the magnitude and complexity of reading difficulties for secondary-school students (Vaughn et al., 2008). These students continue to struggle with phonological, fluency, and comprehension aspects of reading (D. Fuchs, Fuchs, Mathes, & Lipsey, 2000; Savage, 2006; Vellutino, Tunmer, Jaccard, & Chen, 2007), and the extent to which each aspect affects their overall reading performance differs among individuals (Catts, Adlof, & Weismer, 2006; Vaughn et al., 2008).

The magnitude and complexity of reading difficulties for secondary-school students creates unique challenges for educators as they strive to develop effective interventions for these students. Educators could be helped in their efforts if they were to have access to a tool that would be sensitive to small improvements over time, and could be used to evaluate the effects of specialized interventions for individual students. Such a tool would be of practical importance because, though effects may be small, minor improvements in reading during the adolescent years might translate into major improvements in success and satisfaction in adult life. A tool that is potentially suitable for monitoring the progress of secondary-school students with reading difficulties, and for evaluating the effects of interventions on an individual basis, is Curriculum-Based Measurement (CBM; Deno, 1985, 2003).

Curriculum-Based Measurement

CBM is a system designed for teachers to monitor the progress of students with learning difficulties, and to evaluate the effects of interventions on the students' progress (Deno, 1985, 2003). CBM often is used within Multi-Tiered Systems of Supports (MTSS), and is uniquely suited for monitoring progress for 'Tier 3'-students, that is, students who are in need of individualized, intensive instruction. CBM involves frequent (e.g., weekly)

administration of short timed probes of equivalent difficulty. Scores from probes are plotted on progress graphs that depict student growth. Student growth is continuously compared to an expected (desired) rate of growth to determine whether the instruction should be changed or the goal raised. The 'expected' rate of growth is depicted by a straight goal line drawn on the graph that extends from the baseline to the anticipated level of performance at the end of the school year. Students are monitored frequently so that teachers can evaluate growth and make instructional decisions in a timely fashion. When teachers use CBM data to make instructional decisions, they affect significant improvements in student performance (L. S. Fuchs, Fuchs, & Hamlett, 2007; Stecker et al., 2005).

The majority of research on CBM in reading has been conducted at the elementary-school level. In recent years more attention has been directed toward the secondary-school level (see Wayman et al., 2007, for a review), but that research has focused primarily on scores as indicators of performance rather than growth (Espin et al., in press, or see also Chapter 2). To be used for instructional decision-making within Tier 3 settings, scores from CBM measures must be shown to be reliable, sensitive, and valid indicators of growth.

In this study, we examine the reliability, sensitivity, and validity of scores from CBM reading measures as indicators of growth for secondary-school students. We focus on one particular CBM measure: maze-selection. A maze is a passage in which every 7th word is deleted and replaced with a multiple-choice item consisting of the correct word and two distracters. Students read silently for 2-3 min, selecting the word that restores meaning to the text. The number of correct selections is counted and placed on the graph. To date, studies examining the technical adequacy of maze scores have produced tentative support for their use as indicators of reading progress for secondary-school students, but these studies have been limited in several respects.

Research on CBM Maze for Monitoring Progress of Secondary-School Students

Espin et al. (2010) and Tichá et al. (2009) examined the technical adequacy of maze scores administered weekly over a period of 10 weeks. Participants were 31 (Espin et al., 2010) and 35 (Tichá et al., 2009) 8th-grade students who completed 4-min maze passages. Scores were the number of correct and correct-minus-incorrect choices for 2, 3, and 4 min. Criterion variables were scores on a state reading test (Espin et al., 2010; Tichá et al., 2009), reading group status, and scores and changes in scores on a standardized reading measure (Tichá et al., 2009).

Results were similar across the two studies, and provided support for the reliability, sensitivity, and validity of the maze scores. Alternate-form reliability coefficients ranged from r = .69 to .91, with the majority of coefficients above .80. Validity coefficients ranged from r = .75 to .86. Few differences in reliability or validity were found related to administration time or scoring procedures, with the exception that reliability increased somewhat with an increase in administration time. Scores were sensitive to (linear) growth, and growth on the

maze was related to scores on the state reading test (Espin et al., 2010), to group status, and to growth on the standardized reading measure (Tichá et al., 2009).

Although the results of Tichá et al. (2009) and Espin et al. (2010) supported the use of maze scores for monitoring student progress in reading at the secondary-school level, there were limitations to the studies. The sample sizes were small and included only 8th-grade students. Monitoring occurred for only 10 weeks, and the same set of passages were used across the two studies. It was unclear whether findings from these two studies would generalize to a larger and more diverse sample, to monitoring across the entire school year, or to a different set of maze passages.

In 2012, Tolar and colleagues (Tolar et al., 2012, 2014) examined the technical adequacy of maze scores with a large sample (N = 1,343) of students in grades 6 to 8 who were monitored across the entire school year using AIMS-web passages (https://www.aimsweb.pearson.com). Students in the study were monitored five times across the school year. Scores on the maze were the number of correct-minus-incorrect choices in 3 minutes. Reliability coefficients for maze scores ranged from r = .64 to .91. Validity coefficients ranged from r = .45 to .73 (Tolar et al., 2012). Scores reflected linear growth (Tolar et al., 2012, 2014), but not quadratic growth (Tolar et al., 2012). Finally, growth in maze scores was related to reading performance (Tolar et al., 2012, 2014), but not to growth on other reading measures (Tolar et al., 2014).

The results of the Tolar et al. (2012) and (2014) studies provided further support for the technical adequacy of CBM maze scores for monitoring student progress, however, the study included only five measurement moments, too few for timely instructional decision-making within Tier 3 settings. It not clear whether the results from the Tolar et al. (2012, 2014) studies would generalize to the more frequent progress monitoring.

In sum, although the handful of studies that have been conducted to date at the secondary-school level have provided tentative support for the technical adequacy of maze scores for measuring progress in reading, there is a need for a large-scale study to examine the technical adequacy of scores from frequently administered CBM mazes. It would be important to examine whether earlier results replicate, and to more closely examine the relation between growth on the maze and growth on other reading measures. Whereas Tichá et al. (2009) found a significant relation between growth on the maze and growth on reading measures, Tolar et al. (2014) did not.

Finally, the present study should examine the extent to which CBM scores reflect linear growth over an academic school year. As described earlier, in practice, an assumption is made within CBM that students grow in a linear fashion across the school year. That is, the expected rate of growth is represented by a linear goal line extending across the school year. The line is used to make judgements about student progress and about the effectiveness of instruction on that progress. However, it is possible that growth is not linear. For example, research at the elementary-school level has shown that growth trajectories

produced by CBM reading-aloud measures are nonlinear, with more growth seen in the first half than in the second half of the school year (e.g., Christ, Silberglitt, Yeo, & Cormier, 2010). In addition, studies demonstrate that the nature of growth in reading across school years is nonlinear, with more rapid growth in the first few grades, and less rapid growth in grades 3-8 (e.g., Kieffer, 2011). If the growth trajectories produced by weekly maze scores would prove to be nonlinear, it would have implications for the use of the data for instructional decision-making.

Purpose

The purpose of the current study was to examine the reliability, sensitivity, and validity of scores from weekly administered CBM maze measures as indicators of growth in reading for secondary-school students. Multiple criterion variables assumed to represent students' reading proficiency were included in the study, first because there is evidence that results vary widely across standardized reading tests (e.g., Cutting, & Scarborough, 2006; Jenkins & Pany, 1978), and second, because the use of multiple measures allows us to examine whether evidence converges across various measures, building what Cronbach and Meehl (1955) refer to as a nomological network of evidence (Messick, 1989; see Espin & Deno, 2016, for a description specific to CBM). The study considers both linear and nonlinear (logistic) growth patterns to examine whether the assumption of linear growth underlying CBM implementation is warranted.

The following research question was addressed in the study: "What is the technical adequacy of CBM maze-scores as indicators of reading level and growth for secondary-school students?". To examine this research question, three sub-questions were addressed:

1. What is the alternate-form reliability of scores on maze passages?
2. What is the sensitivity to growth of maze scores?
    a. Do maze scores increase over time?
    b. Do students show individual differences in growth trajectories?
    c. Do students with higher initial maze scores show greater growth than students who start with lower maze scores?
    d. What type of growth model, linear vs. nonlinear (logistic), best fits weekly maze-scores?
3. Are maze scores valid indicators of reading level and growth?
    a. Are maze scores and change in maze scores related to group status (school level and dyslexia status)?
    b. Are maze scores and change in maze scores related to scores and change in scores on a standardized reading test?

Method

Participants

Participants were 452 7th-grade students (233 male) from three secondary-schools in the Netherlands. Mean age for the participants was 12.63 (*SD* = 0.63; range 12 - 15) years. Fifty-four participants were students with dyslexia. Dyslexia is defined in the Netherlands as a disorder characterized by a persistent problem with the learning and/or the application of skills in reading and/or spelling at the word level (Stichting Dyslexie Nederland, 2008).

Participants in the study were from a range of school levels. In the Netherlands, secondary-schools are organized into different levels, referred to (in order from lowest- to highest-level) as: practical, pre-vocational (low, intermediate, high), senior general secondary, and pre-university education (Dutch Ministry of Education, Culture and Science, 2005). Instruction and curriculum differ between school levels (i.e., students at higher school levels are required to process more complex information and perform more in-depth thinking than students at lower school levels). In reading, students at all school levels are provided instruction in which increasingly more complex texts are offered, but instruction is provided at a different pace. In addition, all school levels lead to differentiated criteria for graduation at the end of secondary school. For instance, pre-vocational students are expected to read expository passages about common topics either related or unrelated to their daily life, and to read simple narrative literature for adolescents on a surface level by the end of their secondary school, whereas, pre-university students are expected to read and understand a variety of expository passages about different topics within their curriculum and/or about socially relevant topics, and to be able to read and interpret narrative literature for adults by the end of secondary school (Dutch Ministry of Education, Culture and Science, 2009). In the year that the study took place, placement into school level was based on students' academic performance during elementary school and on scores on a national achievement test.

Participants in our study often were placed in classes that combined school levels, thus for purposes of the analyses, we grouped students into three levels: low (practical and low pre-vocational), intermediate (intermediate and high pre-vocational), and high (senior general secondary and pre-university). Low achieving students were overrepresented in our study due to practical reasons.[2] The distribution per school level in our sample vs. the country was approximately 46.7% vs. 14% for low, 36.5% vs. 43% for intermediate, and 16.8% vs. 43% for high levels (CBS StatLine, 2015).

[2] One school with a large number of students in the high school-level agreed to participate, but had withdrawn their participation in the first few weeks of the study due to unavailable resources to collect the data.

Reading instruction

All participating schools provided reading instruction to their students using a reading-comprehension curriculum called *Nieuwsbegrip* (Understanding the news, CED-groep, 2012). The reading instruction follows a recursive process of six weeks, in which each week one of five reading strategies (i.e., predict, clarify, summarize, generating questions, and making connections) is discussed and applied to an expository reading passage. At week six all strategies are applied to a reading passage (CED-groep, 2012). It was not observed how much time was spend and what the quality of the instruction was.

Instruments and Procedure

Maze

The predicted variable in the study was the number of correct selections on the maze. Mazes were reading passages in which the first sentence was left intact, and thereafter, every seventh word was replaced by a multiple-choice item. Each multiple-choice item contained the correct word and two distracters. Distracters were within one letter in length to the correct choice and were clearly incorrect, that is, the word did not (a) fit contextually in the text, (b) rhyme with the correct choice, or (c) sound or look like the correct choice (L. S. Fuchs & Fuchs, 1992).

The mazes were administered weekly using an online program developed by the researchers called *Mazesonline*® (http://www.mazesonline.nl). To complete the maze, students read silently through the passage, and selected a word at each multiple-choice item. After two minutes the task automatically stopped and the number of correct and incorrect choices were registered. The number of correct maze choices (CMC) were used in the current study.

Mazes for this study were constructed from expository reading passages of approximately 400 words. Passages were long enough to ensure that students did not finish the passage before time was up. Reading passages were written by the research team, and focused on general topics thought to be appropriate for and of interest to secondary-school students. Passages were equivalent in terms of scores on a common reading index used in the Netherlands (van den Berg & te Lintelo, 1977). This index is based on the average number of words in each sentence and the average number of syllables per word. The index level for the passages fell within the range of 69-73, a level considered comparable to the reading level of an average performing 5th-grade student at the end of the school year.

Reading passages were converted into maze passages. Multiple-choice items were placed between brackets and in bold. Passages were formatted so that all choices were on one line. The correct option was randomly placed in the first, second or third position for each item. To examine the suitability of the passages two small pilots were conducted, one with nine secondary-school students not participating in the study, and one with eight

graduate students. Based on the results, passages that were too difficult (i.e., relatively small number of words read) were removed from the passage set. A set of 15 maze passages remained to be used in the study. These 15 passages also showed a distribution in scores among 127 7th- to 11th-grade low-performing students (i.e., practical education) in an unpublished study conducted prior to the current study, indicating that the passages were not too difficult for the lower performing students. For the current study, 17 passages were needed. Thus, two additional maze texts were written at the same reading index level as the other passages. Information gleaned from the development of the other passages was used to write the two additional passages.

Reading proficiency

In line with Messick's (1989) approach to establishing construct validity for maze scores as indicators of reading proficiency, we examined the pattern of relations between scores from the maze with various measures assumed to represent reading level and reading growth. These measures included group status (school level and dyslexia status) and scores on the standardized reading test: CITO-VVO. School level and dyslexia status have been already described, thus we only describe the CITO-VVO in the following section.

CITO-VVO

The Dutch reading subtest on the *Cito Volgsysteem Voortgezet Onderwijs* (Cito Progress Monitoring System for Secondary Schools; Cito, 2010), hereafter named CITO-VVO, is a nationally-normed reading test in which students read 6-8 narrative and expository reading passages and answer 40-50 multiple-choice questions. It is administered once a year via pencil and paper. Different forms of the test are made for school level and grade levels. Scaled scores allow for comparison across school and grade levels and measurement occasion. The test for 7th-grade students was administered at the beginning and end of the school year, and given across two sessions at each measurement occasion. Administration time per session is 45 minutes (Cito, 2010).

Internal-consistency (Cronbach's alpha) was reported to be $\alpha > .70$. Construct validity included differences in school levels and discriminant validity: students in higher school levels performed higher than students in lower school levels, and CITO-VVO scores correlated with theoretical related constructs (i.e., vocabulary) and not with unrelated constructs (i.e., mathematics, Egberink, Janssen, & Vermeulen, 2015a, 2015b). Scores on the measure were obtained from the schools. In Table 4.1 the number of participants who completed the CITO-VVO per school level and descriptive statistics are provided.

Table 4.1
Descriptive Statistics of the CITO-VVO

CITO-VVO	*n*	*M*	*SD*	Min	Max
Pre-test (Sept)	393	205.57	18.04	168	267
Low	166	192.93	11.77	168	221
Intermediate	152	207.14	11.44	182	245
High	75	230.39	12.69	206	267
Post-test (Jun/Jul)	386	211.16	29.48	148	309
Low	164	192.30	21.71	148	254
Intermediate	147	211.27	18.15	158	266
High	75	252.21	18.39	215	309

Data Collection

Data were collected between January and June of one school year. At the beginning of the study, teachers were informed about the purpose, background and instruments of the study. Teachers organized and supervised the electronic administration of the maze passages in *Mazesonline®* (http://www.mazesonline.nl). Students completed a total of 17 parallel maze passages, and received one maze weekly (with the exception of vacation weeks) over a period of 23 weeks between January to June. The research team was present for the initial administration to ensure that the system worked. During the first session teachers gave students a short introduction on the background of the study and instructions on how to complete the maze. Students then signed into the system and completed a practice task, after which they completed the first maze. Each week thereafter, students completed one maze via the online system.

Data Analyses

The first research question addressed the *alternate-form reliability* of maze passages. This was assessed via Pearson correlations in IBM SPSS Statistics 24.

To address the research questions related to the growth and validity of maze scores, multilevel analyses with maximum likelihood estimation were performed in the statistical software R, using the lme4 package (Bates & Maechler, 2010). Multilevel analysis is especially suitable for examining longitudinal data because it controls for dependence between measurements, missing data, and unequal groups in categorical variables (Hox, 2010). A multilevel model for linear growth consists of a within-individual level:

$$Y_{ti} = \pi_{0i} + \pi_{1i}\, maze\ session_{ti} + e_{ti} \quad (1)$$

where Y_{ti} is the maze score for individual i at time t. π_{0i} is the maze performance for individual i at the beginning of the study (intercept). π_{1i} is the maze growth rate per week (slope) for individual i. e_{ti} is the error term at the within-individual level.

The multilevel model also contains a between-individual level:

$$\pi_{0i} = \beta_{00} + \beta_{01} Z_i + u_{0i} \quad (2)$$

$$\pi_{1i} = \beta_{10} + \beta_{11} Z_i + u_{1i} \quad (3)$$

where β_{00} and β_{01} are respectively the mean intercept and slope to predict π_{0i} from a between-individual level variable Z_i (e.g., reading test score) for individual i. The u_{0i} is the error term for the overall intercept π_{0i}, i.e., the difference between mean intercept and an individual's intercept. The β_{10} and β_{11} are, respectively, the mean intercept and slope to predict π_{1i} from a between-individual level variable Z_i for individual i. The u_{1i} is the error term for the overall slope π_{1i}, i.e., the difference between mean slope and an individual's slope.

The multilevel model presented in equation (1) can be adapted to fit nonlinear growth as follows:

$$Y_{ti} = \pi_{0i} + \pi_{1i} (\ln(maze\ session_{ti} + 1)) + e_{ti} \quad (4)$$

In essence, equation (4) represents the same model as the linear model in equation (1), but the change of maze scores is expected to follow a logarithmic curve rather than a straight line.

To examine the *growth*, four steps were taken. First, the data were examined to determine whether the maze scores were sensitive for measuring growth in general. For this first step, a model with zero growth was compared to a model with a fixed effect for growth. Second, the data were examined to determine whether the maze scores were sensitive for measuring individual differences in growth between students (i.e., different growth trajectories), and compared to the model from step one. Third, a model in which the intercept and slope were correlated was included to examine whether students with higher levels of reading performance showed higher rates of reading growth compared to students with lower reading levels. This model was then compared to the second step model. The models were compared using a Likelihood Ratio Test (LRT), which was a Chi-squared test that assessed change in goodness of model fit from a nested model to a more complex model containing additional parameters; a significant result indicated that the new model better fitted the data and that one could statistically infer that the additional parameters were of value (Hox, 2010). The analyses were performed for two growth models; linear versus nonlinear. Fourth, the two growth models were compared to determine which model best fitted the maze data. Because the LRT only can be used in nested models, the linear versus nonlinear models were compared by looking at the fit indices: AIC, BIC, Log likelihood and Deviance, where smaller indices indicated better fitting models (Hox, 2010).

Two approaches were used to examine *validity*. First, the relations between maze scores and group status (school level and dyslexia status) were examined. With regard to school level, we examined whether maze scores were higher and growth was greater for students at higher school levels than for students in lower levels. We applied a contrast forward difference coding to compare the three categories separately. With regard to dyslexia, we examined whether maze scores were higher and growth was greater for students without dyslexia than for students with dyslexia. The fixed effects of the linear and nonlinear models were interpreted at group performance levels. We examined whether the

maze session, group level, and the interaction maze session X group level were significant predictors of maze scores.

Second, we examined the relations between maze scores (both performance levels and growth) and performance level and/or growth on the CITO-VVO. We examined whether the students who obtained higher scores and achieved greater growth on the maze also obtained higher scores and achieved greater growth on the CITO-VVO. The significance of the fixed effects was interpreted for the predictors maze session, reading test score or growth, and interaction maze session X reading test score or growth.

Results

Data Inspection

Data were inspected,[3] and scores were removed if students had a zero score due to the fact they made no selections, or had a potentially inflated score due to the fact they made random selections.[4] We verified that the assumptions for multilevel analyses, normality and homoscedasticity at within- and between-individual levels were met. The intra-class correlation (ICC) was .70, which indicated that the variance explained at the between-individual level was high.

Alternate-Form Reliability of Scores on Maze Passages

Table 4.2 provides an overview of the correlations between scores on maze passages for adjacent administration weeks. The correlations ranged between $r = .67$ and .83.

Growth of Maze Scores

The growth of maze scores was examined by fitting different models for linear and nonlinear growth.[5] Table 4.3 presents the fit indices per model, and the Likelihood-Ratio-test results for each model comparison, and Table 4.4 (last row) presents the estimated mean initial maze scores (intercept) and progress (growth) for both growth models.

[3] For three maze passages, in the latter part of the passage the Mazesonline system registered the fact that the student had made *a* selection, but not *which* word was selected. Because the glitch occurred in the latter part of the passage, it affected only a small number of students. (Most students did not reach this point in the 2-minute time limit.) For the students who were affected, we examined their scores on all other maze passages, and noted that they made few if any incorrect choices on their maze. We thus made the decisions to assume that their selections had been correct, and counted these items as correct choices.

[4] There was the possibility that students could just randomly click on answers without reading the text, and thereby obtain an inflated correct maze score. To control for such random selections, we identified scores that included a larger than expected number of correct and incorrect choices. "Larger than expected" was defined as scores that were more than two standard deviations larger than the mean correct/ incorrect group score for the passage. This comprised only 0.02% of the scores and these were removed from the analysis. Nine additional scores were removed because students did not make any selections. The removal of scores was considered not a problem, because only a short amount of scores were removed, and were from students from various proficiency levels.

[5] We did not control for school in the analyses because our focus was on differences between school levels rather than differences between schools.

Table 4.2
Correlations between Maze Passages with One-Week Intervals

	Passage comparison							
	0 & 1	1 & 2	2 & 3	3 & 4	4 & 5	5 & 6	6 & 7	7 & 8
n	150	95	169	154	214	65	106	174
r	.75**	.78**	.78**	.79*	.81**	.83**	.68**	.67**
	18 & 19	19 & 10	10 & 11	11 & 12	12 & 13	13 & 14	14 & 15	15 & 16
n	198	59	245	172	134	139	163	95
r	.73**	.72**	.77**	.69**	.76**	.75**	.74**	.69**

**$p < .01$

Table 4.3
Fit Indices and LRT results for Null, Linear and Nonlinear Growth Models

Model	Nested Model	Fit indices				Likelihood Ratio Test (LRT)		
		AIC	BIC	-LL	Deviance	χ^2	*df*	*p*
Null model								
M0		28130	28149	14062	28124			
Linear								
M1A	M0	27961	27986	13976	27953	171.21	1	<.001
M1B	M1A	27859	27891	13924	27849	104.4	1	<.001
M1C	M1B	27858	27897	13923	27846	2.37	1	.12
Nonlinear								
M2A	M0	27908	27934	13950	27900	223.65	1	<.001
M2B	M2A	27800	27832	13895	27790	110.14	1	<.001
M2C	M2B	27800	27838	13894	27788	2.50	1	.11

M0 = zero growth, M1A/M2A = mean growth, M1B/M2B = individual differences, M1C/M2C = correlation intercept and slope, AIC = Akaike information criteria, BIC = Bayesian information criteria, LL = Log likelihood

Table 4.4
Mean Maze Intercept (Initial Performance Level) and Mean Maze Slope (Rate of Growth) per School Level and Growth Model

Maze scores	*n*	Linear model				Nonlinear model			
		Intercept ($\bullet_0$)[1]		Slope ($\bullet_1$)[1]		Intercept ($\bullet_0$)[2]		Slope ($\bullet_1$)[2]	
		M	SE	*M*	SE	*M*	SE	*M*	SE
School level									
Low	211	22.32	0.42	0.10	0.02	21.41	0.43	0.80	0.15
Intermediate	165	26.54	1.05	0.17	0.06	25.53	1.05	1.28	0.39
High	76	33.43	1.23	0.24	0.07	31.19	1.22	1.99	0.43
Dyslexia									
Yes	54	21.30	1.37	0.15	0.06	20.33	1.35	1.16	0.42
No	398	26.39	0.35	0.15	0.02	25.34	0.35	1.19	0.11
Total	452	25.72	0.36	0.16	0.01	24.70	0.38	1.22	0.08

[1]See equation 1
[2]See equation 4

To measure whether students' maze scores *improved over time*, a linear and nonlinear (logistic) growth model were fitted to the data and compared to a null model where no growth was assumed (M0). For linear growth, the parameters of equation (1) at the within-individual level and (2) at the between-individual level were estimated; π_{1i} was a constant. For nonlinear growth, equation (4) was used at the within-individual level. A linear mean growth model (M1A) was found to be a significantly better model than the null model, indicating that, on average, students changed over time on maze scores. The mean intercept

was 25.72 correct choices, and mean growth was an increase of 0.16 correct choices per week. Students made on average 25.72 CMC during the first session, and after for instance ten weeks increased with 1.60 CMC to 27.32 CMC. A nonlinear growth model (M2A) was also found to be a significantly better model than the null model. The parameters of equation (4) were estimated and resulted in the formula: 24.70 + 1.22 x ln(maze session + 1); see Figure 4.1 for a depiction of the nonlinear mean growth rate on maze scores. Students made on average 24.70 CMC during the first session and the increase of scores became smaller with each week. For instance, after five weeks, students made on average 26.89 CMC (an average increase of 0.44 per week), whereas after ten weeks, students made on average 27.63 CMC (an average increase of 0.29 per week).

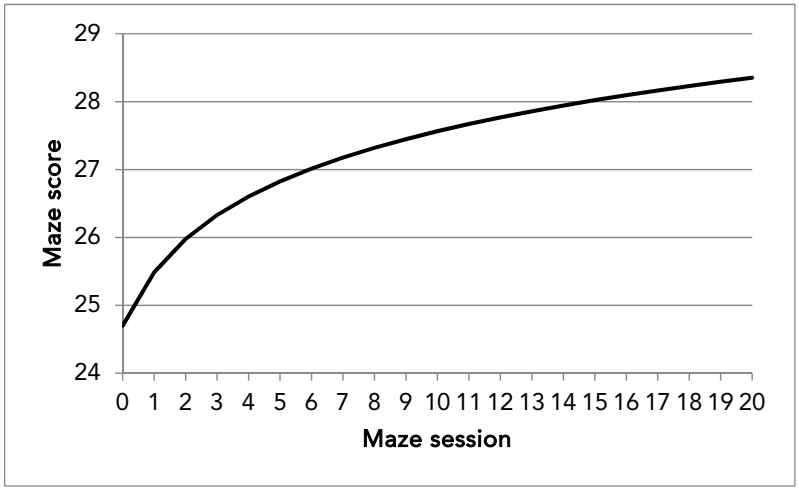


Figure 4.1. The estimated mean growth curve in the nonlinear model

To measure whether students showed *individual differences in growth trajectories* on maze scores, random growth rates were added to the fixed growth model resulting in M1B and M2B for linear and nonlinear growth respectively (see Table 4.3). In both cases, the parameters of equation (3) at the between-individual level were also estimated. For both growth models, students showed significantly different growth trajectories. Thus, students' individual growth rates on the maze differed.

Next a model in which the initial maze scores (intercept) and maze growth (slope) were correlated was fitted to examine whether students with higher initial maze scores grew more over time than those with lower initial maze scores (see Table 4.3). Initial maze scores were not significantly related to the rate of growth, indicating that students with higher initial maze scores did not show greater growth than students with lower initial scores.

Comparing fit indices (AIC, BIC, -LL, and Deviance; smaller is better, Hox, 2010) revealed that the nonlinear growth model was a better fit than the linear growth model for the students' maze scores. Thus, the students' growth on the mazes was better represented by a nonlinear than a linear model.

Validity of Maze Scores

To address the research questions related to validity, relations between maze scores (both level and growth) and group status, and between maze scores and scores on the CITO-VVO were examined for both linear and nonlinear growth models. The difference (growth) score on the CITO-VVO was calculated by subtracting the pretest score from the posttest score. In Table 4.5, an overview of the estimated parameters of the fixed effects and its significance for both growth models is presented.

Table 4.5
Fixed Effects and Significance of Group Status and Reading Tests for Linear and Nonlinear Growth Models

	n	Linear			Nonlinear		
		B	*SE*	*t*	*B*	*SE*	*t*
School level							
Maze level: low vs. intermediate contrast	452	-6.36	0.61	-10.42***	-6.13	0.60	-10.24***
Maze level: intermediate vs. high contrast	452	-9.19	0.79	-11.62***	-8.85	0.78	-11.35***
Maze growth: low vs. intermediate contrast	452	-0.09	0.03	-2.88**	-0.70	0.20	-3.49***
Maze growth: intermediate vs. high contrast	452	-0.10	0.04	-2.62**	-0.97	0.26	-3.67***
Dyslexia							
Maze level: dyslexia	452	-5.07	1.02	-4.98***	-5.01	1.00	-5.02***
Maze growth: dyslexia	452	0.004	0.05	0.08	-0.03	0.31	-0.11
CITO-VVO							
Maze level: CITO-VVO level	393	0.24	0.02	15.00***	0.23	0.02	14.83***
Maze growth: CITO-VVO level	393	0.002	0.001	1.91	0.02	0.01	2.88**
Maze level: CITO-VVO growth	386	0.11	0.02	5.99***	0.10	0.02	5.71***
Maze growth: CITO-VVO growth	386	0.002	0.001	2.71**	0.02	0.01	3.40***

*** $p < .001$, ** $p < .01$, * $p < .05$

Group status

Mean group differences were examined for maze performance level and growth for more and less proficient groups, as defined by both school level and dyslexia status. School level was significantly related to both performance level and growth on maze for the linear and nonlinear growth model (see Table 4.5). Examining contrasts revealed that there were significant differences in performance level and growth for students at adjacent (low vs. intermediate vs. high) school levels. Students at higher school levels obtained higher maze scores and displayed steeper growth rates than students at lower school levels, see Table 4.4 for mean initial scores and growth rates per school level.

Dyslexia status was found to be significantly related to performance level, but not to growth on maze (see Table 4.5). This result was found for the linear and nonlinear growth model. Students with dyslexia performed lower on maze scores than students without dyslexia; however, the two groups did not differ in growth rates, see Table 4.4 for mean initial scores and growth rates per dyslexia status.[6]

CITO-VVO

The relations between maze performance level and growth, and performance level and growth on the CITO-VVO was examined. The scores were centered
for easier interpretation of the fixed effects. The parameter estimates and the significance of the relations are reported in Table 4.5. Maze performance level was significantly related to performance level on the CITO-VVO for both growth models. Students with higher scores on the CITO-VVO had higher initial maze scores than students with lower scores on the CITO-VVO, see for example the results for an average score of 206 vs. a high score of 260 on the CITO-VVO in Table 4.6. Students with higher scores on the CITO-VVO also showed higher growth rates on the maze than students with lower scores on the CITO-VVO. However, this relation was not significant in the linear model, whereas it was significantly related in the nonlinear model.

Both maze performance level and growth were significantly related to CITO-VVO growth, and this was the case for both linear and nonlinear growth models. Students who showed steeper growth on the CITO-VVO had higher initial maze scores and steeper maze growth than students with a lower growth rate on the CITO-VVO, see for example the results for an average growth rate of 5 vs. a high growth rate of 20 on the CITO-VVO in Table 4.6.

Table 4.6
Maze Intercept (Initial Performance Level) and Maze Slope (Rate of Growth) per Growth Model for CITO-VVO Average and High Scores and Growth Rates

	CITO-VVO level		CITO-VVO growth	
	Average (206)	High (260)	Average (5)	High (20)
Linear				
Maze level	26.42	39.25	25.75	27.35
Maze growth	0.15	0.24	0.13	0.17
Nonlinear				
Maze level	25.30	37.82	24.77	26.27
Maze growth	0.28	0.50	0.26	0.32

[6] As suggested by one of the anonymous reviewers, we further unpacked the results relating to dyslexia status and school level. More specifically, we looked at whether there was a three-way interaction between school level, dyslexia status and maze level and growth. There was no reliable interaction between the three. The results are available upon request.

Discussion

Results of the current study supported the reliability, sensitivity to growth, and validity of maze scores as indicators of reading level and growth for secondary-school students.

Reliability of Maze Passages

The alternate-form reliability of 2-min maze passages was moderate to moderate-good, with correlations ranging from $r = .67$ to .83. Reliability coefficients were slightly lower than those reported in Espin et al. (2010) and Tichá et al. (2009) where reliability coefficients were near $r = .80$, but were similar to those reported by Tolar et al. (2012). Both Espin et al. (2010) and Tichá et al. (2009) found that reliability coefficients increased with an increase in administration time. Thus for screening purposes, it might be wise to combine scores across two maze passages and/or to increase the length of the administration time to 4 minutes. For the purpose of measuring reading growth, the reliability coefficients can be considered acceptable to good; however, it will be important in future research to examine the effects of duration and schedules on the stability of the maze growth estimates (see e.g., Christ et al., 2013, on CBM reading-aloud measures).

Growth

The second research question addressed growth. We tested both linear and nonlinear growth models. In both cases, the models reflected that students improved over time, and that there were substantial individual differences in initial performance level and growth. These findings are consistent with findings from previous studies (Espin et al., 2010; Tichá et al., 2009). However, unlike previous studies, the linear growth rates in the current study were small: 0.16 correct choices per week compared to 2.17 for Espin et al. (2010) and 0.86 for Tichá et al. (2009).[7] The obtained growth rates were similar to those found by Tolar et al. (2012)[6] where growth rates were 0.13 choices per week (although note that scores in that study were correct-minus-incorrect choices).

The inconsistent findings across studies might be due to several factors, including the composition of the sample, the nature and amount of reading instruction given to students, and the study design. With regard to the sample composition, both the current study and the Tolar et al. (2012) study had a relatively large proportion of low-performing students in their samples (47% for our study; 56% for Tolar et al., 2012). The percentage of low-performers in the Tichá et al. (2009) was 37%. In the Espin et al. (2010) study, performance levels were not reported, however, the mean score on the state reading test was similar to the mean score for all students in the state, indicating that there likely was not a disproportionate number of low-performing students in the sample. It will be important in

[7] To compare across studies, growth rates for 3-minute maze probes were prorated into 2-minute growth rates by multiplying the 3-minute growth rate statistic with 0.67.

sample of students at various grade levels in order to establish normative scores for both reading level and growth.

An additional reason for differences in growth rates might be the amount and nature of the reading instruction provided to the students participating in the studies. Previous research has demonstrated that under typical reading conditions, growth rates for struggling readers (i.e., students in special education) are lower than for non-struggling readers, but under optimal reading instruction conditions, growth rates for the two groups are the same (Deno, Fuchs, Marston, & Shin, 2001). Unfortunately, little information was provided in the studies regarding the amount and nature of the reading instruction provided to the participants. More research is needed into the effects of intensive, individualized reading instruction on the growth trajectories produced by maze scores.

A final reason for differences in growth rates might be the study design, more specifically, the duration, and schedule employed in the studies. Christ et al. (2013) found that the stability of growth trajectories produced by reading-aloud scores differed with duration (number of weeks) and schedule (weekly vs. biweekly data collection). Espin et al. (2010) and Tichá et al. (2009) collected data over a short duration (10 weeks) using a dense schedule (weekly), whereas Tolar et al. (2012) collected data over a long duration (school year) using a less dense schedule (every 6-8 weeks). In our study, we collected data over a relatively long duration (half a school year) using a dense schedule (weekly). The various combinations of duration and schedule might influence the precision of growth estimates. Research is needed to examine the effects of duration and schedule on the growth trajectories produced by the maze scores.

Linear vs. Nonlinear Growth

We found that maze reading growth was best represented within a nonlinear growth model. This result could be explained by the nature of reading growth. For example, Kieffer (2011) demonstrates a plateauing effect of students' reading growth across school years, with rapid reading growth in the first few grades, and less rapid growth through grades 3 to 8. Although the results supported the use of a nonlinear growth model, the use of such a growth model produces a unique challenge to data interpretation. Within CBM, instructional decisions are made by comparing the student's rate of growth to a goal line, which is typically represented with a linear growth line. Use of a logistic learning curve might better represent students' reading growth, but it is likely to complicate data interpretation, and recent research has demonstrated that interpretation of CBM graphs can be difficult for teachers (van den Bosch, Espin, Chung, & Saab, 2017). A solution to this dilemma might be found in the use of electronic progress-monitoring programs. Given the current development and use of online and software programming possibilities, it is imaginable that programs could be developed that incorporate nonlinear long-range goals to enhance teachers' ability to interpret and use progress data for instructional decision-making. However, it will be important in the future

to examine the effects of using nonlinear growth models on the interpretation and use of the data.

A second factor to consider before recommending use of a nonlinear goal line relates to teachers' expectations and the effects of these expectations on student growth. Although logistic growth curves might better represent students' 'typical' reading trajectories, adoption of a logistic growth curve also might lead to less ambitious teacher expectations, and in turn, to less intensive instruction in the latter part of the school year. Research on CBM has demonstrated that when teachers have higher expectations and set more ambitious goals, students learn more (Allinder, 1995; L. S. Fuchs, 1989). It may be especially important for students who struggle to maintain good quality reading instruction throughout the school year. If teachers are pressed to provide students the instruction they need, and to ignore possible slow incremental reading growth toward the end of the school year, it might lead to better student performance.

Validity of Maze Scores

The third research question addressed the validity of maze scores as indicators of reading performance level and growth. We examined four types of relations: 1) maze performance level with performance level on criterion measures, 2) maze growth with performance level on criterion measures, 3) maze performance level with growth on criterion measures, and 4) maze growth with growth on criterion measures. On the whole, our results revealed that initial maze performance level and growth was related to performance level and growth on criterion reading measures. These results were consistent with what was found in previous studies (Espin et al., 2010; Tichá et al., 2009; Tolar et al., 2014; Tolar et al., 2012). However, our result that growth on maze was related to growth on scores on the reading test was similar to the results of Tichá et al. (2009), but not to Tolar et al. (2014). These inconsistencies across studies might be explained by the fact that it is more difficult to statistically establish relations between more complicated relations (i.e., between two growth curves) as opposed to somewhat simpler relations (i.e., between growth and static scores or between static scores). The effects of more complicated relations could be influenced by several factors, including the chosen analysis method, the technical adequacy of scores on criterion measures and its sensitivity in detecting reading growth, or the assumption that reading growth is linear for the criterion measures (when using only pre- and posttest scores, see, for example, McArdle, Grimm, Hamagami, Bowles, and Meredith, 2009, for a discussion on factors that should be considered when modeling growth data). There is a need to examine this issue more closely in future research.

Limitations

One limitation of the study is that we administered the maze in the second half of the school year, so the results could only be based on the second part of the school year. Our purpose

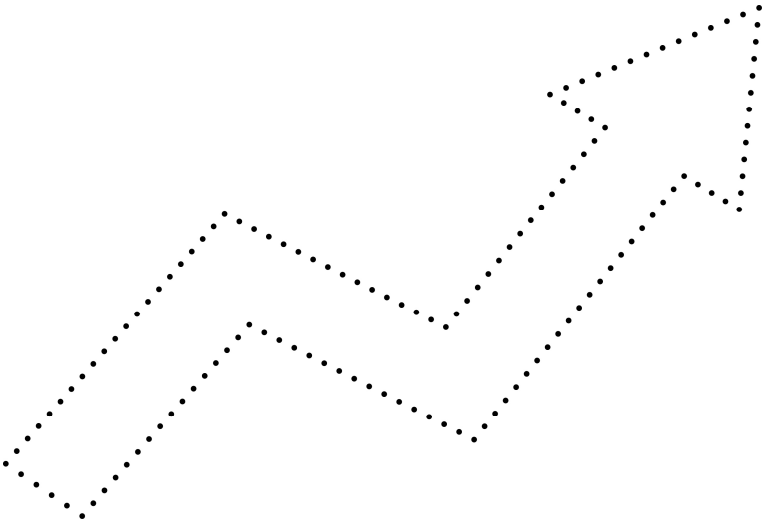
was to examine the technical adequacy of maze scores that could be used within a CBM framework in which frequent measurement is required. Such research is costly and time-intensive, thus the choice was made on first examining the technical adequacy of maze scores for a somewhat shorter period of time. A second limitation concerns the inclusion of a relatively large group of students with low proficiency. This means that our results best generalize to this population. This may also explain why the estimated weekly growth was rather low. Given that CBM is especially suitable for students who are in need of more intensive instruction, thus the low performing students, it is good that the results could be interpreted for this group of students. Future research should focus on a more representative sample administered throughout the whole school year to establish normative scores for maze reading level and growth. A third limitation is that no data was gathered on the amount and quality of the reading instruction. We recommend that further investigations into the technical adequacy of CBM maze growth rates also include data on the quality of reading instruction within the classroom.

Conclusion and Future Directions

In conclusion, our results provide support for the reliability, sensitivity to growth and validity of maze scores for secondary-school students. Yet, several questions remain and need to be examined more closely in future research. A number of these points already have been raised throughout the discussion, including the use of a more representative sample, the examination of reading growth throughout the school year, and the inclusion of the quality of the provided reading instruction. In addition, future research should focus on how maze growth can be used for its intended purpose; that is, to examine if and how secondary-school educators use maze growth to evaluate the effectiveness of reading instruction and inform their instructional decisions for older struggling readers.

Chapter 5

Stability of maze reading slopes: Implications for use of CBM maze scores for tracking the progress of secondary-school students



Partly based on:

Siuman Chung

Christine Espin

Claire Stevenson

Manuscript in preparation

Abstract

Curriculum-based measurement (CBM) is a procedure that teachers can use to monitor student progress and evaluate the effectiveness of instruction. Despite the large body of literature supporting the technical adequacy of CBM measures, an area of concern remains the stability of the individual slopes or growth rates produced by the measures. Most research on this topic has focused on scores from CBM reading-aloud measures. In this paper, the stability of slopes produced by scores from CBM maze measures is examined. Specifically, we examine (1) the influence of duration, schedule, and variation in maze scores on the stability of CBM maze reading slopes, and (2) the effects of assuming a linear vs. nonlinear growth rate on SEE. For research question 1, stability, operationalized as the standard error of the slope (SEb), was calculated for different durations, schedules, and levels of SEE. Results revealed that the denser the schedule and lower the SEE, the shorter duration needed to achieve stable slopes. Large instability was seen in the first few weeks of data collection, regardless of schedule and SEE. For research question 2, the effects of assuming linear vs. nonlinear growth rates were examined. Assuming a nonlinear growth rate did not improve SEE. Implications for teachers' use CBM maze scores as indicators of growth in reading are discussed.

Introduction

Curriculum-Based Measurement (CBM) is a tool that teachers can use to monitor students' learning progress and evaluate the effectiveness of instruction on that progress (Deno, 1985, 2003). Short, timed CBM measures are administered frequently, and the scores are used to represent students' academic progress. For each student, CBM scores are plotted in a progress graph in which a linear growth line, referred to as the slope line, is drawn. Teachers read and interpret the graphs to determine whether the student is improving/growing at a desirable rate, and thus, whether the instruction is effective. If growth is not at a desirable rate, instruction is changed. If growth is at a rate better than expected, the goal is raised.

Over the last 40 years, a large amount of research has been conducted on the technical adequacy of CBM measures across various academic areas, including reading, mathematics, written expression, and content-area learning (see, for example, reviews by Wayman et al., 2007; Foegen et al., 2007; McMaster & Espin, 2007; Espin et al., in press, or see also Chapter 2). The majority of this research has focused on the technical adequacy of CBM scores as performance level indicators, that is, on whether "static" scores from the measures adequately represent students' levels of performance. This research is referred to by L. S. Fuchs (2004) as *Stage 1* research. Much less research has been conducted at what L. S. Fuchs (2004) refers to as *Stage* 2. Stage 2 research focuses on the technical adequacy of CBM scores as progress or growth indicators, that is, on whether the slopes generated from scores on frequently administered CBM probes provide stable, sensitive, and valid estimates of growth (see Wayman et al., 2007).

The paucity of research at Stage 2 might be due in part to the technical challenges associated with carrying out this research. It is only in recent years that advancements in technology and statistical methods have made it possible to conduct in-depth and comprehensive investigation of the properties of CBM slopes. Nonetheless, as early as the mid-1980s and 1990s there were attempts to examine slope. For example, Good and Shinn (1990) and Shinn, Good, and Stein (1989) compared ordinary-least-squares and split-middle techniques for estimating reading growth with CBM reading-aloud scores, and found that ordinary-least-squares was more accurate and precise for growth estimation than was the split-middle technique. Skiba, Deno, Marston, and Wesson (1986) reported on the characteristics of level, slope, and standard error of estimate (SEE) for CBM reading-aloud data for students in grades 2 through 5, providing typical SEE levels at each grade level. L. S. Fuchs & Fuchs (1992) examined the ratio of SEE to the mean growth rate for scores from two different types of CBM reading measures: reading-aloud and maze-selection. (A maze is a text in which every seventh word is replaced by a multiple-choice item with one correct and two incorrect choices. Students read silently and select words as they read.) Results of L. S. Fuchs and Fuchs (1992) demonstrated that the ratio of SEE to mean growth was similar for scores from the two measures.

In recent years, advances in technology and statistics have allowed for a more comprehensive look at the technical characteristics of CBM slopes. Most of this research has focused on students at the elementary-school level and on CBM reading-aloud measures (see Petscher, Cummings, Biancarosa, & Fien, 2013; Wayman et al., 2007, and work by Christ and colleagues reviewed later in the paper). Much less research has focused on students at the secondary-school level or has included CBM maze-selection measures.

Research on CBM Reading Growth at the Secondary-School Level

A handful of studies have examined the technical adequacy of reading growth trajectories produced by frequent (weekly) CBM measurement at the secondary-school level (Chung, Espin, & Stevenson, 2018, or see also Chapter 4; Espin et al., 2010; Tichá et al., 2009).[8] Espin et al. (2010) and Tichá et al. (2009) examined the reliability, sensitivity, and validity of reading growth produced by scores from two CBM measures – reading aloud (also referred to as oral reading fluency) and maze – for 31 and 35 8th-grade students respectively. For reading aloud, students read aloud for 1 to 3 min from a passage, and the number of words read correctly (WRC) was scored. For maze, students read silently for 2 to 4 min from a maze passage, and the number of correct maze choices (CMC) was scored. In both studies, growth was monitored for 10 weeks, with one reading-aloud and one maze probe administered per week.

Alternate-form reliability coefficients, calculated across scores on adjacent weeks, were on average r = .86 for 1-min reading aloud and r = .80 for 3-min maze scores (Espin et al., 2010).[9] Hierarchical linear modeling (HLM) was used to examine sensitivity and validity of the growth trajectories. Sensitivity analyses focused on whether the mean growth rate was significantly different from zero, which would indicate that the measures were sensitive to change in performance. Scores from reading-aloud reflected minimal to no change over time, whereas the scores from maze reflected significant growth at 2-, 3-, and 4-minute time frames (Espin et al., 2010; Tichá et al., 2009). The validity of the growth trajectories was then examined to determine whether growth on CBM maze was related to performance level and growth on the criterion measures. The CBM maze growth trajectories were found to be significantly related to performance level and growth on standardized reading measures and to reading group membership (Espin et al., 2010; Tichá et al., 2009).

Although the results from Espin et al. (2010) and Tichá et al. (2009) provided initial support for the use of maze to monitor reading growth for secondary-school students, the studies involved small sample sizes and included only 10 weeks of progress monitoring.

[8] Other studies at the secondary-school level have examined growth, but have not included weekly measurements (Codding et al., 2015; Tolar et al., 2014; Tolar et al., 2012; Yeo et al., 2012). However, frequent data collection is necessary if CBM progress monitoring data are to be used for ongoing evaluation of instruction for students with learning difficulties as illustrated in the opening paragraph of this paper. Thus, in the current study we only reviewed the studies that included frequent measurements.

[9] In the study by Tichá et al. (2009), the alternate-form reliabilities were reported for scores on reading-aloud and maze measures that have been administered at one measurement point, and were on average r = .96 for 1-min reading-aloud scores and r = .87 for 3-min maze scores.

Thus, in a subsequent study, Chung et al. (2018, or see also Chapter 4) replicated this early research with a larger sample and for a longer progress-monitoring time period. In addition, Chung et al. (2018, or see also Chapter 4) fit both linear and nonlinear growth models to the data. Participants in the study were 452 7th-grade students who completed one 2-min maze task each week for 23 weeks. Alternate-form reliability, calculated for maze scores on adjacent administration weeks, was on average r = .75. Results of multilevel analyses revealed that both linear and nonlinear (logistic) growth models fit the data, although the fit was better for the nonlinear than the linear growth model. For both linear and nonlinear models, level and growth on the maze were found to be significantly related to scores on multiple criterion measures, including school level, dyslexia status, and performance level and growth on three external reading measures. The results of Chung et al. (2018, or see also Chapter 4), Espin et al. (2010), and Tichá et al. (2009) provide support for the alternate-form reliability, sensitivity, and validity of maze scores as indicators of reading growth for secondary-school students. However, missing from these studies is consideration of one of the most essential factors in CBM growth measurement – stability. Stability refers to the magnitude of measurement error associated with reading slopes for individual students. The larger the measurement error associated with a slope, the less likely it is that the slope is reflecting "true" growth (Hintze & Christ, 2004). Stability is important within CBM growth measurement to answer questions related to how often and how long students must be monitored before the slopes generated by the CBM measures are "trustworthy". These questions lie at the core of CBM progress monitoring because for teachers to use CBM progress data to inform instruction, they must be able to use the data in a "timely fashion" to make instructional decisions about the effectiveness of their interventions, and to change instruction if it is not effective. In recent years, Christ and colleagues (Christ, 2006; Christ, Monaghen, Zopluoglu, & van Norman, 2012; Christ, Zopluoglu, Long, & Monaghen, 2012; Christ et al., 2013) have examined the question of slope stability for CBM reading-aloud measures at the elementary-school level.

Stability of Slopes for Scores from CBM Reading-Aloud Measures

In 2006, Christ and colleagues began a series of studies designed to examine the stability of slopes generated by CBM reading-aloud scores, and to determine factors that influenced that stability (Christ, 2006; Christ, Monaghen, et al., 2012; Christ, Zopluoglu, et al., 2012; Christ et al., 2013). The factors examined in the studies included duration, schedule, and variation in scores. Duration was the number of weeks of progress monitoring. Schedule was the frequency of progress monitoring, and included the number of occasions (the number of measurements, per week or per month), and the number of observations (the number of passages administered per occasion). Variation was the intra-individual variation in the data around the slopes, and was reported as either SEE or residual variance in scores.

In the first in this series of studies, Christ (2006) examined the effects of duration and SEE on the stability of the slopes. Results revealed that the lower the SEE and the longer the data-collection duration, the more stable the slopes; however, the relation between these variables was curvilinear. Thus, with only a few weeks of data collection (for example, 2 to 4 weeks), SEE had a large influence on the stability, whereas for longer durations (for example, 12 to 15 weeks), SEE had minimal influence on stability.

In subsequent studies, Christ and colleagues (Christ, Monaghen, et al., 2012; Christ, Zopluoglu, et al., 2012; Christ et al., 2013) used a simulated data set with reading-aloud scores for 9000 hypothetical students. Duration and variation were again examined, and a third factor, schedule of data collection, was also considered (Christ et al., 2013). Schedules of data collection included one, two, three or five measurement per week or month (occasions) with either one or three measures administered per occasion (observations per occasion). In addition, reliability and validity coefficients of the slopes were estimated for different durations, schedules, and levels of score variation.

In general, the results of the simulation studies with regard to duration and variation were similar to those found in Christ (2006). The stability of the reading-aloud slopes showed a similar curvilinear trend, with high levels of instability found for shorter progress-monitoring durations and stability increasing rapidly with longer durations. Lower levels of score variation resulted in more stable, reliable and valid reading slopes (Christ, Monaghen, et al., 2012; Christ, Zopluoglu, et al., 2012; Christ et al., 2013). With regard to schedule, results revealed that the denser the schedule, the more stable, reliable, and valid the reading slopes. For instance, when score variation was low, reliable and valid slopes were seen after only 8 weeks of data collection with 5 to 6 observations per week, after 10 weeks of data collection with 3 observations per week, and after 13 weeks with 3 observations per month (Christ et al., 2013).

In sum, the studies by Christ and colleagues demonstrate that duration, schedule, and variation in scores affect the stability of the slopes generated by CBM reading-aloud scores for elementary-school students. What is not yet known is whether and to what extent such factors influence the stability of the slopes generated by CBM maze scores for secondary-school students. In addition, the studies done by Christ and colleagues all assumed linear growth models, yet growth on CBM measures may be nonlinear. For example, studies have shown that growth on CBM measures is greater from fall-to-winter than from winter-to-spring (e.g., Christ et al., 2010; Keller-Margulis, Mercer, Payan, & McGee, 2015; Nese et al., 2012; Shin, Espin, Deno, & McConnell, 2004). Assuming a nonlinear growth model might result in a decrease in score variation.

Purpose of the Study and Research Questions

The purpose of the present study is to examine the stability of the growth trajectories generated by scores on frequently administered CBM maze measures. The study involves a

reanalysis of data from Chung et al. (2018, or see also Chapter 4), and examines the influence of duration, schedules of data collection, and variation (levels of SEE) on slope stability. We adopt the approach used by Christ and colleagues in their examination of slope stability for CBM reading-aloud measures, but extend their work by examining the effects on SEE of assuming a nonlinear growth model. Two research questions are addressed:

1. What are the effects of duration, variation in maze scores, and schedule of data collection on the stability of CBM maze slopes?
2. Can standard error of the estimate be decreased by assuming a nonlinear (logistic) growth rate rather than a linear growth rate?

Method

The current study was a reanalysis of data from Chung et al. (2018, or see also Chapter 4). A short overview of the participants, instruments and procedure relevant to this study are described here. For more elaborate details, see Chung et al. (2018, or see also Chapter 4).

Participants

Participants were 452 (233 male) Dutch 7th-grade students from three secondary schools. Students were *M* = 12.63 (*SD* = 0.63; range: 12-15) years old. Participants spanned a wide range of performance/school levels. In the Netherlands, secondary schools are divided into four levels: practical education, pre-vocational, senior general secondary, and pre-university education (ordered from highest to lowest, Dutch Ministry of Education, Culture and Science, 2005). The number of participants per level were: practical (*n* = 55), pre-vocational (*n* = 321), and senior general secondary/pre-university (*n* = 76). Students with dyslexia (12%) were represented in all school levels.

Instruments and Procedure

Student growth was measured using CBM maze tasks. A maze is a reading passage in which every seventh word is deleted and replaced by a multiple-choice item with the correct word and two distractors. Students read silently through the passage for 2 min selecting words as they read. The number of correct choices was scored.

Mazes were administered and scored through an online progress-monitoring system called *Mazesonline* (http://www.mazesonline.nl). Students completed one maze per week for a period of 23 weeks. The maze was not administered during vacation or exam weeks, resulting in a total passage set of 17 maze passages.

Data Analyses

Variation in maze scores

Based on the observed maze scores, the beginning level of performance at the start of the study (intercept) and the rate of growth (slope) were calculated for each individual student, after which the variation around the slope (i.e., standard error of the estimate; SEE) was estimated per student. The SEE was estimated by applying the following sample estimation formula:

$$SEE_i = \sqrt{\frac{\sum(Y_{it} - Y'_{it})^2}{n_i - 2}}, \quad (1)$$

where Y_{it} was the observed maze score of individual i at time t, and Y_{it}' was the predicted maze score of individual i at time t, and n_i was the total number of maze tasks administered to individual i. We calculated the mean SEE from the sample, and used this level of SEE to represent typical levels of SEE for 7th-grade students completing mazes on a weekly basis. We then varied SEE relative to this typical level.

Stability of maze slopes

To examine the influence of duration, schedule and various levels of SEE on the stability of maze growth rates, we calculated the standard error of the slope (SEb; stability) by applying the following formula for various scenarios:

$$\text{SEb} = \frac{SEE}{SD_x\sqrt{n}}, \quad (2)$$

where *SEE* was the standard error of the estimate, x was the duration, and n was the number of maze tasks collected. The SEb was calculated in Excel for SEEs ranging from two to five, and durations of up to 30 weeks.[10] Schedules included one observation twice a week, once a week, once in two weeks, and once a month.

Linear vs. nonlinear growth

To address research question 2, SEE was compared for linear vs. nonlinear (logistic) growth. For both functional forms (linear and nonlinear), two approaches were used to estimate the predicted slope lines of individual students: ordinary least squares regression (OLS) and multilevel modeling (ML; Hox, 2010). Both estimation approaches represent individual learning growth, but differ in practicality and the accuracy of "true" learning growth. While

[10] We first calculated the daily SEb, and then converted this to weekly SEb by multiplying the daily SEb by 7. For all schedules, the first administration was set on day one. The twice a week schedule was set with intervals of 3 to 4 days, thus administrations occurred on day 1, 4, 8, 11, 15, etc. The once a week schedule was set with intervals of 7 days; thus administrations occurred on day 1, 8, 15, 22, 29, etc. The once every two weeks schedule was set with intervals of 14 days, thus administrations occurred on day 1, 15, 29, 43, 57, etc. The once a month schedule was set with intervals of 30 to 31 days, thus administrations occurred on day 1, 32, 62, 93, 123, etc. Missing values were imputed by interpolating the data with Excel.

ML is more accurate in representing "true" growth, because the method controls for missing data and dependence between measurements (Hox, 2010), OLS is more practical, because it can be used more easily in the educational practice by teachers. OLS slopes can be graphed from individual student's scores by using, for instance, free templates that are available online (e.g., http://www.measuredeffects.com). For graphing ML slopes, additional data from a large representative sample is required, and thus, in practice, less easy to use. Because both arguments – accuracy of "true" learning progress and practical use – are important, both estimation approaches were included in the analyses. The difference in SEE for linear vs. nonlinear growth were conducted separately per estimation approach (OLS and ML). We will describe both estimation approaches in more detail.

OLS consisted of only a within-individual level, whereas ML consisted of both within- and between-individual levels. At the within-individual level, the following formulas were applied to estimate the predicted maze scores for linear and nonlinear (logistic) growth respectively:

$$Y'_{it} = \pi_{0i} + \pi_{1i} * maze\ session_{it}, \text{ and} \quad (1)$$

$$Y'_{it} = \pi_{0i} + \pi_{1i} (\ln(maze\ session_{it} + 1)) \quad (2)$$

where the π_{0i} and π_{1i} were respectively the intercept and slope of the predicted growth line in both formulas. The between-individual level expanded the formulas above with:

$$\pi_{0i} = \beta_{00} + u_{0i} \quad (3)$$

$$\pi_{1i} = \beta_{10} + u_{1i} \quad (4)$$

where β_{00} was the overall mean intercept to predict π_{0i} from a between-individual level for individual i. The u_{0i} was the error term for π_{0i}, thus the difference between the overall mean intercept and the intercept of individual *i*. The β_{10} was the mean slope to predict π_{1i} from a between-individual level for individual i. The u_{1i} was the error term for π_{1i}, thus the difference between the overall mean slope and slope of individual *i*.

Differences in standard errors of estimate (SEEs) were examined with the non-parametric Wilcoxon Signed Rank test. Wilcoxon Signed Rank tests were used because assumptions for the paired samples t-tests were not met. The formula used to estimate SEE consisted of an observed maze score Y_{it} and a predicted maze score Y_{it}' of individual *i* at time *t*, see equation 1.

The intercepts and slopes for linear and nonlinear growth for OLS and ML estimation were calculated by using the packages plyr (Wickham, 2011) and lme4 (Bates & Maechler, 2010) in the software program R. The SEEs were calculated in plyr in R or manually with Excel. The Wilcoxon Signed Rank tests were performed in IBM SPSS.

Results

Selecting "Typical" and "Optimal" Standard Error of the Estimates (SEEs)

We first determined which values to use to represent "typical" and "optimal" SEEs. The standard error of the estimate (SEE) of reading slopes for the weekly administered 2-min maze passages for the sample was *M* = 4.21 (*SD* = 1.67), with a range of 0.92 – 13.01 SEE, with 73.5% of the cases falling below an SEE = 5. Based on the obtained mean, SEEs of 4 and 5 were selected to represent "typical" SEEs, whereas SEEs of 2 and 3 were used to represent more "optimal" SEEs. (Recall that a lower SEE indicates that there is less error between observed and "true" scores).

Stability of Reading Slopes

To examine the stability of maze slopes, the standard error of the slope (SEb) was calculated for different durations, levels of SEEs, and schedules. We report on the effects of duration and SEE, and on duration and schedule, on the stability of maze slopes.

Duration and SEE

In Figure 5.1 the SEb's are plotted for weekly-administered maze passages with more optimal (SEE= 2 and 3) and typical (SEE= 4 and 5) SEEs. At all levels of SEE, curves showed a steep decelerating trend in the first weeks, followed by stable straight lines thereafter. Stable lines were achieved sooner (i.e., SEb was lower) for lower levels of SEE than for higher levels of SEE. For instance, given an optimal SEE of 2, the SEb at week 8 was 0.26, and reductions in SEb were minimal (smaller than 0.05) each week thereafter. In contrast, given a typical SEE of 4, the SEb at week 8 was 0.52, and it was not until week 10 that the weekly reduction was smaller than 0.05 per week. At week 22 the differences in SEb for different levels of SEE were negligible (less than 0.10).

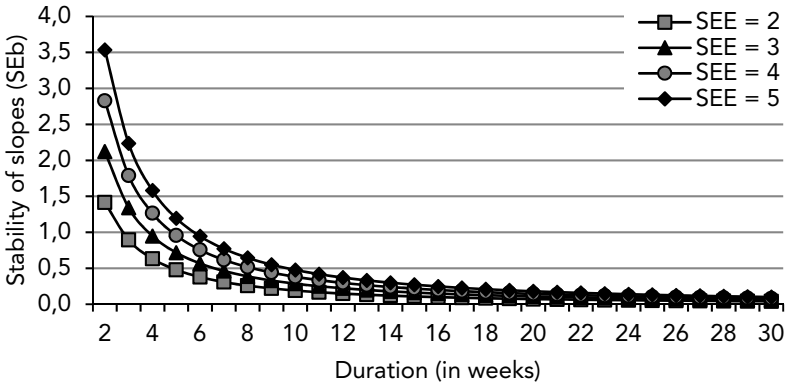


Figure 5.1. Stability of the slopes (SEb) for different SEEs on weekly administered maze passages

Duration and schedule

In Figure 5.2 the stability of the slopes for different schedules are plotted given a typical SEE of 4. For all schedules, curves showed a steep decelerating trend in the first weeks, followed by stable straight lines thereafter. Stable lines were achieved sooner (i.e., SEb was lower) for higher density schedules. For instance, given a dense schedule of data collection (twice per week), the SEb at week 9 was 0.34, and reductions in SEb were minimal (smaller than 0.05) each week thereafter. In contrast, given a less dense, but more typical schedule of data collection (once per week), the SEb at week 9 was 0.44, and it was not until week 10 that the weekly reduction was smaller than 0.05 per week. At week 27 the differences in SEb for different data collection schedules were negligible (less than 0.10).

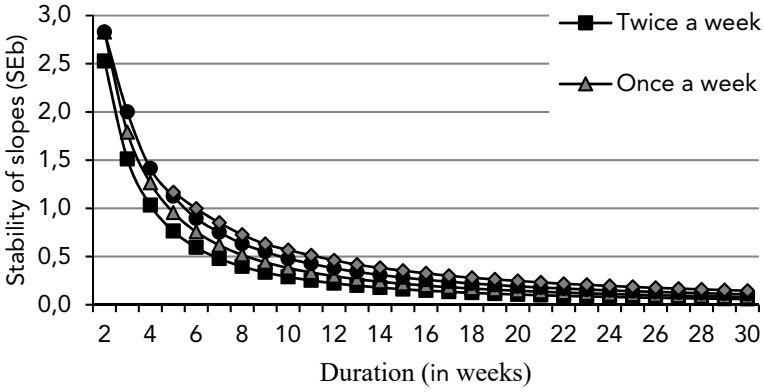


Figure 5.2. Stability of the slopes (SEb) for different schedules and SEE = 4

Linear vs. Nonlinear Growth

In Table 5.1 descriptive statistics are provided for the intercepts, slopes, and SEE for linear and nonlinear growth as estimated with ordinary least squares regression (OLS) and multilevel modeling (ML). Average SEEs ranged from 4.2 to 4.5, and were similar for linear and nonlinear growth. The obtained SEE was higher for ML estimation (SEE = 4.5) than for OLS estimation (SEE = 4.2). The range of the SEE was large, ranging from approximately 1 to 15. Results revealed that SEEs did not differ significantly for linear and nonlinear growth using an OLS estimation (Z = -.82, p = .41) or ML estimation (Z = -1.19, p = .24).

Table 5.1
Intercepts, slopes, and SEEs for linear and nonlinear slopes with OLS and ML estimation

	M	*SD*	Min	Max
OLS				
Intercept linear	25.70	7.35	9.47	52.32
Slope linear	0.16	0.39	-1.19	2.20
Intercept nonlinear	24.77	7.29	6.43	51.06
Slope nonlinear	1.16	2.20	-7.06	10.87
SEE linear	4.21	1.67	0.92	13.01
SEE nonlinear	4.23	1.68	1.01	12.64
ML				
Intercept linear	25.71	6.40	11.36	47.66
Slope linear	0.15	0.14	-0.40	0.62
Intercept nonlinear	24.70	5.83	11.87	44.70
Slope nonlinear	1.20	0.94	-2.14	4.22
SEE linear	4.50	1.74	0.94	14.50
SEE nonlinear	4.50	1.75	1.14	14.52

Discussion

This research addressed the second stage of CBM research (L. S. Fuchs, 2004) - the technical adequacy of CBM slopes for secondary-school students. Specifically, we examined the stability of the reading slopes produced by maze scores and factors that affect that stability. We focused on the influence of duration, schedule, and variation of maze scores (SEE) on the stability of maze slopes, and examined whether assumption of a nonlinear growth model would result in reduced variation in the data.

Results revealed that the stability of maze slopes was affected by duration, schedule, and variation in maze scores (SEE). The denser the schedule and the lower the variation in scores, the shorter the duration needed to achieve stable maze slopes. Relatively large instability was seen in the first few weeks of data collection, regardless of schedule or variation in scores. These results are similar to those of Christ and colleagues in the sense that duration, schedule, and variation in scores affected the stability of the slopes (Christ, 2006; Christ et al., 2013), and that stability increased rapidly with longer duration.

With regard to the effects of SEE and duration, given a schedule of once a week data collection, it was not until approximately weeks 8 to 11 that the effects of SEE began to decrease for maze slopes. At week 12, the decrease of stability for each week was minimal. Such results suggest that if monitoring is being conducted with a known passage set that has been shown to produce relatively low levels of SEE for a given population, then instructional decisions could reasonably be made on the basis of slope after a relatively short period of time, for example, perhaps after only 9 or 10 weeks of progress monitoring. If, however, the SEE for a passage set is unknown for a given population, then it would be advisable to monitor the progress for a longer period of time, for example, 12 to 14 weeks, before making instructional decisions on the basis of the maze slope.

Decisions about duration can also be influenced by the schedule of data collection. Given an SEE of 4, if data are collected on a "thin" schedule, for example, once every two weeks or once a month, then it is necessary to collect data over a longer period of time before slopes stabilize. However, for denser schedules of data collection, for example, once or twice a week, slopes stabilize earlier (that is, stability does not decrease substantially with additional weeks of data collection). Thus, for thin schedules of data collection, slopes do not stabilize until about 12 to 13 weeks, whereas for dense schedules of data collection, slopes stabilize at about 10 to 11 weeks. Differences in slope stability between once- and twice-a-week data collection disappear fairly quickly, with only with minimal differences seen between the two schedules of data collection at 10 to 11 weeks.

Although we found that after approximately 12 weeks, slope stability levels off, it is still worrisome that the standard error of the slope given a "typical" SEE of 4 was about 0.3 for once- or twice-a-week data collection. Although 0.3 seems quite small, it is large when one considers that the mean rate of growth for the 7th-grade students included in this study was approximately 0.2 correct maze choices per week (Chung et al., 2018, or see also Chapter 4). Such variation could lead to incorrect instructional decisions for some students. For instance, if a student (with an average amount of variation in his or her slope) has an observed growth rate of 0.3 correct maze choices, the "true" rate of growth might lie between 0 to 0.5. An educator might decide to not change the student's instruction based on the observed growth rate of growth of 0.3 given that this growth exceeds the typical rate of growth for students at this age; however, if the "true" rate of growth for the student is actually 0, then the correct decision would be to change instruction. Similarly, educators may unnecessarily change instruction that has been effective if the student's "true" growth rate exceeds their observed growth rate. An additional consideration is the fact that the *average* SEE was approximately 4. As mentioned earlier, about 25% of the sample had SEEs of 5 or higher, with some students having SEEs as high as 13. The chances of incorrect instructional decision-making is especially large for students with much variation in their slopes.

To address problems of variation, one can implement longer progress-monitoring durations with denser schedules of data collection, but it is also important to try and reduce the maze variation seen in slopes. As can be seen in the current study, the less variation, the faster stable maze slopes are achieved. One way to reduce variation in maze slopes is to provide optimal conditions within the classroom when CBM mazes are administered. For instance, educators should be sure to administer the probes according to guidelines and as intended, to provide a quiet setting in which students can complete the CBM mazes with enough concentration, to motivate students' to do their best, and to carefully score and graph the CBM maze scores (Haladyna & Downing, 2004). In addition, there is also room for improvement for developers and/or (future) researchers to reduce variation in maze scores via attention to the characteristics of the passages used to monitor

progress. One factor that likely influences the variation in maze scores is the extent in which frequently administered passages are equivalent to one another. Approaches to addressing problems related to passage characteristics have been described for scores from CBM reading-aloud measures, and include removing outlier passages from a passage set in which scores on passages had the largest Euclidean distance (Christ & Ardoin, 2009), or by statistical equating of the scores in the passage set (Albano & Rodriguez, 2012; Santi, Barr, Khalaf, & Francis, 2016).

Although reducing the variation in maze slopes may increase the stability, our data reveal that even for low levels of SEE, for example SEE = 2, the standard error of the slope is still approximately 0.15. Considering a mean growth rate of 0.2, this standard error might still be considered to be too large. It will be important in future research to examine ways to increase the rates of growth typically seen for maze data. One obvious approach is to adjust the maze passage slightly. For instance, by increasing the number of possible maze choices per passage by, for example, replacing every fifth word in the passage rather than every seventh word, scores per passage would be greater. It also might be possible to administer longer maze passages with a longer administration time, or to administer more passages at each occasion. A different approach would be to support the maze score with an additional score from a different CBM probe, and to combine the two scores as one indicator of reading progress. Both approaches would increase the score on the CBM progress measure, and possible also the growth rates. However, it is not yet known whether the amount of variations in the slopes would remain constant, or whether such measures would be reliable and valid indicators of reading level and growth, thus additional research should be conducted if these approaches would be considered.

A final consideration with regard to growth rates is to examine growth under "optimal" rather than under "typical" conditions. In our study, we did not implement reading interventions for participating students. It may be that under "optimal" instructional conditions where students receive evidence-based reading instruction, their rates of growth on the CBM measures would be much larger. Supporting this idea is research by Deno et al. (2001) who demonstrated that growth on CBM reading measures was much larger for students with learning disabilities under optimal instructional conditions than under typical instructional conditions.

A second purpose of our study was to examine whether assumption of a nonlinear growth rate would result in reduced SEE. Recent studies have suggested that CBM growth measures might not be linear but nonlinear (Christ et al., 2010; Chung et al., 2018, or see also Chapter 4; Keller-Margulis et al., 2015; Nese et al., 2012; Shin et al., 2004); If students' reading growth is in fact nonlinear, then the variation in scores around the nonlinear growth line (SEE) might be smaller. Our results revealed, however, that variation was not reduced by assuming a nonlinear growth model. The variation in maze scores were the

same for both linear and nonlinear growth models, regardless of whether OLS or ML estimation approaches were used.

Our results are seemingly at odds with the original data analyses by Chung et al. (2018, or see also Chapter 4), which demonstrated that a nonlinear growth model provided a better fit to the data than a linear growth model. The seemingly conflicting results might be explained by the choice of analysis method. Our study used the Wilcoxon Signed Rank test that tests differences based on matched pairs of residuals (SEE linear and nonlinear), whereas Chung et al. (2018, or see also Chapter 4) compared fit indices that were based on the sum of residuals for the two full growth models (AIC, BIC, Deviance of linear vs. nonlinear). At the very least, the results of the present study support the use of a linear growth model for monitoring students' progress with CBM measures. A linear growth model is more practical in terms of generating progress graphs and ease of interpretation for educators. It is worthwhile to note however, that in a recent study that compared linear and nonlinear growth for students, not all students showed one type of growth trajectory. On reading-aloud scores, some students followed a linear learning curve, whereas other students followed a nonlinear learning pattern (van Norman & Parker, 2016). Such results indicate that variation in scores might be smaller if the individual growth trajectories were taken into account. The issue of individual growth trajectories should be considered in future research.

Limitations

A limitation of the study was the selection of "typical" levels of the variation in maze scores (SEE). These levels would preferably be set based on a range of studies and data sets to be able to gauge what the expected amount of variation in maze scores is for secondary-school students. However, to date there are only a handful studies that have examined reading growth for secondary-school students. Of those studies, only one has reported the SEE of the reading slopes (Tolar et al., 2012), but in that study reading growth was based on only five measurements occasions throughout the school year. If more data were available, "typical" levels of SEE might be different. That said, results related to the effects of duration, variation, and schedule on the stability of the slopes would not change. More research is needed on the use of CBM reading measures to monitor the growth of secondary-school students, and especially to establish levels of SEE for student growth.

Conclusions and Future Directions

The current study extended CBM research at the secondary-school level, and expanded the research conducted at the second stage (technical adequacy of CBM reading slopes) of CBM measure development. This study was the first to examine the stability of maze slopes for secondary-school students and to examine an approach for reducing variation in maze scores. The development of CBM progress monitoring in reading for secondary-

school students is a relatively new area of research, and many questions remain. Our study illustrates the need to examine methods for reducing variation in maze scores and thereby improving slope stability. It also illustrates the importance of examining growth under optimal rather than just typical instructional conditions. Finally, our present study focused solely on the technical characteristics of scores from the CBM measures. It did not address the essential question of CBM progress monitoring: *Does teacher use of CBM progress-monitoring result in improved reading instruction for and improved performance of students with learning disabilities?* Using the CBM progress data for instructional decision-making is the essence of CBM. This issue has yet to be addressed at the secondary-school level.

Chapter 6

Summary and general discussion

Introduction

Curriculum-Based Measurement (CBM) is a progress-monitoring system designed to be used by educators to screen performance, to monitor growth, and to evaluate the effectiveness of interventions on the growth of students with learning difficulties (Deno, 1985, 2003). Implementation of CBM for monitoring growth and evaluating instruction involves frequent measurement of student performance. Scores from the measures are placed on graphs that depict growth over time and provide information to educators about how effective instruction has been.

Although there has been a large body of research conducted on CBM (Espin et al., in press, or see also Chapter 2; Foegen et al., 2007; McMaster & Espin, 2007; Wayman et al., 2007), there are notable gaps. First and foremost, relatively few studies have been conducted at the secondary-school level, and of those, few have focused on reading, and none on foreign-language learning. Second, relatively few studies have been conducted at what L. S. Fuchs (2004) refers to as Stage 2 or Stage 3 CBM research. Stage 2 studies focus on the technical adequacy of scores from the measures as indicators of growth, and Stage 3 on the instructional utility of CBM, that is, on the effects of implementing CBM on teacher instruction and student achievement. The majority of CBM studies have been Stage 1 studies, which focus on the technical features of static (performance) scores at one point in time. The disproportionate focus on Stage 1 may be due in part to the popularity of using CBM measures for screening, or to the fact that studies at Stage 1 are less costly and time-intensive than studies at Stages 2 or 3 (L. S. Fuchs, 2004; Wayman et al., 2007). In addition, Stage 1 studies are often the first step in CBM research: promising measures are identified in Stage 1 studies, which then can be more closely examined in the more costly and time-intensive Stage 2 and 3 studies.

The studies in this dissertation addressed gaps in CBM research. First, the research focused on the development of CBM progress-monitoring measures at the secondary-school level in reading and foreign-language learning. Second, the studies in reading were Stage 2 studies, and focused on the technical features of the slope/growth rates produced by reading CBM measures.

The essential characteristics of CBM measures as outlined by Deno (1985) served as a framework for this research project. First, the measures had to be simple, efficient, and inexpensive so they could be used by educators to frequently and repeatedly (for example, weekly) monitor student growth. Second, scores from the measures had to be easy to understand so that the results could be clearly and correctly communicated to parents, teachers or students. Finally, and most importantly, scores from the measures had to be reliable and valid *indicators* of performance level and growth in an academic area so that improvement in scores reflected general improvement in the academic area. In this way, the measures could be used by educators to monitor the general growth of students in an

academic area (for example, reading), and to evaluate the effects of their instruction on that growth. In our studies we wanted to obtain a broad view of the validity of CBM scores by adopting Messick's (1989) recommendation to build evidence for validity across various criterion variables. Thus, we included multiple criterion variables that were assumed to represent performance level or growth in the academic area of interest and examined the pattern of relations of the criterion measures with the CBM scores.

In the following sections, the results of the studies included in this dissertation are briefly summarized and reviewed. This is followed by a general discussion of the implications of the findings.

Development of CBM Measures in Foreign-Language Learning

To our knowledge, Chapter 3 on the development of CBM measures in foreign-language learning was the first CBM study to be published in this area. The study was Stage 1 CBM research (L. S. Fuchs, 2004) and focused on the technical adequacy of CBM scores as indicators of (static) performance in foreign-language learning. Specifically, we examined the reliability and validity of CBM maze and Dutch-to-English and English-to-Dutch word-translation scores for 8th- and 9th-grade students. Stage 1 research was appropriate given the fact that there has been little previous CBM research in the area of foreign-language learning and with students in secondary school. The results of this initial study can be used to guide future Stage 2 and 3 studies.

In general, scores on 2-min maze and 2-min word translation (either Dutch-to-English or English-to-Dutch) were reliable and valid indicators of performance level in foreign-language learning for secondary-school students. Different administration times and scoring procedures were examined. The results of the reliability analyses revealed that scores on all three measures that were administered for two minutes were more reliable than for one minute administration times. With regard to the scoring procedure, the correct-minus-incorrect maze scores were slightly more reliable than the number of correct maze scores, whereas for both word-translation tasks the number of correct translations were more reliable than the correct-minus-incorrect translations. The use of a guessing rule for the maze or a spelling rule for word translation did not affect the reliability of the scores on the measures.

Based on the results of the reliability analysis, and on factors such as ease, efficiency of scoring, previous research, and so forth, a subset of scores were selected for subsequent validity analysis. In general, the pattern of results supported the validity of scores from 2-min maze and 2-min word-translation tasks (either English-to-Dutch or Dutch-to-English) as indicators of general foreign-language performance level. No clear pattern of differences emerged across scores from the three potential measures; however, validity results were generally better for students in their 3rd year of English than for students in their 2nd year of English. Finally, we also explored the effects of combining scores across measures for

students in their 3rd year of English (9th grade). A combination of scores from maze and the English-to-Dutch word translation provided the best prediction of English course grades, suggesting that, if the measures are to be used for screening purposes, the use of scores from two CBM measures might be better than scores from one measure alone. Including the scores from the Dutch-to-English word translation did not improve prediction.

The extent to which scores from these measures can be used to monitor student growth and inform teachers' decision-making needs to be investigated in future Stage 2 and 3 studies.

Development of CBM Measures in Reading

Two studies were conducted to examine CBM measures in reading for secondary-school students. The studies were Stage 1 and 2 CBM research (L. S. Fuchs, 2004) and focused on the technical adequacy of CBM maze scores as indicators of performance level and growth in reading. Specifically, the alternate-form reliability, sensitivity, validity, and stability of scores or change in scores from CBM maze were examined for 7th-grade students. Both linear and nonlinear (logistic) growth trajectories were examined.

In the first study that we conducted on the development of CBM measures in reading, scores on 2-min mazes were found to be reliable, sensitive, and valid indicators of performance level and growth in reading for secondary-school students. Results for alternate-form reliability analyses revealed that the reliability was moderate-good. Sensitivity analyses showed that maze scores were sensitive to change in performance, and students differed in growth trajectories. This was the case for both linear and nonlinear growth, but nonlinear growth fit the maze data better. The overall pattern of results showed that the CBM maze task was valid and supported the use of CBM maze scores as indicators of performance level and growth in reading. In general, students with higher reading proficiency, as assumed by their dyslexia status, school level, and scores and change in scores on a reading test, had higher scores and grew steeper on the maze than students with lower reading proficiency. The pattern of results were similar for linear and nonlinear growth.

The findings from the study provided tentative support for the technical adequacy of scores from CBM maze measures as indicators of reading performance level and growth for secondary-school students; however, the results were based on 23 weeks of progress monitoring. If educators are to use progress data to inform instruction, then they must be able to make decisions in a timely fashion, for example, after 8 or 10 weeks. Therefore, the next study investigated the stability of the maze growth rates or slopes under various conditions. This was the first study to examine the stability of maze growth, and the first to examine the stability produced from CBM reading measures at the secondary-school level.

Results of this subsequent study revealed that the stability of maze slopes was dependent on a number of factors, including the duration of the progress-monitoring

period, the data-collection schedule (i.e., how many data points and often data were collected), and the variation in individual maze scores (reflected in the standard error of estimate). Longer durations, denser schedules, and lower variation in maze scores increased the stability of the maze slopes. However, relatively large instability was seen in the first few weeks of data collection, regardless of the schedule or maze variation, but decreased with longer duration. The implications are that educators should monitor students' growth for at least 10 to 12 weeks – assuming that mazes were administered once per week under 'typical' conditions – before the maze slope would stabilize. Finally, the effects of assuming nonlinear growth on maze variation were examined, but did not result in reduced variation in scores.

The Use of CBM Scores and Future Directions

The results of our studies contributed to the development of CBM progress-monitoring measures at the secondary-school level, and were Stage 1 CBM research (L. S. Fuchs, 2004) in foreign-language learning, and Stage 1 and 2 CBM research (L. S. Fuchs, 2004) in reading. These stages focused on whether CBM scores were suitable for monitoring learning progress. In this section, the results are compared with previous research and the use of the CBM scores are discussed separately for foreign-language learning and reading.

Foreign-language learning

It is difficult to compare the results of our study on the development of CBM measures for foreign-language learning to previous studies because there have been few (if any) such studies. However, it is possible to compare our results to CBM secondary-school studies in related areas, for instance, studies on the development of measures for English-language learners (ELLs) in the United states (McMaster et al., 2006), studies that have examined maze scores as indicators of reading performance level (Espin et al., 2010; Tichá et al., 2009), or studies that have examined scores from a vocabulary-matching task as indicators of content-area learning (for example, Busch & Espin, 2003; Espin et al., 2001; Espin et al., 2013; Espin & Deno, 1994-95). Such studies provide a context within which to evaluate the results of our study.

The reliability results of the foreign-language measures could be considered good, given that the reliability is either similar or somewhat better than what has been found in previous studies (Espin et al., in press, or see also Chapter 2; McMaster et al., 2006). However, the validity results were somewhat weaker than those found in previous studies (Espin et al., in press, or see also Chapter 2). It is worthwhile to note that the 'weaker' relations were mostly seen for 8th-grade students, whereas the validity results for 9th-grade students were similar to those of previous studies (Espin et al., in press, or see also Chapter 2; McMaster et al., 2006). It is possible that students need to reach a basic level of general language-proficiency before the CBM scores can predict performance in foreign-language learning adequately, and that the 8th-grade students who were in their second year of English

instruction, had not yet reached this basic level. Perhaps 9th-grade students had reached this level because they had received an extra year of English instruction, and for instance the English language learners in the previous studies had also reached this basic level because they lived in a 'rich' environment where English was spoken regularly. Nevertheless, it is too soon to draw firm conclusions, especially given that the validity analyses in the foreign-language learning study were conducted within grade level and school- or reading test level, and sample sizes were therefore small.

An interesting result from the foreign-language study was that a combination of scores from two CBM measures, maze and English-to-Dutch word translation, predicted performance in foreign-language learning better than a single measure alone. Although this result should be interpreted carefully due to small sample sizes, this finding was similar to that of Espin and Foegen (1996), in which a combination of scores from maze and vocabulary-matching predicted the performance of content-area tasks better than scores from only one of the measures. It might be worthwhile to examine such predictions on performance in future research.

The results of the study on CBM foreign-language learning are promising, but it is important that such results be replicated in future research, if possible with a larger sample size and wider range of grade levels. Most importantly, additional research should be conducted to examine the CBM scores as indicators of growth in foreign-language learning at secondary-school level, and to examine the effects of teachers' use of such progress-monitoring measures on their instruction and on student achievement.

Reading

Although the technical adequacy of scores from CBM measures in reading at secondary-school level has been examined more extensively than for foreign-language learning, the research has still been sparse, especially compared to the amount of research in CBM reading that has been conducted at elementary-school level (see Wayman et al., 2007). Thus, in our study, we focused on the alternate-form reliability, sensitivity, validity, and stability of maze scores as indicators of reading performance level and growth at the secondary-school level.

The reliability results of the 2-min maze scores were slightly weaker than the 3-min maze scores used in studies that also included frequent (i.e., weekly) measurement of CBM measures (Espin et al., 2010; Tichá et al., 2009). Such results were acceptable given that the reliability in general was moderate-good. Nonetheless, a lower reliability may affect the variation in maze scores, and in turn, the stability of maze slopes of individual students. The sensitivity and validity results were similar to what was found in most previous studies (Espin et al., 2010; Ticha et al., 2009; Tolar et al., 2014; Tolar et al., 2012). The replication of these results was important and necessary because previous studies included only small sample sizes (Espin et al., 2010; Tichá et al., 2009), or included only five measurements that were

too few for making timely instructional decisions (Tolar et al., 2014; Tolar et al., 2012), whereas in our study, progress monitoring occurred frequently (i.e., weekly) for 23 weeks for a large sample of students. Finally, since our study was the first to examine the stability of scores from CBM reading measures at the secondary-school level, and the first to examine the stability of scores produced by the maze measure, we compared our results to studies at elementary-school level focused on CBM reading-aloud scores. The results on the stability of maze slopes were similar to the findings for the stability of reading-aloud slopes at the elementary-school level (Christ, 2006; Christ et al., 2013).

It is positive that our results confirmed earlier findings. However, although the results were consistent with previous studies, it is important to note that the maze growth rates differed greatly across studies. For instance, growth rates in the current study were similar to the growth rates found in Tolar et al. (2012), but more than five times smaller than growth rates found in other studies (Espin et al., 2010; Tichá et al., 2009). The various possible explanations for these differences were discussed in detail in Chapter 4; however, we emphasize here that the differences are of concern, especially in relation to the stability of the growth rates. If expected growth rates are smaller than the amount of instability in individual maze slopes, students' reading growth might not be adequately represented, and this might lead to incorrect interpretations and instructional decisions. A more detailed discussion of this issue and possibilities to reduce maze score variation in order to improve the stability of maze slopes are provided in Chapter 5. It is important that in future research normative growth rates are established by administering a validated maze passage set to a representative sample of students. In addition, future research should also focus on reducing maze score variation around the slope. Such research would provide educators with information regarding what they should expect from their students, and thus would help with setting long range goals in learning when using CBM.

Although there is much that remains to be examined in future research, we did find that, in general, scores or change in scores on CBM maze were reliable, sensitive, and valid indicators of reading performance level and growth at secondary-school level. The amount of stability given the expected growth rates is still a matter of concern, which needs to be addressed in future research.

One final related issue is the functional form of the expected growth of students. Although we found that nonlinear growth better represented the data, it appears that the difference is small given that the variation in maze slopes did not differ when either linear or nonlinear was assumed. It would be worthwhile to examine this issue in future research.

Conclusion and Final Remarks

This dissertation project contributed to the knowledge base of Stage 1 and 2 CBM research (L. S. Fuchs, 2004) at secondary-school level, namely to the suitability of using scores from CBM measures as indicators of performance levels and/or growth in foreign-language

learning and reading. Scores from CBM maze and word translation emerged as potential measures for CBM progress-monitoring in foreign-language learning. Moreover, the results in reading for CBM maze measures strengthens earlier research, and provides additional evidence for the use of maze scores in CBM progress-monitoring at secondary-school level.

Based on our results and given the intended use of the scores, we carefully recommend that 2-min CBM maze or CBM word-translation scores could be used as indicators of foreign-language learning performance levels, although more research is necessary, especially research focused on growth. For monitoring reading growth, we recommend the use of 2-min CBM maze scores for 7th-grade students, but, the measure or administration conditions must change slightly in order to improve the stability of the maze slopes. Moreover, we currently recommend that CBM reading growth can be presented as a linear rather than nonlinear growth trajectory, because from our research it could not be concluded that one approach was *much* better than the other, and linear growth trajectories provide higher practical advantages and possibly higher expectations from educators for their students' performance (see also Chapter 4).

As a last point, we want to emphasize that the 'real test' for CBM progress monitoring lies at Stage 3 research. It is one thing to develop a measurement system that has evidence of reliability and validity for monitoring the growth of students in an academic area, but quite another to determine whether implementation of such a system leads to improvements in the instruction offered to students who struggle, and ultimately to improvements in their achievement. The ultimate test of CBM is to examine whether teachers' use of CBM measures for monitoring the growth of secondary-school students who struggle in reading and/or foreign-language learning leads to improvement in teachers' instruction and to improvements in student performance. A related question is whether implementation of CBM progress monitoring leads to changes in teachers' beliefs about the extent to which they can influence the performance and growth of students who struggle.

In sum, the ultimate goal of CBM implementation is to provide teachers a method that allows them to 'scientifically' examine and evaluate the effectiveness of the instruction they deliver to students who struggle, and to create in teachers a sense of responsibility and a sense of urgency with regard to performance and growth for secondary-school students who struggle. Relatively speaking, educators have little time to influence the learning of students who struggle. Time cannot be wasted on the implementation of ineffective instruction for these students. In addition, educators must strive to develop in such students the maximal levels of skills possible in order to improve the likelihood of success for these students in their post-secondary school years. Perhaps CBM implementation can be a part of the development of effective and powerful instructional programs for students who struggle.

Nederlandse samenvatting (Dutch summary)

Introductie

Curriculum-Based Measurement (CBM; Continue voortgangsmonitoring, CVM) is een voortgangsmonitoringstool die gebruikt kan worden in het onderwijs om de leerprestaties en -groei van leerlingen te monitoren en om de effectiviteit van de gegeven instructies te evalueren (Deno, 1985). Om de leergroei te kunnen monitoren, dienen de leerprestaties van een leerling frequent (bijvoorbeeld wekelijks) gemeten te worden met een CBM-taak. De behaalde scores op een CBM-taak worden vervolgens geplaatst in een voortgangsgrafiek die weergeeft of een individuele leerling voortgang maakt. Op basis hiervan kan geïnterpreteerd worden in hoeverre de lesinstructie en/of interventie effectief is en of het eventueel nodig is om de instructie en/of interventie aan te passen.

CBM-taken dienen aan een aantal voorwaarden te voldoen (Deno, 1985). Ten eerste dienen de taken simpel, efficiënt en goedkoop te zijn en per type taak dienen meerdere parallelversies beschikbaar te zijn, zodat ze door leerkrachten en/of andere onderwijs-professionals frequent afgenomen kunnen worden. Ten tweede dienen de scores van de taken gemakkelijk te begrijpen zijn, zodat de resultaten duidelijk uitgelegd kunnen worden aan leerkrachten, ouders en/of leerlingen. Tot slot dienen de scores op de taken betrouwbare en valide *indicatoren* te zijn van leerprestaties en -groei in een vakgebied in het onderwijs, zodat de scores een representatie zijn van algemene leergroei en gekoppeld kunnen worden aan de instructies in dat vakgebied.

Ondanks dat er een breed scala aan CBM-onderzoek is uitgevoerd en gerapporteerd (Espin et al., in press, of zie ook Hoofdstuk 2; Foegen et al., 2007; McMaster & Espin, 2007; Wayman et al., 2007), is in de wetenschappelijke literatuur nog onvoldoende aandacht besteed aan twee belangrijke aspecten. Zo heeft slechts een beperkt aantal studies zich gericht op CBM-onderzoek in het voortgezet onderwijs. Daarnaast hebben relatief weinig studies zich gericht op de tweede en derde fase van CBM-onderzoek (L. S. Fuchs, 2004). Fase-2-onderzoek heeft betrekking op de technische adequaatheid van scores op CBM-taken als indicatoren van leergroei. Fase-3-onderzoek betreft het gebruik van CBM, oftewel de effecten van de implementatie van CBM op de instructie van de leerkracht en/of leerlingprestaties. Het overgrote deel van de onderzoeken is gericht op de eerste fase van CBM-onderzoek dat betrekking heeft op de technische adequaatheid van scores op een CBM-taak als indicator van het leerniveau (i.e., leerprestaties op één tijdsmoment). Het is mogelijk dat deze studies vaker worden uitgevoerd, omdat dergelijke onderzoeken relatief goedkoper en minder tijdsintensief zijn dan onderzoeken in de tweede en derde fasen van CBM-onderzoek. Daarnaast is fase-1-onderzoek ook een eerste stap in het uitvoeren van CBM-onderzoek, zodat op basis van deze onderzoeksresultaten geëvalueerd kan worden of het de moeite waard is om vervolgonderzoek uit te voeren in fasen 2 en 3.

De studies in de huidige dissertatie focusten zich op CBM-onderzoek in zowel fase 1 en 2 op het gebied van leesvaardigheid en het leren van een vreemde taal in het voortgezet onderwijs. In de onderstaande paragrafen worden de onderzoeksresultaten van de studies kort samengevat.

Ontwikkeling van CBM-taken in Engelse taalvaardigheid

In hoofdstuk 3 werd de technische adequaatheid van scores op potentiële CBM-taken als indicatoren van Engelse taalvaardigheid onderzocht. Er werd gekeken naar de parallelvorm betrouwbaarheid en validiteit van scores op CBM maze, Nederlands-naar-Engels en Engels-naar-Nederlands woordvertaaltaken voor leerlingen uit de tweede en derde klas van de middelbare school. Deze studie was het eerste gepubliceerde artikel gericht op het leren van een vreemde taal en was gericht op fase-1 CBM-onderzoek (L. S. Fuchs, 2004), zodat op basis van deze resultaten aanbevelingen gegeven konden worden voor vervolgonderzoek in fasen 2 en 3.

In dit onderzoek werd gevonden dat scores op 2-min maze and 2-min woordvertaaltaken (zowel Nederlands-naar-Engels als Engels-naar-Nederlands) betrouwbare en valide indicatoren zijn van prestaties in het leren van Engels als vreemde taal voor middelbare scholieren. Verschillende afnametijden en scoringsprocedures werden vergeleken voor iedere taak. Uit de betrouwbaarheidsanalyses kwam naar voren dat scores met een hogere afnameduur van 2 min betrouwbaarder waren dan scores met een lagere afnameduur van 1 min op de drie onderzochte CBM-taken. Met betrekking tot de scoringsprocedure werd gevonden dat de correct-minus-incorrecte maze scores iets betrouwbaarder waren dan het aantal correcte maze scores, terwijl voor beide woordvertaaltaken het aantal correcte vertalingen betrouwbaarder waren dan het aantal correct-minus-incorrecte vertalingen. Het gebruik van een gokregel voor de maze of een spellingsregel voor de vertaaltaken had geen invloed op de betrouwbaarheid van de meetinstrumenten.

Op basis van de resultaten van de betrouwbaarheidsanalyses en factoren als gemak, efficiëntie en eerder onderzoek, werd een selectie gemaakt van scores op de CBM-taken die nader onderzocht zouden worden in de validiteitsanalyses. Over het algemeen kwam uit deze validiteitsanalyses naar voren dat de validiteit van de scores op 2-min maze- en 2-min woordvertaaltaken (zowel Engels-naar-Nederlands als Nederlands-naar-Engels) ondersteund werd als indicatoren van algemene taalvaardigheden in het Engels als vreemde taal. Er waren geen duidelijke verschillen zichtbaar tussen de scores van de drie CBM-taken. Echter, de validiteitsresultaten waren over het algemeen beter voor leerlingen in de derde klas dan voor leerlingen in de tweede klas. Tot slot werden de effecten van het combineren van de scores van de CBM-taken onderzocht voor leerlingen uit de derde klas. Uit de resultaten kwam naar voren dat een combinatie van scores van zowel de maze als de Engels-naar-Nederlands woordvertalingstaak de beste voorspelling

gaf aan het eindcijfer voor het vak Engels. Dit suggereerde dat het gebruik van scores van twee CBM-taken de prestatie op Engelse taalvaardigheid beter voorspelde dan het gebruik van de scores van slechts één CBM-taak. Het toevoegen van de scores op de Nederlands-naar-Engels woordvertaling verbeterde de voorspelling niet.

De mate waarin de scores op de CBM-taken gebruikt kunnen worden om de leergroei te monitoren en/of hoe leerkrachten de data gebruiken om hun instructies aan te passen, dient onderzocht te worden in toekomstig fase 2 en 3 CBM-onderzoek.

Ontwikkeling van CBM-taken voor het Monitoren van Groei in Lezen

Twee studies werden uitgevoerd om CBM-taken in algemene leesvaardigheid te onderzoeken. Deze studies waren zowel fase 1 en 2 CBM-onderzoek (L. S. Fuchs, 2004) en focusten zich op de technische adequaatheid van maze scores als indicatoren van leesvaardigheidsniveau en –groei. De parallelvorm betrouwbaarheid, sensitiviteit, validiteit en stabiliteit van maze-scores werd onderzocht bij leerlingen uit de eerste klas van het voortgezet onderwijs. Zowel lineair als non-lineair (logistische) groei werden meegenomen in de analyses.

In de eerste studie werd gevonden dat scores van 2-min frequent afgenomen mazes betrouwbare, sensitieve en valide indicatoren waren van zowel leesvaardigheidsniveau als leesgroei. Uit de resultaten van de betrouwbaarheidsanalyses kwam naar voren dat de betrouwbaarheid gemiddeld-goed waren. De sensitiviteitsanalyses toonden aan dat maze scores sensitief waren om verandering in prestaties te detecteren en dat leerlingen individuele verschillen lieten zien in groeitrajecten. Dit was het geval voor zowel lineaire als non-lineaire groei, maar de maze data had een betere fit met het non-lineaire groeimodel. De patronen in de resultaten toonden aan dat scores van de CBM maze taak valide indicatoren waren van leesvaardigheidsniveau en -groei. Leerlingen met een hoge vaardigheid in lezen, zoals verwacht kan worden op basis van hun dyslexiestatus, schoolniveau, en scores en verandering in scores op een gestandaardiseerde leesvaardigheidstest, hadden hogere scores en groeiden steiler op de maze dan leerlingen met een lagere vaardigheid in lezen. De patronen waren vergelijkbaar voor lineaire en non-lineaire groei.

Uit de resultaten van de studie kan met voorzichtigheid geconcludeerd worden dat de technische adequaatheid van CBM maze scores ondersteund worden als indicatoren van leesvaardigheidsniveau en -groei voor leerlingen uit het voortgezet onderwijs. Echter waren de resultaten gebaseerd op de voortgang van 23 weken. Indien de maze-voortgangsdata door onderwijsprofessionals gebruikt zou worden om (les)instructies formatief te evalueren, dan zouden beslissingen tijdig gemaakt moeten kunnen worden, bijvoorbeeld na 8 of 10 weken. Daarom werd in een vervolgstudie onderzocht wat de stabiliteit van maze groeisnelheden of maze trendlijnen zou zijn onder

verschillende condities. Dit was de eerste studie die de stabiliteit van groei op de maze en die de stabiliteit van groei op CBM leestaken in het voortgezet onderwijs onderzocht.

In deze vervolgstudie werd gevonden dat de stabiliteit van maze trendlijnen afhankelijk waren van een aantal factoren, namelijk de duur van de voortgangsmonitoringsperiode, de dataverzamelingsplanning (i.e., het aantal datapunten per keer en hoe vaak de data per week was verzameld) en de variatie in individuele maze scores (gereflecteerd in de 'standard error of the estimate'). Een langere duur, een strakkere planning en een lagere variatie in maze scores verhoogde de stabiliteit van de maze trendlijnen. Echter, een relatief hoge mate van instabiliteit was zichtbaar tijdens de eerste dataverzamelingsweken, ongeacht de planning of variatie in maze scores, maar de instabiliteit verminderde naarmate er voor een langere duur data werd verzameld. De resultaten impliceerden dat onderwijsprofessionals de leesgroei van leerlingen voor minimaal 10 tot 12 weken dienen te volgen voordat de maze trendlijnen zouden stabiliseren, indien mazes één keer per week afgenomen zouden worden onder 'typische' omstandigheden. Tot slot werd onderzocht of de variatie in maze scores zou verminderen, indien een non-lineaire groei verondersteld zou worden. De variatie in maze scores veranderde niet bij het veronderstellen van non-lineaire groei.

Conclusie

De huidige dissertatie draagt bij aan de kennis van fase 1 en 2 CBM onderzoek (L. S. Fuchs, 2004) in het voortgezet onderwijs, namelijk in hoeverre CBM scores gebruikt kunnen worden als indicatoren van leerprestatieniveau en/of leergroei in het leren van een vreemde taal en leesvaardigheid. Zowel CBM maze als CBM woordvertaaltaken bleken veelbelovende taken om de voorgang in het leren van Engels als vreemde taal te monitoren. Daarnaast vullen de onderzoeksresultaten gericht op de CBM mazetaak eerder onderzoek naar CBM leesinstrumenten voor het voortgezet onderwijs aan en versterken de resultaten het bewijs dat de CBM mazetaak gebruikt kan worden om de leesvoortgang van leerlingen in het voortgezet onderwijs te monitoren.

Op basis van de onderzoeksresultaten en gegeven het doel van CBM, bevelen we met voorzichtigheid het gebruik van scores van een 2-min CBM maze of 2-min CBM woordvertaaltaak aan om de prestatie in het leren van Engels als vreemde taal in kaart te brengen, alhoewel meer onderzoek nodig is, met name gericht op de leergroei. Voor het monitoren van leesgroei bevelen we het gebruik van 2-min maze scores aan voor brugklasleerlingen, maar de taak of afnamecondities zouden licht aangepast moeten worden om de stabiliteit van de maze trendlijn te verbeteren. Daarnaast bevelen we op dit moment aan om de maze leesgroei lineair te benaderen, omdat uit het huidige onderzoek niet geconcludeerd kan worden dat de non-lineaire benadering veel beter is dan de lineaire benadering en omdat lineaire groei praktische voordelen biedt en mogelijk hogere

verwachtingen van leerkrachten over de leerprestaties van hun leerlingen kan opleveren (zie ook Hoofdstuk 4).

Tot slot willen we benadrukken dat de 'echte test' voor het ontwikkelen van een voortgangssysteem als CBM in de derde fase van CBM-onderzoek ligt. Enerzijds is het van belang om een meetsysteem te ontwikkelen dat ondersteund wordt met een sterke betrouwbaarheid en validiteit voor het monitoren van de leergroei van leerlingen. Anderzijds is het ook van belang of het gebruik van een dergelijk voortgangssysteem leidt tot verbeteringen in de instructie van leerlingen en uiteindelijk leidt tot verbeteringen in de leerprestaties van leerlingen. Het ultieme doel van de implementatie van CBM is om leerkrachten (en andere onderwijsprofessionals) een tool te bieden waarmee ze op een 'wetenschappelijke' wijze de effectiviteit van hun instructies aan leerlingen met leerproblemen kunnen evalueren en zo het maximale uit de leerlingen kunnen halen, waar de leerlingen tijdens maar ook na hun middelbare school nog lang profijt van kunnen hebben.

References

References

Albano, A. D., & Rodriguez, M. C. (2012). Statistical equating with measures of oral reading fluency. *Journal of School Psychology, 50*(1), 43-59. doi:10.1016/j.jsp.2011.07.002

Alderson, J. C., & Banerjee, J. (2002). State of the art review: Language testing and assessment (part two). *Language Teaching, 35*(2), 79-113.

Allinder, R. M. (1995). An examination of the relationship between teacher efficacy and curriculum-based measurement and student achievement. *Remedial and Special Education, 16*(4), 247-254. doi:10.1177/074193259501600408

Antonacci, P. A., O'Callaghan, C. M., & Berkowitz, E. (2015). *Developing content area literacy: 40 strategies for middle and secondary classrooms* (2nd ed). Thousand Oaks, CA: Sage.

Armbruster, B. B., & Anderson, T. H. (1988). On selecting "considerate" content area textbooks. *Remedial and Special Education, 9,* 47-52. doi:10.1177/074193258800900109

Baker, D. L, Biancarosa, G., Park, B. J., Bousselot, T., Smith, J., Baker, S. K., Kame'enui, E. J., Alonzo, J., & Tindal, G. (2015). Validity of CBM measures of oral reading fluency and reading comprehension on high-stakes reading assessments in grades 7 and 8. *Reading and Writing: An Interdisciplinary Journal, 28(1*), pp 57-104. doi:10.1007/s11145-014-9513-4

Barth, A. E., Stuebing, K. K., Fletcher, J. M., Cirino, P. T., Romain, M., Francis, D., & Vaughn, S. (2012). Reliability and validity of oral reading fluency median and mean scores among middle grade readers when using equated texts. *Reading Psychology, 33*, 133-161. doi:10.1080/02702711.2012.631863

Barth, A. E., Stuebing, K. K., Fletcher, J. M., Denton, C. A., Vaughn, S., & Francis, D. (2014). The effect of reading duration on the reliability and validity of middle school students' ORF performance. *Assessment for Effective Intervention, 40*, 53-64. doi:10.1177/1534508414545643

Bates, D., & Maechler, M. (2010). *lme4: Linear Mixed-Effects models using S4 Classes.* [Computer program and manual]. Retrieved from: http://r-forge.r-project.org/projects/lme4/

Betts, J., Pickart, M., & Heistad, D. (2009). An investigation of the psychometric evidence of CBM-R passage equivalence: Utility of readability statistics and equating for alternate forms. *Journal of School Psychology, 47*, 1–17. doi:10.1016/j.jsp.2008.09.001

Beyers, S. J., Lembke, E. S., & Curs, B. (2013). Social studies progress monitoring and intervention for middle school students. *Assessment for Effective Intervention, 38,* 224-235. doi:10.1177/1534508413489162

Borsuk, E. R. (2010). Examination of an administrator-read vocabulary-matching measure as an indicator of science achievement. *Assessment for Effective Intervention, 35,* 168-177. doi:10.1177/1534508410372081

Busch, T. W. & Espin, C. A. (2003). Using curriculum-based measurement to prevent failure and assess learning in the content areas. *Assessment for Effective Intervention, 28,* 49-58. doi:10.1177/073724770302800306

Catts, H. W., Adlof, S. M., & Weismer, S. E. (2006). Language deficits in poor comprehenders: A case for the simple view of reading. *Journal of Speech and Hearing Research, 49*, 278–293. doi:10.1044/1092-4388(2006/023)

CBS StatLine. (2015). VO leerlingen, onderwijssoort in detail, leerjaar [Secondary-school students, school level in detail, grade level] [data file]. Retrieved from http://www.cbs.nl/

CED-groep. (2012). Nieuwsbegrip [Understanding the news]. Rotterdam, The Netherlands: CED-groep.

Christ, T. J. (2006). Short-term estimates of growth using curriculum-based measurement of oral reading fluency: Estimating of standard error of the slope to construct confidence intervals. *School Psychology Review, 35*(1), 128-133.

Christ, T. J., & Ardoin, S. P. (2009). Curriculum-based measurement of oral reading: Passage equivalence and probe-set development. *Journal of School Psychology, 47*(1), 55-75. doi:10.1016/j.jsp.2008.09.004

Christ, T. J., Monaghen, B. D., Zopluoglu, C., & van Norman, E. R. (2012). Curriculum-based measurement of oral reading: Evaluation of growth estimates derived with pre-post assessment methods. *Assessment for Effective Intervention, 38*(3), 139-153.

Christ, T. J., Silberglitt, B., Yeo, S., & Cormier, D. (2010). Curriculum-based measurement of oral reading: An evaluation of growth rates and seasonal effects among students served in general and special education. *School Psychology Review, 39*(3), 447-462.

Christ, T. J., Zopluoglu, C., Long, J. D., & Monaghen, B. D. (2012). Curriculum-based measurement of oral reading: Quality of progress monitoring outcomes. *Exceptional Children, 78*(3), 356-373. doi:10.1177/001440291207800306

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. doi:10.1016/j.jsp.2012.11.001

Chung, S., Espin, C. A., & Stevenson, C. E. (2018). CBM maze-scores as indicators of reading level and growth for seventh-grade students. *Reading and Writing: An Interdisciplinary Journal, 3*, 627-648. doi:10.1007/s11145-017-9803-8

Cito. (2010). *Cito volgsysteem voortgezet onderwijs [Cito monitoring system for secondary education]*. Arnhem, The Netherlands: Centraal Instituut voor Toetsontwikkeling.

Codding, R. S., Petscher, Y., & Truckenmiller, A. (2015). CBM reading, mathematics, and written expression at the secondary level: Examining latent composite relations among indices and unique predictions with a state achievement test. *Journal of Educational Psychology, 107*(2), 437-450. doi:10.1037/a0037520

Conoyer, S. J., Lembke, E. S., Hosp, J., Espin, C. A., Hosp, M. &, Poch, A. (2017). Getting more from your maze: Examining differences in distractors. *Reading & Writing Quarterly: Overcoming Learning Difficulties, 33*(2), 141-154. doi:10.1080/10573569.2016.1142913

Cranney, A. G. (1972-73). The construction of two types of cloze reading tests for college students. *Journal of Reading Behavior, 5*(1), 60-64.

Cronbach, L. J., & Meehl, P. E. (1955). Construct validity in psychological tests. *Psychological bulletin, 52*(4), 281-302.

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

Decker, D. M., Hixson, M. D., Shaw, A., & Johnson, G. (2014). Classification accuracy of oral reading fluency and maze in predicting performance on large-scale reading assessments. *Psychology in the Schools, 51,* 625-635. doi:10.1002/pits.21773

Deno, S. L. (1985). Curriculum-based measurement: The emerging alternative. *Exceptional Children, 52*, 219-232. doi:10.1177/001440298505200303

Deno, S. L. (1990). Individual differences and individual difference: The essential difference of special education. *The Journal of Special Education, 24*(2), 160-173. doi:10.1177/002246699002400205

Deno, S. L. (2003). Developments in curriculum-based measurement. *The Journal of Special Education*, *37*(3), 184-192. doi:10.1177/00224669030370030801

Deno, S. L. (2013). Problem-solving assessment. In R. Brown-Chidsey and K. J. Andren (Eds.), *Assessment for Intervention: A problem-solving approach* (pp. 10-38). New York, NY: Guilford Press.

Deno, S. L. & Fuchs, L. S. (1987). Developing curriculum-based measurement systems for data-based special education problem solving. *Focus on Exceptional Children, 19*(8), 1-16.

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.

Deno, S. L., & Mirkin, P. K. (1977). *Data-based Program Modification: A Manual.* Reston, VA: Council for Exceptional Children.

Denton, C. A., Barth, A. E., Fletcher, J. M., Wexler, J., Vaughn, S., Cirino, P. T., Romain, M., & Francis, D. J. (2011). The relations among oral and silent reading fluency and

comprehension in middle school: Implications for identification and instruction of students with reading difficulties. *Scientific Studies of Reading, 15,* 109-135. doi:10.1080/10888431003623546

Dutch Inspectorate of Education. (2015). *Opbrengstgericht werken aan taal en rekenen in het voortgezet onderwijs [data-based decisions in language arts and mathematics at the secondary-school level]*. Utrecht: Dutch Inspectorate of Education. Retrieved from: https://www.rijksoverheid.nl/binaries/rijksoverheid/documenten/rapporten/2015/02/01/opbrengstgericht-werken-aan-taal-en-rekenen-in-het-voortgezet-onderwijs/opbrengstgericht-werken-aan-taal-en-rekenen-in-het-voortgezet-onderwijs.pdf

Dutch Ministry of Education, Culture and Science. (2005). *The Education System in the Netherlands 2005*. The Hague, The Netherlands: Dutch Eurydice Unit. Retrieved from: https://www.rijksoverheid.nl/binaries/rijksoverheid/documenten/rapporten/2005/12/23/education-system-in-the-netherlands/eurydice-en.pdf

Dutch Ministry of Education, Culture and Science. (2009). *Referentiekader taal en rekenen. [Reference framework language and mathematics]*. Enschede, The Netherlands: Doorlopende leerlijnen Taal en Rekenen. Retrieved from: http://www.taalenrekenen.nl/downloads/referentiekader-taal-en-rekenen-referentieniveaus.pdf/

Dutch Ministry of Education, Culture and Science. (2011). *Actieplan VO beter presteren: Opbrengstgericht en ambitieus. [Action plan to improve performance at secondary-school level: data-based decision-making and ambitious goals]*. The Hague: Dutch Ministry of Education, Culture and Science. Retrieved from: https://www.rijksoverheid.nl/binaries/rijksoverheid/documenten/kamerstukken/2011/05/23/actieplan-vo-beter-presteren/actieplan-vo-beter-presteren.pdf

Edmonds, M. S., Vaughn, S., Wexler, J., Reutebuch, C., Cable, A., Tackett, K. K., & Schnakenberg, J. W. (2009). A synthesis of reading interventions and effects on reading comprehension outcomes for older struggling readers. *Review of Educational Research*, *79*(1), 262-300. doi:10.3102/0034654308325998

Egberink, I. J. L., Janssen, N. A. M., & Vermeulen, C. S. M. (2015a). COTAN beoordeling 2015, Cito LOVS toets 0 [COTAN review 2015, Cito student educational progress monitoring system test 0]. Retrieved from www.cotandocumentatie.nl

Egberink, I. J. L., Janssen, N. A. M., & Vermeulen, C. S. M. (2015b). COTAN beoordeling 2015, Cito LOVS toets 1 [COTAN review 2015, Cito student educational progress monitoring system test 1]. Retrieved from www.cotandocumentatie.nl

Espin, C. A., Busch, T. W., Lembke, E. S., Hampton, D. D., Seo, K., & Zukowski, B. (2013). Curriculum-based measurement in science learning: Vocabulary-matching as an indicator of performance and progress. *Assessment for Effective Intervention, 38,* 203-213. doi:10.1177/1534508413489724

Espin, C. A., Busch, T. W., Shin, J., & Kruschwitz, R. (2001). Curriculum-based measurement in the content areas: Validity of vocabulary-matching as an indicator of performance in

social studies. *Learning Disabilities Research and Practice, 16,* 142-151. doi:10.1111/0938-8982.00015

Espin, C. A., & Campbell, H. (2012). They're getting older . . . but are they getting better? The influence of CBM on programming for secondary-school students with learning disabilities. In C. Espin, K. McMaster, S. Rose, & M. Wayman (Eds.) *A Measure of Success: The Influence of Curriculum-Based Measurement on Education* (pp. 149-164). Minneapolis, MN: University of Minnesota Press.

Espin, C. A., Chung, S., Foegen, A. & Campbell, H. (in press). Curriculum-based measurement for secondary-school students. In P. C. Pullen, & M. J. Kennedy (Eds.), *Handbook of Response to Intervention and Multi-Tiered Systems of Support*. London: Taylor & Francis/Routledge.

Espin, C. A., & Deno, S. L. (1993a). Content-specific and general reading disabilities of secondary-level students: Identification and educational relevance. *Journal of Special Education, 27*, 321-337.

Espin, C. A., & Deno, S. L. (1993b). Performance in reading from content-area text as an indicator of achievement. *Remedial and Special Education,14*(6), 47-59. doi:10.1177/074193259301400610

Espin C. A., Deno S. L. (1994–95). Curriculum-based measures for secondary students: Utility and task specificity of text-based reading and vocabulary measures for predicting performance on content-area tasks. *Diagnostique, 20*, 121–142

Espin, C. A., & Deno, S. L. (2016). Oral reading fluency or reading aloud from text: An analysis through a unified view of construct validity. In K. D. Cummings & Y. Petscher (Eds.), *The Fluency Construct: Curriculum-Based Measurement Concepts and Applications* (pp. 365-384). New York, NY: Springer.

Espin, C. A., Deno, S. L., Maruyama, G., & Cohen, C. (1989). *The Basic Academic Skills Samples (BASS): An instrument for the screening and identification of children at risk for failure in regular education classrooms*. Paper presented at the National Convention of the American Educational Research Association, March, 1989.

Espin, C. A., & Foegen, A. (1996). Validity of three general outcome measures for predicting secondary students' performance on content-area tasks. *Exceptional Children, 62*, 497-514.

Espin, C. A., Shin, J., & Busch, T. W. (2005). Curriculum-Based Measurement in the content areas: Vocabulary matching as an indicator of progress in social studies learning. *Journal of Learning Disabilities, 38,* 353-363. doi: 10.1177/00222194050380041301

Espin, C. A., & Tindal, G. (1998). Curriculum-based measurement for secondary students. In M. R. Shinn (Ed.), *Advanced applications of curriculum-based measurement* (pp. 214-253). New York, NY: Guilford Press.

Espin, C. A., 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 and Practice*, *25*(2), 60-75. doi:10.1111/j.1540-5826.2010.00304.x

Eurydice (2001). *Foreign language teaching in schools in Europe*. Brussels: European Commission.

Fewster, S., & MacMillan, P. (2002). School-based evidence for the validity of Curriculum-Based Measurement of reading and writing. *Remedial and Special Education, 23,* 149-156. doi:10.1177/07419325020230030301

Foegen, A., Jiban, C., & Deno, S. L. (2007). Progress monitoring measures in mathematics: A review of the literature. *Journal of Special Education, 41*(2), 1-139. doi:10.1177%2F00224669070410020101

Fuchs, D., Fuchs, L. S., Mathes, P. G., & Lipsey, M. W. (2000). Reading differences between low-achieving students with and without learning disabilities: A meta-analysis. In R. Gersten, E. Schiller, & S. Vaughn (Eds.), *Research synthesis in special education* (pp. 81–104). Mahwah, NJ: Erlbaum.

Fuchs, L. S. (1989). Effects of systematic observation and feedback on teachers' implementation of curriculum-based measurement. *Teacher Education and Special Education, 16*(2), 178-187. doi:10.1177/088840649301600210

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

Fuchs, L. S., & Deno, S. L. (1994). Must instructionally useful performance assessment be based in the curriculum? *Exceptional Children, 61*(1), 15-24. doi: 10.1177/001440299406100103

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

Fuchs, L. S., Fuchs, D., & Hamlett, C. L. (2007). Using curriculum-based measurement to inform reading instruction. *Reading and Writing: An Interdisciplinary Journal*, *20*(6), 553-567. doi:10.1007/s11145-007-9051-4

Good, R. H., & Shinn, M. R. (1990). Forecasting accuracy of slope estimates for reading curriculum-based measurement: Empirical evidence. *Behavioral Assessment, 12*, 179-193.

Groves, F. H. (1995). Science vocabulary load of selected secondary science textbooks. *School Sciences & Mathematics, 95,* 231-235. doi:10.1111/j.1949-8594.1995.tb15772.x

Haladyna, T. M., & Downing, S. M. (2004). Construct-irrelevant variance: A threat in highstakes testing. *Educational Measurement: Issues and Practice, 23*(1), 17-27.

Hale, A. D., Henning, J. B., Hawkins, R. O., Sheeley, W., Shoemaker, L., Reynolds, J. R., & Moch, C. (2011). Reading assessment methods for middle-school students: An investigation of reading comprehension rate and maze accurate response rate. *Psychology in the Schools, 48,* 28-36. doi:10.1002/pits.20544

Hale, G. A., Stansfield, C. W., Rock, D. A., Hicks, M. M., Butler, F. A., & Oller, J. W. (1989). The relation of multiple-choice cloze items to the test of English as a foreign language. *Language Testing, 6*(1), 47-76.

Hallahan, D. P., Lloyd, J. W., Kauffman, J. M., Weiss, M. P., & Martinez, E. A. (2005). *Learning disabilities: foundations, characteristics, and effective teaching* (3rd edition). Boston: Pearson/Allyn and Bacon.

Hintze, J. M., & Christ, T. J. (2004). An examination of variability as a function of passage variance in CBM progress monitoring. *School Psychology Review, 33*(2), 204-217.

Hox. J. (2010). *Multilevel analysis. Techniques and applications* (2nd edition). New York, NY: Routledge. (see http://joophox.net/mlbook2/MLbook.htm)

Jenkins, J. R. & Fuchs, L. S. (2012). Curriculum-Based Measurement: The paradigm, history, and legacy. In C. Espin, K. McMaster, S. Rose, & M. Wayman (Eds.) *A Measure of success: The influence of Curriculum-based Measurement on education* (pp. 7-26). Minneapolis, MN: University of Minnesota Press.

Jenkins, J. R., Graff, J. J., & Miglioretti, D. L. (2009). Estimating reading growth using intermittent CBM progress monitoring. *Exceptional Children, 75,* 151-163. doi:10.1177/00144

Jenkins, J. R., & Pany, D. (1978). Curriculum biases in reading achievement tests. *Journal of Reading Behavior, 10*, 345–357. doi:10.1080/10862967809547288 0290907500202

Johnson, E. S., Semmelroth, C., Allison, J., & Fritsch, T. (2013). The technical properties of science content maze passages for middle school students. *Assessment for Effective Intervention, 38,* 214-233. doi:10.1177/ 1534508413489337

Keller-Margulis, M. A., Mercer, S. H., Payan, A., & McGee, W. (2015). Measuring annual growth using written expression curriculum-based measurement: An examination of seasonal and gender differences. *School Psychology Quarterly, 30*(2), 276-288. doi:10.1037/spq0000086

Ketterlin-Geller, L. R., McCoy, J. D., Twyman, T., & Tindal, G. (2006). Using a concept maze to assess student understanding of secondary-level content. *Assessment for Effective Intervention, 31,* 39-50. doi:10.1177/ 073724770603100204

Kieffer, M. J. (2011). Converging trajectories: Reading growth in language minority learners and their classmates, kindergarten to grade 8. *American Educational Research Journal, 48*, 1187–1225. doi:10.3102/0002831211419490

Kim, Y., Petscher, Y., & Foorman, B. (2015). The unique relation of silent reading fluency to end-of-year reading comprehension: Understanding individual differences at the student, classroom, school, and district levels. *Reading and Writing: An Interdisciplinary Journal, 28,* 131-150. doi:10.1007/s11145-013-9455-2

Mandinach, E. B. (2012). A perfect time for data use: Using data-driven decision making to inform practice. *Educational Psychologist, 47*, 71–85.

Mazesonline. (2013). Available from http://www.mazesonline.nl/

McArdle, J. J., Grimm, K. J., Hamagami, F., Bowles, R. P., & Meredith, W. (2009). Modeling life-span growth curves of cognition using longitudinal data with multiple samples and changing scales of measurement. *Psychological Methods, 14*(2), 126-149. doi:10.1037/a00015857

McCane-Bowling, S. J., Strait, A. D., Guess, P. E., Wiedo, J. R., & Muncie, E. (2014). The utility of maze accurate response rate in assessing reading comprehension in upper elementary and middle school students. *Psychology in the Schools, 51,* 789-800. doi:10.1002/pits.21789

McMaster, K., L. & Espin, C. (2007). Technical features of curriculum-based measurement in writing: A literature review. *Journal of Special Education, 41*(2), 68-84. doi:10.1177/00224669070410020301

McMaster, K. L., Wayman, M. M., & Cao, M. (2006). Monitoring the reading progress of secondary-level English learners: Technical features of oral reading and maze tasks. *Assessment for Effective Intervention, 31*(4), 17-31. doi:10.1177/073724770603100402

Messick, S. (1989). Validity. In R. L. Linn (Ed), *Educational measurement* (3rd edition), (pp. 13-103). New York, NY: MacMillan.

Mooney, P., McCarter, K. S., Russo, R. J., & Blackwood, D. L. (2013). Examining an online content general outcome measure: Technical features of the static score. *Assessment for Effective Intervention, 38,* 249-260. doi:10.1177/ 1534508413488794

Mooney, P., McCarter, K. S., Schraven, J., & Callicoatte, S. (2013). Additional performance and progress validity findings targeting the content-focused vocabulary matching. *Exceptional Children, 80,* 85-100. doi:10.1177/ 001440291308000104

Mooney, P., McCarter, K. S., Schraven, J., & Haydel, B. (2010). The relationship between content area general outcome measurement and statewide testing in sixth-grade world history. *Assessment for Effective Intervention, 35*(3), 148-158. doi:10.1177/1534508409346052

Muijselaar, M. M. L., Kendeou, P., de Jong, P. F., & van den Broek, P. W. (2017). What does the CBM-maze test measure? *Scientific Studies of Reading, 21*(2), 120-132. doi:10.1080/10888438.2016.1263994

Nation, I. S. P. (1982). Beginning to learn foreign vocabulary: A review of the research. *RELC Journal, 13*(1), 14-36.

Nederlands Instituut van Psychologen (2011). *Documentatie van tests en testresearch in Nederland: Researchbeschrijving Volg- en Adviessysteem (VAS) [Documentation of assessments and research on assessments in the Netherlands: Description of research on Progress Monitoring and Advice system].* Amsterdam: Boom test uitgevers.

Nese, J. F. T., Biancarosa, G., Anderson, D., Lai, C., Alonzo, J., & Tindal, G. (2012). Within-year oral reading fluency with CBM: A comparison of models. *Reading and Writing: An Interdisciplinary Journal, 25*(4), 887-915. doi:10.1007/s11145-011-9304-0

Nese, J. F. T., Biancarosa, G., Cummings, K., Kennedy, P., Alonzo, J., & Tindal, G. (2013). In search of average growth: Describing within-year oral reading fluency growth across Grades 1-8. *Journal of School Psychology, 51,* 625-642. doi:10.1016/j.jsp.2013.05.006

NWEA. (2003). *Technical manual for the Northwest Evaluation Association (NWEA) Measures of Academic Progress and Achievement Level Tests.* Portland, OR: Northwest Evaluation Association.

OECD. (2016). PISA 2015 results (Volume I): Excellence and equity in education, PISA, OECD Publishing, Paris. doi:10.1787/9789264266490-en

Petscher, Y., Cummings, K. D., Biancarosa, G., & Fien, H. (2013). Advanced (measurement) applications of curriculum-based measurement in reading. *Assessment for Effective Intervention, 38*(2), 71-75. doi:10.1177/1534508412461434

Sandberg, K. L., & Reschly, A. L. (2011). English learners: Challenges in assessment and the promise of curriculum-based measurement. *Remedial and Special Education, 32*(2), 144-154.

Santi, K. L., Barr, C., Khalaf, S., & Francis, D. J. (2016). Different approaches to equating oral reading fluency passages. In: K. D. Cummings, & Y. Petscher (Eds), *The fluency construct: Curriculum-based measurement concepts and applications* (pp. 223-265). New York: Springer.

Savage, R. (2006). Reading comprehension is not always the produce of nonsense word decoding and linguistic comprehension: Evidence from teenagers who are extremely poor readers. *Scientific Studies of Reading, 10*, 143–164. doi:10.1207/s1532799xssr1002_2

Scammacca, N. K., Roberts, G. J., Cho, E., Williams, K. J., Roberts, G., Vaughn, S. R., & Caroll, M. (2016). A century of progress: Reading interventions for students in grades 4-12, 1914-2014. *Review of Educational Research, 86*(3), 756-800. doi: 10.3102/0034654316652942

Seifert, K., & Espin, C. A. (2012). Improving reading of science text for secondary students with learning disabilities: Effects of text reading, vocabulary learning, and combined approaches to instruction. *Learning Disability Quarterly, 35*, 236-247. doi:10.1177/0731948712444275

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. doi:10.1007/BF03024951

Shinn, M. R., Good, R. H., & Stein, S. (1989). Summarizing trend in student-achievement: A comparison of methods. *School Psychology Review, 18*(3), 356-370.

Silberglitt, B., Burns, M. K., Madyun, N., & Lail, K. E. (2006). Relationship of reading fluency assessment data with state accountability test scores: A longitudinal comparison of grade levels. *Psychology in the Schools, 43,* 527-535. doi:10.1002/pits.20175

Skiba, R. J., Deno, S. L., Marston, D., & Wesson, C. (1986). Characteristics of time-series data collected through curriculum-based reading measurement. *Diagnostique, 12*, 3-15. doi:10.1177/073724778601200101

Skinner, M. E., & Smith, A. T. (2011). Creating success for students with learning disabilities in postsecondary foreign language courses. *International Journal of Special Education, 26*(2), 42-57.

Sparks, R. L. (2006). Is there a disability for learning a foreign language? *Journal of Learning Disabilities, 39*(6), 544-557.

Sparks, R. L., Javorksy, J. & Philips, L. (2005). Comparison of the performance of college students classified as ADHD, LD, and LD/ADHD in foreign language courses. *Language Learning, 55*(1), 151–177.

Solis, M., Miciak, J., Vaughn, S., & Fletcher, J. M. (2014). Why intensive interventions matter: Longitudinal studies of adolescents with reading disabilities and poor reading comprehension. *Learning Disability Quarterly, 37*(4), 218-229. doi:10.1177/0731948714528806

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,* 795-819. doi:10.1002/pits.20113

Stevenson, N. A. (2015). Predicting proficiency on statewide assessments: A comparison of Curriculum-Based Measures in middle school. *The Journal of Educational Research, 108*(6), 492-503. doi:10.1080/0022067.2014.910161

Stichting Dyslexie Nederland (2008). *Dyslexie: Diagnose en behandeling van dyslexie [Dyslexia: Diagnoses and remediation of dyslexia].* Retrieved from http://www.stichtingdyslexienederland.nl/media/183/sdnbrochure2008.pdf.

Taylor, W. L. (1956). Recent developments in the use of 'cloze procedure'. *Journalism Quarterly, 33*, 42-48.

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. doi:10.1111/j.1540-5826.2009.00287.x

Tolar, T. D., Barth, A. E., Fletcher, J. M., Francis, D. J., & Vaughn, S. (2014). Predicting reading outcomes with progress monitoring slopes among middle grade students. *Learning and Individual Differences, 30*, 46-57. doi:10.1016/j.lindif.2013.11.001

Tolar, T. D., Barth, A. E., Francis, D. J., Fletcher, J. M., Stuebing, K. K., & Vaughn, S. (2012). Psychometric properties of maze tasks in middle school students. *Assessment for Effective Intervention, 37*(3), 131-146. doi:10.1177/1534508411413913

U. S. Department of Education. (2016a). *The nation's report card: 2011 Reading assessments: Classroom context: Time spent on language arts*. Retrieved from http://www.nationsreportcard.gov/reading_2011/context_1.aspx

U. S. Department of Education. (2016b). *The nation's report card: 2015 Reading Assessment, Overall achievement levels 8th grade*. Retrieved from https://www.nationsreportcard.gov/reading_math_2015/#reading/acl?grade=8

van den Berg, R., & te Lintelo, H. (1977). Analyse van Individualiseringsvormen (AVI)[Assessment of Dutch reading grade level]. Den Bosch: KPC-Groep.

van den Bosch, R. M., Espin, C. A., Chung, S., & Saab, N. (2017). Data-based decision-making: Teachers' comprehension of curriculum-based measurement progress-monitoring graphs. *Learning Disabilities Research & Practice, 32*(1), 46-60. doi:10.1111/ldrp.12122

van Norman, E. R., & Parker, D. C. (2016). An evaluation of the linearity of curriculum-based measurement of oral reading (CBM-R) progress-monitoring data: Idiographic considerations. *Learning Disabilities Research & Practice, 31*, 199-207. doi:10.1111/ldrp.12108

Vannest, K. J., Parker, R., & Dyer, N. (2011). Progress monitoring in grade 5 science for low achievers. *Journal of Special Education, 44*(4), 221-233. doi:10.1177/0022466909343121

Vaughn, S., Fletcher, J. M., Francis, D. J., Denton, C. A., Wanzek, J., Wexler, J., …Romain, M. A. (2008). Response to intervention with older students with reading difficulties. *Learning and Individual Differences, 18*, 338-345. doi:10.1016/j.lindif.2008.05.001

Vellutino, F. R., Tunmer, W. E., Jaccard, J. J., & Chen, R. (2007). Components of reading ability: Multivariate evidence for a convergent skills model of reading development. *Scientific Studies of Reading, 11*, 3–32. doi:10.1207/s1532799xssr1101_2

Verspoor, M., de Bot, K., & van Rein, E. (2010). Binnen- en buitenschools taalcontact en het leren van Engels. *Levende Talen Tijdschrift, 11*(4), 14-33.

Wang, M. (2011). Learning a second language. In: R. E. Mayer, & P. A. Alexander (Eds.), *Handbook of Research in Learning and Instruction* (pp.127-147). New York: Routledge.

Wallace, C. (2008). Vocabulary: The key to teaching English language learners to read. *Education Digest, 73*(9), 36-39.

Wayman, M. M., Wallace, T., Wiley, H. I., Tichá, R., & Espin, C. A. (2007). Literature synthesis on curriculum-based measurement in reading. *Journal of Special Education, 41*(2), 85-120. doi:10.1177/00224669070410020401

Wexler, J., Vaughn. S., Edmonds, M., & Reutebuch, C. K. (2008). A synthesis of fluency interventions for secondary struggling readers. *Reading and Writing: An Interdisciplinary Journal*, *21*, 317-347. doi:10.1007/s11145-007-9085-7

Wickham, H. (2011). The split-apply-combine strategy for data analysis. *Journal of Statistical Software, 40*(1), 1-29. doi:10.18637/jss.v040.i01

Yager, R. E. (1983). The importance of terminology in teaching K-12 science. *Journal of Research in Science Teaching, 20,* 577-588. doi:10.1002/tea.3660200610

Yeo, S., Fearrington, J. Y., & Christ, T. J. (2012). Relation between CBM-R and CBM-mR slopes: An application of latent growth modeling. *Assessment for Effective Intervention, 37*(3), 147-158. doi:10.1177/1534508411420129

Yovanoff, P., Duesbery, L., Alonzo, J., & Tindal, G. (2005). Grade-level invariance of a theoretical causal structure predicting reading comprehension with vocabulary and oral reading fluency. *Educational Measurement: Issues and Practice, 24,* 4-12. doi:10.1111/j.1745-3992.2005.00014.x

Dankwoord (Acknowledgements in Dutch)

Dankwoord (Acknowledgements in Dutch)

Dit proefschrift zou niet tot stand zijn gekomen zonder de begeleiding, ondersteuning, samenwerking en medewerking van verschillende personen. Ik zou dan ook graag van deze gelegenheid gebruik willen maken om de personen te bedanken die een bijdrage hebben geleverd aan mijn promotieonderzoek.

Allereerst wil ik graag alle leerlingen, leerkrachten en contactpersonen van de scholen bedanken voor de inzet die ze leverden tijdens hun deelname aan het onderzoek of het in goede banen leiden van het onderzoek binnen hun school. Zonder jullie hulp hadden we de onderzoeken niet kunnen uitvoeren. Daarnaast waren ook alle studenten en Robin onmisbaar. Ik wil jullie bedanken voor het helpen ontwikkelen van het onderzoeksmateriaal, het bezoeken van de scholen en/of het verwerken van de gegevens.

Mijn dank gaat uit naar mijn begeleiders. Chris, dank voor het vertrouwen dat je in mij had (en nog steeds hebt). Je was niet alleen betekenisvol voor mijn promotietraject, maar ook op persoonlijk gebied. Vooral jouw (of Deno's) wijze levenslessen heb ik altijd zeer gewaardeerd. Heel veel dank hiervoor! Claire, ik heb altijd geroepen dat ik het jammer vond dat je niet vanaf het begin betrokken was bij mijn begeleiding. Je efficiënte aanpak, duidelijke uitleg, altijd snelle reactie, prettige manier van feedback geven, maar vooral ook je positieve aanmoedigingen waren zeer waardevol voor mij, bedankt!

Uiteraard wil ik ook mijn lieve (oud-) collega's van Pedagogische Wetenschappen in Leiden bedanken. Ik vind het erg fijn dat ik jullie heb mogen leren kennen, omdat jullie veel voor mij hebben betekend (en nog steeds betekenen). Jullie waren er altijd om even over het onderzoek te kunnen sparren, om een gezellige tijd mee te hebben, maar voornamelijk ook om lief en leed mee te delen. Dank voor het luisterend oor, de schouder om op te huilen, de lieve attenties, de gezellige lunches/etentjes, de dropjes en de gezellige gesprekjes op weg naar de wc/het kopieerapparaat.

Roxette en Merel, naast bovenstaande wil ik jullie ook nog extra bedanken voor het feit dat jullie mij bij willen staan als paranimfen tijdens de promotie, zoals jullie er ook voor mij waren als kamergenootjes (en in de periode erna).

Tot slot zou ik mijn familie en vrienden willen bedanken. Dank voor jullie vertrouwen in mij en de aanmoedigingen. In het bijzonder wil ik Hidde bedanken. Het promotietraject was een uitdagend proces en jij was er altijd om mij te stimuleren het beste uit mijzelf te halen. Dank je wel!

About the author

Curriculum Vitae

Siuman Chung was born on December 13th 1984 in Haarlem, the Netherlands. After completing her secondary education in 2003 at Kaj Munk College in Hoofddorp, she studied Educational Sciences (Pedagogische Wetenschappen) at VU University in Amsterdam. During her master, she completed a clinical internship as a school psychologist at an elementary school, where she worked with students with special educational needs. In 2009, she obtained her master's degree, and thereafter, in 2010, she started her PhD research at the department of Learning Problems and Impairments of Leiden University under supervision of prof. dr. Christine Espin and dr. Claire Stevenson. Her PhD project focused on Curriculum-based measurement in reading and foreign-language learning for secondary-school students. To broaden her knowledge in the area of Curriculum-based measurement research, she joined the research teams of the department of School Psychology and Special Education at the University of Minnesota for two months in 2014, and worked on one of her research papers with dr. Theodore Christ from the University of Minnesota.

In addition to her PhD project, Siuman was also involved in teaching and ancillary activities. At the department of Education and Child Studies at Leiden University, she supervised theses and internships of bachelor- and master students. Moreover, as a member of the NVO Students and Starters committee she organized workshops and meetings for students and recent graduates. Finally, as a member of the ISED PhD committee she organized conferences for PhD students.

Currently, Siuman is an instructor at the institute of Education and Child Studies at Leiden University. She is involved in the teacher preparation training (Academische Pabo), and supervises theses and internships of bachelor- and master students.

List of Publications

Chung, S., Espin, C. A., & Stevenson, C. E. (2018). Stability of maze reading slopes: Implications for use of CBM maze scores for tracking the progress of secondary-school students. *Manuscript in preparation*.

van den Bosch, R. M., Espin, C. A., Sikkema-de Jong, M. T., **Chung, S.**, Boender, P., & Saab, N. (2018). Teachers' inspection patterns of curriculum-based measurement progress-monitoring graphs: An eye-tracking study. *Manuscript submitted for publication*.

Espin, C. A., **Chung, S.**, Foegen, A. & Campbell, H. (in press). Curriculum-based measurement for secondary-school students. In P. C. Pullen, & M. J. Kennedy (Eds.), *Handbook of Response to Intervention and Multi-Tiered Systems of Support*. London: Taylor & Francis/Routledge.

Chung, S., Espin, C. A., & Stevenson, C. E. (2018). CBM maze-scores as indicators of reading level and growth for seventh-grade students. *Reading and Writing: An Interdisciplinary Journal, 3*, 627-648. doi:10.1007/s11145-017-9803-8

van den Bosch, R. M., Espin, C. A., **Chung, S.**, & Saab, N. (2017). Data-based decision-making: Teachers' comprehension of curriculum-based measurement progress-monitoring graphs. *Learning Disabilities Research & Practice, 32*(1), 46-60. doi: 10.1111/ldrp.12122

Chung, S., & Espin, C. A. (2013). CBM progress monitoring in foreign-language learning for secondary school students: Technical adequacy of different measures and scoring procedures. *Assessment for Effective Intervention, 38,* 236-248. doi:10.1177/1534508413489723

