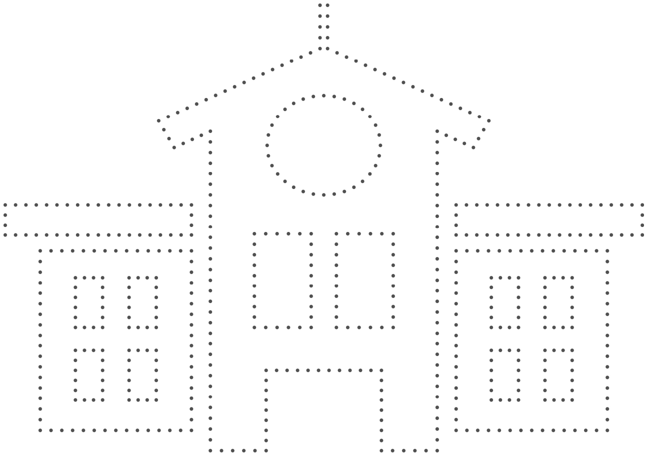
CBM progress monitoring in reading and foreign-language learning for secondary-school students

Chung, S.

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.