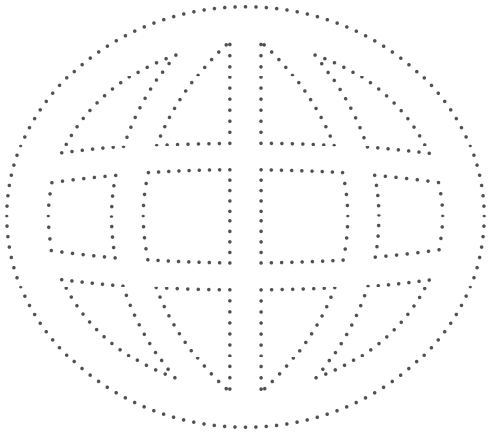
Chapter 3

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.