CBM Language Measures as Indicators of Foreign-Language Learning: Technical Adequacy of Scores for Secondary-School Students

Laura Hoefnagel, Christine A. Espin, and Ralph Rippe
Leiden University

Abstract

Students with and without learning disabilities often struggle to learn a foreign language (FL). Teachers could benefit from a measure designed to screen and identify students at risk for FL learning difficulties. In this study, we examined the reliability and validity of scores from four curriculum-based measures (CBM) as potential indicators of English FL learning: reading aloud, maze selection, and English-to-Dutch and Dutch-to-English word translation. Participants were 133 Dutch students in Grade 8. Criterion variables were English course grades and scores on a standardized achievement test (Cito-VAS). Alternate-form reliability ranged from r = .77 to .87. Correlations between CBM and criterion measure scores ranged from r = -.04 to .65. Scores from maze selection and reading aloud alone predicted English-language proficiency better than a combination of scores from the four measures, explaining 29.7% and 23.6% of the variance, respectively. Implications for the use of CBM for FL screening and progress-monitoring are discussed.

Keywords: Foreign-language learning, curriculum-based measurement, progress monitoring, technical adequacy, secondary school

In our globalized society, mastering languages other than one's native language is essential. For example, many universities in the United States require foreign-language credits for admission and/or graduation (see Campus Explorer, 2019; Grove, 2019). In 2016, more than 1.4 million students were enrolled in foreign-language courses at institutions of higher education in the U.S. (Looney & Lusin, 2018). In Europe, the ambition is to have 75% of young citizens master two foreign languages (Dutch Education Council, 2008). In 2015, 98.6% of lower secondary-level students in the European Union studied at least one foreign language, of which English was by far the most common (Eurostat, 2017).

Difficulties in Foreign-Language Learning

Although many students learn a foreign language without difficulty, others struggle. One of the best predictors of foreign-language (FL) learning ability is native language ability (Domínguez de Ramírez & Shapiro, 2007; Ganschow et al., 1998; Sparks, 2008; Sparks et al., 2006). It is perhaps not surprising, then, that students with learning disabilities (LD) often are considered to be at risk for FL learning difficulties (Skinner & Smith, 2011), and are granted waivers or substitutions for FL courses. However, not all students with LD experience difficulties with FL learning, and not all students who have difficulties with FL learning have LD (DiFino & Lombardino, 2004; Sparks, 2006, 2016; Wight, 2015).

Sparks (2006, 2009) suggested that FL learning ability be viewed as occurring along a continuum, and that identification of students at risk for FL difficulties be made on the basis of performance rather than labels. Students who are identified as being at risk could then be monitored and, if necessary, provided specialized teaching methods and accommodations to enhance their FL learning (Skinner & Smith, 2011; Sparks et al., 2002). A tool that could be used to screen and identify

Hoefnagel, L., Espin, C. A., & Rippe, R. (2021). CBM Language Measures as Indicators of Foreign-Language Learning: Technical Adequacy of Scores for Secondary-School Students. *International Journal for Research in Learning Disabilities, 5*(1), 42-57. https://doi.org/10.28987/ijrld.5.1.42

at-risk students, monitor their progress, and evaluate the effectiveness of specialized methods and accommodations is Curriculum-Based Measurement (CBM).

Curriculum-Based Measurement

CBM is a simple procedure for repeated measurement of student growth toward long-range instructional goals in academic areas (Deno, 1985). Using CBM, teachers measure student progress on a frequent basis (e.g., once a week) using brief samples of work, and place the scores on a graph that depicts progress. They subsequently examine the progress graph to determine the effectiveness of instruction.

CBM measures are designed to be practical (simple, time efficient, easy to administer and score) and to produce scores that serve as valid and reliable indicators of performance and progress in an academic area (Deno, 1985; Espin & Deno, 2016). A considerable body of research has examined the validity and reliability of scores from CBM measures in reading and writing (see McMaster & Espin, 2007; Wayman et al., 2007), but this research has primarily been conducted with students in their native language. And while research has been carried out on the development of CBM measures for English Learners (EL; e.g., Baker & Good, 1995; Campbell et al., 2013; Domínguez de Ramírez & Shapiro, 2006, 2007; Sandberg & Reschly, 2011), the findings cannot automatically be generalized to FL learning as the situations under which EL and FL students learn a second language differ.

In searching the literature, we located only one study that examined CBM measures of FL learning (Chung & Espin, 2013). Chung and Espin examined the technical adequacy of scores from maze selection, Dutch-to-English word translation, and English-to-Dutch word translation, both alone and in combination, as indicators of FL learning for middle-school students. For each measure, different time frames and scoring procedures were compared. Criterion variables in the study were English course grades and scores on a standardized English reading test. Results varied somewhat across grade and skill level but provided tentative support for scores from maze-selection (2 min, correct minus incorrect choices) and word-translation measures (English-to-Dutch- or Dutch-to-English, 2 min, correct translations) as indicators of FL performance. In addition, results demonstrated that a combination of scores from maze and English-to-Dutch word translation accounted for a greater proportion of variance in English course grades than scores from either measure alone.

Despite these important findings, the Espin and Chung study (2013) has limitations. First, the study did not include a reading-aloud measure. CBM reading aloud often is used to monitor progress in one's native language (Wayman et al., 2007). Second, for some of the analyses, sample sizes were small because different CBM and criterion measures were used across grade and educational levels. Finally, there was a ceiling effect for the maze scores.

The Present Study

Given the importance of FL learning in today's globalized society, and the number of students who struggle to learn an FL, it seems important to replicate the Chung and Espin (2013) study, addressing the limitations of the study wherever possible.

The present study was a replication and extension of Chung and Espin (2013). Specifically, the study examined the reliability and validity of scores from four CBM measures, alone and in combination, as potential indicators of FL learning. The four measures were maze selection, Dutch-to-English word translation, English-to-Dutch word translation, and reading aloud. To avoid some of the limitations of the Chung and Espin study, we increased the length of the maze passages to avoid ceiling effects and, to the extent possible, used identical measures across educational levels.

Two research questions were addressed in the study:

1. What are the reliability and validity of scores from four CBM measures as potential indicators of English FL performance?
2. Does a combination of scores predict English FL performance better than scores from a single measure?

Based on the results of Chung and Espin (2013), we expected that scores from maze-selection and word-translation measures (both English-to-Dutch and Dutch-to-English) would be reliable and would significantly relate to scores on the criterion measures. Further, we expected that a combination of scores from maze selection and English-to-Dutch word translation would account for a greater proportion of variance in the criterion variables than scores from either measure alone. Because Chung and Espin (2013) did not include a reading-aloud measure, we had no expectations related to scores from the reading-aloud measures.

The present study can be characterized as Stage 1 CBM research (Fuchs, 2004), where the focus is on the technical adequacy of scores as indicators of performance. Given the focus on technical adequacy, partici-

pants in our study represented a range of performance levels. Results provide information on the extent to which the CBM scores accurately rank order students on their English FL performance, and have implications for the use of the measures to screen and identify at-risk students. The findings also inform future Stage 2 research, where the focus is on the use of the measures for progress monitoring.

Method

Participants

Participants were 133 eighth-grade students (67 males, 66 females; M_{age} = 13.60, SD = .69, age range 12–16 years) from 15 classrooms in three secondary schools in The Netherlands. Schools were located in three middle-large to large cities in the west and central part of the country. Schools were selected via the researchers' networks. Participants were recruited via their English-language courses. All students were invited to participate.

Secondary education in The Netherlands is divided into different educational levels, which are (from the lowest to highest): vocational-low, vocational-high, professional, and university preparation. English as a FL is mandatory in all Dutch secondary schools, and the national curriculum for English consists of reading, writing, listening, and speaking. The curriculum follows the Common European Framework of Reference for Languages (CEFR), with CEFR target levels set for the end of secondary school for each educational level (www.erk.nl).

Participants in the study were in their second year of formal English education and represented all educational levels: vocational-low (12.8%), vocational-high (11.3%), professional (16.5%), university (36.1%), and combined professional/university (23.3%) levels. Fourteen students (10.5%) were enrolled in a bilingual education program in which at least 50% of core courses were provided in English for the first three years of secondary school. These students were from the university education levels. Home-language information was available for 40% of the students. For all these students, Dutch was spoken at home.

Fifteen percent of participants were students with dyslexia. The diagnosis of dyslexia in The Netherlands is based on significant delays in reading and/or spelling, despite systematic and frequent intervention, where delays are not due to an intellectual or sensory disability, or to inadequate education (De Jong et al., 2016). The prevalence of dyslexia is estimated to be 3.6% at the end of primary school; that is, sixth grade (Blomert, 2005). However, the Dutch Inspectorate of Education (2019) reports that the percentage of students actually labeled with dyslexia in sixth grade is 7.5%. This number increases sharply at secondary school, to 11.9% and 13% of seventh- and ninth-grade students, respectively. These percentages are similar to the 15% of students with dyslexia in the current sample.

Predictor Variables

Predictor variables were scores from four CBM FL measures: reading aloud, maze selection, English-to-Dutch word translation, and Dutch-to-English word translation.

Reading Aloud

Reading-aloud passages were two English narrative texts selected from Children's Educational Services passages (Deno & Marston, 1987). Passages were 488 and 498 words in length, written for students in Grade 4, and were non-culturally specific. Students read aloud from each passage for 1 min, whereupon the number of correct (WRC) and correct minus incorrect (WRCI) words read were scored. Incorrect words included mispronunciations, word substitutions, omissions, reversals, and words supplied by the examiner when a student did not know a word.

Maze Selection

Maze selection passages were constructed from the same English narrative texts used for reading aloud to minimize differences in results due to text effects. To create the maze, the first sentence was left intact, after which every seventh word was deleted and replaced by the correct word and two distractors. The three choices were placed in bold print and underlined in the text, and were not split across lines. If the seventh word was a proper noun, it was left and, instead, the next word was deleted. The placement of the correct choice varied. Distractors were approximately equal in length (within one letter) to the correct choice, and were clearly incorrect (for guidelines, see Conoyer et al., 2017; Fuchs & Fuchs, 1992). Students read each maze text silently for 2 min, circling the word that restored meaning to the passage. Scores were the number of correct (MCC) and correct minus incorrect maze (MCCI) choices. Scoring was carried out with and without a guessing rule. With the guessing rule, scoring was stopped after three consecutive incorrect choices.

CBM maze selection is similar to the modified- or multiple-choice cloze measures often used in FL assessment (Hale et al., 1989; Porter, 1976) with one key difference. In typical FL modified-cloze measures, distractors are similar in meaning and syntax to the target word (Porter, 1976). In CBM mazes, on the other hand, distractors are selected to be clearly different in meaning and syntax from the target word so that one answer is obviously correct (Fuchs & Fuchs, 1992). Chung and Espin (2013) reported alternate-form reliability for CBM maze scores ranging from r = .69 to .78 and validity from r = .20 to .79, with higher reliabilities reported for MCCI than for MCC.

Word-Translation Measures

Dutch-to-English and English-to-Dutch word-translation measures consisted of a list of 50 words (25 words per page) with a blank next to each word. Words were randomly selected from an English-language curriculum used in Dutch secondary schools. All parts of speech were represented on each measure. Students wrote as many translations as possible in 2 min. Scores were the number of correct (WTC) and correct minus incorrect (WTCI) translations.

Based on the results of Chung and Espin (2013), a decision was made to count translations as correct only if they were spelled correctly. Chung and Espin (2013) reported alternate-form reliability for word translation scores ranging from r = .76 to .88 for WTC and from r = .59 to .78 for WTCI, and validity from r = .44 to .77 for WTC. Validity for WTCI was not examined in Chung and Espin (2013) because the reliabilities were low.

Criterion Variables

Criterion variables in the study were English course grades and scores on a standardized English-language test (Cito-VAS).

English Course Grades

English course grades were average grades across three grading periods in the school year. Grades were based on the students' performance in reading, listening, writing, speaking, vocabulary, and grammar within the individual student's educational level. Grades ranged from 1 (low) to 10 (high), and were reported to one decimal point. A grade of 5.5 was passing. Grades are assigned within educational level; thus, a grade of 7 in English in a vocational-low level program was not equivalent to a grade of 7 in a university-preparation level program. Analyses involving grades were, therefore, carried out within educational level.

Cito-VAS Scores

The Cito-VAS test (Cito, 2015) is a standardized achievement test administered in many Dutch secondary schools in the middle of the school year. In our study, two of the three participating schools administered the Cito-VAS. (The school with students in combined professional/university levels did not administer the test.) Scores from the English reading and English vocabulary subtests of the Cito-VAS were used in the study.

The English reading subtest consisted of expository passages, each with 1-3 multiple-choice questions, for a total of 35 questions. The English vocabulary subtest consisted of multiple-choice items in which students had to choose (a) the correct Dutch translation of the underlined word in an English sentence, (b) the correct English word to complete a sentence, (c) a synonym or an antonym for an English word, or (d) the word that did not belong in a set of words. The vocabulary subtest included a total of 45 items. Each subtest took approximately 50 minutes to complete. Different forms of the Cito-VAS were administered at different educational levels. Therefore, standard scores were used in the analysis, enabling comparisons across test levels.

Technical adequacy information for the Cito-VAS was available only for an earlier version of the test that did not include the English vocabulary subtest (Van Til & Van Boxtel, 2015). Cronbach's alphas for the English reading subtest were reported to be .76, .78 and .80. With regard to validity, a consistent increase in mean scores across grade and educational levels was reported, and correlations between subtests measuring different constructs were found to be weaker (r = .25 - .42) than between subtests measuring overlapping constructs (r = .58 to .72). Finally, standard scores on the English reading subtest for eighth-grade students were found to predict educational level one year later (r = .63).

Procedure

Participants completed the measures in the following order: maze selection, English-to-Dutch word translation, Dutch-to-English word translation, and reading aloud. For all CBM measures, two parallel forms were administered, with the order of the forms counterbalanced. Maze-selection and word-translation measures were administered in a group setting. Reading aloud was administered individually on the same day or within the same week. If a student was absent for part of the data collection, every attempt

was made to schedule a make-up session. Data were collected and scored by four master's-level students who were trained in two 1.5-hour training sessions. Two data collectors were present for all data collection. English course grades, Cito-VAS scores, and student background information were collected from the schools at the end of the school year.

Scoring

All measures were scored by two coders. Interscorer agreement was calculated by dividing the smaller by the larger score and multiplying by 100. Agreement was calculated separately for each score. For reading aloud, agreement was 99.2% (WRC) and 98.9% (WRCI). For maze selection, agreement was 99.8% (MCC) and 99.8% (MCCI). For English-to-Dutch word translation, agreement was 99.0% (WTC) and 97.4% (WTCI). For Dutch-to-English word translation, agreement was 98.2% (WTC) and 94.7% (WTCI). Disagreements were discussed and resolved before coming to a final score.

Results

Data Inspection

Data inspection indicated normal distributions for all independent variables and no substantial univariate outliers. To check for bivariate outliers, multivariate scatterplots were inspected. The patterns in the scatterplots revealed approximately linear associations between the independent and dependent variables. One possible bivariate outlier was detected in nearly every scatterplot. For this student, who was at the university preparation level and was diagnosed with dyslexia, Cito-VAS scores and English course grades were relatively high, whereas scores on the CBM measures were relatively low. Removal of this outlier yielded a change in explained variances (for example from $R^2 = .35$ to $R^2 = .42$ for the relation between maze selection MCC and Cito-VAS scores). Because of the disproportionally large effects of the student's scores on the strength of the correlations, analyses were conducted both with and without the outlier. The association patterns were the same with and without the outlier, but the results were somewhat stronger (i.e., correlation coefficients increased) when the outlier was removed. We report results *without* the outlier for the validity analyses.

Handling of Missing Data and Assumptions

Patterns of missing observations were checked. Little's MCAR test (Little, 1988) showed that no patterns in missingness could be detected; $\chi 2(8, N = 133) = 5.83$, $p = .666$; therefore, any missingness was considered to be completely at random. Analyses were based on full information maximum likelihood (FIML) estimation (Graham et al., 1996), with which missingness is commonly handled within the analysis model (Dempster et al., 1977) as it yields the most likely parameter values given all available data in the model, regardless of their level of completeness.

Descriptive Analyses

Means and standard deviations for scores on the CBM measures (alternate forms and combined) are reported in Table 1. On average, students read aloud approximately 150 correct and 5 incorrect words in 1 min, made approximately 22 correct and 0.5 to 1.0 incorrect maze choices in 2 min (depending on whether a guessing rule was applied or not), translated approximately 25 words correct and 4 incorrect from English to Dutch and approximately 19 words correct and 5 incorrect from Dutch to English. Means and standard deviations broken down by gender are reported in Table 2. Girls tended to score higher on the CBM measures than boys, but differences were not large.

There were significant differences in mean scores between Forms A and B for reading aloud (WRC, $t(121) = 6.55$, $p < .001$; WRCI, $t(121) = 6.42$, $p < .001$) and English-to-Dutch translation (WTC, $t(129) = 11.42$, $p < .001$; WTCI, $t(129) = 9.72$, $p < .001$; Bonferroni correction applied), but not for maze selection or Dutch-to-English translation. Further, no significant differences in mean scores were found between scoring with or without use of a guessing rule for the maze.

Means and standard deviations for Cito-VAS English vocabulary and English reading subtests, broken down by educational level, are reported in Table 3. (Recall that scores were not available for one school.) Means and standard deviations for the English course grades are reported by educational level in Table 3. Individual grades ranged from 4.30 to 9.27 ($M = 7.14$, $SD = 1.03$).

Table 1
Means and Standard Deviations of the CBM Measures Form A, Form B, and Mean of A and B

	Form A		Form B		Mean (A+B)	
Measure / Score	*N*	*M (SD)*	*N*	*M (SD)*	*N*	*M (SD)*
Reading Aloud						
WRC	122	155.46 (37.72)	122	144.29 (31.75)	122	149.87 (33.56)
WRCI	122	150.43 (39.46)	122	139.22 (33.08)	122	144.82 (35.11)
Maze, guessing rule						
MCC	131	21.91 (8.70)	132	21.96 (7.53)	132	21.88 (7.84)
MCCI	131	21.35 (9.00)	132	21.36 (7.75)	132	21.30 (8.10)
Maze, no guessing rule						
MCC	131	22.34 (7.99)	132	22.30 (7.04)	132	22.26 (7.24)
MCCI	131	21.37 (8.97)	132	21.31 (7.91)	132	21.28 (8.13)
English-to-Dutch						
WTC	130	26.68 (7.08)	130	23.15 (5.71)	130	24.92 (6.19)
WTCI	130	23.33 (8.42)	130	18.80 (6.71)	130	21.05 (7.13)
Dutch-to-English						
WTC	131	19.70 (8.61)	131	18.63 (8.16)	131	19.17 (8.09)
WTCI	131	14.16 (9.92)	131	13.30 (9.63)	131	13.73 (9.31)

Note. WRC = words read correct. WRCI = words read correct minus incorrect. MCC = maze choices correct. MCCI = maze choices correct minus incorrect. WTC = words translated correct. WTCI = words translated correct minus incorrect.

Table 2
Means and Standard Deviations of the CBM Measures by Gender

	Males		Females	
Measure / Score	*N*	*M (SD)*	*N*	*M (SD)*
Reading Aloud				
WRC	63	148.58 (35.27)	58	151.76 (31.92)
WRCI	63	143.67 (36.52)	58	146.51 (33.92)
Maze, guessing rule				
MCC	66	20.95 (8.43)	65	22.88 (7.17)
MCCI	66	20.31 (8.57)	65	22.35 (7.54)
Maze, no guessing rule				
MCC	66	21.65 (7.42)	65	22.94 (7.08)
MCCI	66	20.27 (8.65)	65	22.36 (7.54)
English-to-Dutch				
WTC	64	24.17 (6.52)	65	25.71 (5.82)
WTCI	64	20.44 (7.38)	65	21.70 (6.92)
Dutch-to-English				
WTC	65	18.45 (7.90)	65	19.95 (8.31)
WTCI	65	13.48 (8.70)	65	14.01 (10.01)

Note. WRC = words read correct. WRCI = words read correct minus incorrect. MCC = maze choices correct. MCCI = maze choices correct minus incorrect. WTC = words translated correct. WTCI = words translated correct minus incorrect.

Table 3
Means and Standard Deviations for Cito-VAS English Vocabulary and English Reading Subtests and for English Course Grades by Educational Level

		Cito-VAS Vocabulary	Cito-VAS Reading	English course grades	
Educational level	*N*	*M (SD)*	*M (SD)*	*N*	*M (SD)*
Vocational-low	11	152.91 (21.49)	136.36 (17.83)	17	7.27 (.79)
Vocational-high	15	173.83 (31.59)	155.19 (17.23)	15	7.03 (.72)
Professional	21	173.95 (28.46)	150.95 (14.41)	22	6.64 (1.13)
Professional/university	-	-	-	31	6.57 (.95)
University	24	195.47 (32.66)	169.80 (19.69)	48	7.73 (.88)
Total	71	177.94 (32.55)	155.96 (20.62)	133	7.14 (1.03)

Note. WRC = words read correct. WRCI = words read correct minus incorrect. MCC = maze choices correct. MCCI = maze choices correct minus incorrect. WTC = words translated correct. WTCI = words translated correct minus incorrect.

Reliability Analyses

To assess alternate-form reliability, Pearson correlations between scores on parallel forms of each measure were computed. Reliability coefficients ranged from r = .77 to .87, with all but one coefficient (English-to-Dutch translation) above .82 (see Table 4). All correlations were statistically significant, with all p-values < .001. Alternate-form reliability coefficients were high despite significant mean differences between Forms A and B for reading aloud and English-to-Dutch translation, indicating that, even though students scored higher on Form A than on Form B, the rank ordering of students remained similar across the forms. Mean scores across Forms A and B were used for the subsequent validity analyses to increase the stability of the scores.

Validity Analyses

To reduce the number of statistical tests, a limited number of scores were carried forward for the validity analysis. Selection of scores was based on the reliability coefficients, the efficiency of scoring procedure, and on whether the scoring procedure was typically used in other CBM research. The following scores were selected for the validity analysis: WRC for reading aloud, MCCI with use of a guessing rule for maze selection, and WTC for both word-translation measures. Mean scores across forms A and B were used for all analyses.

Correlations With Criterion Variables

Correlations between CBM scores and the Cito-VAS scores were statistically significant, ranging from r = .31 to .65 (see Table 5). In general, correlations for reading aloud and maze selection were higher than for the translation tasks; the lowest correlations were found for English-to-Dutch translation. Correlations tended to be somewhat higher with the Cito-VAS reading subtest than the vocabulary subtest, but differences were small.

Correlations with English course grades were computed within educational level, resulting in samples ranging from 14 to 48 students per subgroup. Means and standard deviations for the CBM scores, broken down by educational level, are reported in Table 6. In general, as illustrated, mean scores increased across educational level although scores for combined professional/university were higher than for university only. Correlations between CBM scores and English course grades ranged from r = -.04 to .65 (see Table 7). Across educational levels, correlations tended to be lowest for the English-to-Dutch translation, but patterns for the other measures differed somewhat. For example, for vocational and professional educational levels, coefficients tended to be higher for reading aloud and maze selection than for word-translation measures, and for professional/university and university levels, coefficients for word-translation measures were as high as or higher than for reading aloud and maze.

Regression Analyses

To examine whether a combination of measures predicted English-language proficiency better than a single measure, latent linear regression models were tested in different stages, evaluating performance through different compositions of a latent performance score. Multilevel models revealed negligible intra-class correlations on educational level (ICCs for average Cito scores = .06, .03 for Cito vocabulary, and .08 for Cito

Table 4
Alternate-Form Reliability Coefficients of Scores From the CBM Measures

	Reading Aloud				Maze selection	
Scoring	*N*	*r*	Guessing rule	Scoring	*N*	*r*
WRC	122	.87	Rule	MCC	131	.85
WRCI	122	.87		MCCI	131	.85
			No rule	MCC	131	.83
				MCCI	131	.84

English-to-Dutch translation			Dutch-to-English translation		
Scoring	*N*	*r*	Scoring	*N*	*r*
WTC	130	.87	WTC	131	.86
WTCI	130	.77	WTCI	131	.82

Note. All correlations significant at $p < .001$ level.

Table 5
Correlations Between CBM Scores and Cito-VAS Scores

Measure	Cito Vocabulary	Cito Reading
Reading aloud WRC (*N* = 65)	.56***	.65***
Maze selection MCCI (*N* = 69)	.63***	.63***
English-to-Dutch WTC (*N* = 68)	.31*	.34**
Dutch-to-English WTC (*N* = 69)	.50***	.52***

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

Note. WRC = words read correct; MCCI = maze choices correct minus incorrect; WTC = words translated correctly.

Table 6
Means and Standard Deviations for Selected CBM Scores by Educational Level

Educational level	WRC	MCCI	WTC E-D	WTC D-E
Vocational-low	120.00 (31.57)	11.32 (7.86)	16.88 (5.22)	11.88 (5.85)
Vocational-high	139.23 (33.53)	17.11 (8.78)	19.62 (5.25)	13.36 (6.59)
Professional	133.89 (31.11)	22.52 (6.93)	26.36 (4.08)	17.02 (6.18)
Combined professional/university	169.69 (25.95)	25.33 (5.81)	28.15 (5.14)	24.31 (6.94)
University	152.07 (28.49)	22.23 (6.10)	25.47 (4.65)	19.02 (7.44)
Total	150.15 (33.56)	21.42 (8.06)	24.93 (6.20)	19.12 (8.09)

Note. WRC = words read correct; MCCI = maze choices correct minus incorrect; WTC E-D = words translated correct, English-to-Dutch; WTC D-E = words translated correct, Dutch-to-English. Standard deviations are in parentheses.

Table 7
Correlations Between CBM Scores and Average English Grades Within Educational Level

	Reading aloud WRC		Maze selection MCCI	
Educational level	*N*	*r*	*N*	*r*
Vocational low	16	.48	17	.57*
Vocational high	14	.65*	15	.54*
Professional	18	.51*	21	.63**
Professional/university	30	.38*	30	.40*
University	43	.26	47	.54***
	English-to-Dutch WTC		Dutch-to-English WTC	
Educational level	*N*	*r*	*N*	*r*
Vocational low	16	.43	16	.47
Vocational high	14	.35	15	.44
Professional	21	-.04	21	.30
Professional/University	31	.46*	31	.50**
University	48	.30*	47	.53***

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

Note. WRC = words read correct; MCCI = maze choices correct minus incorrect; WTC = words translated correctly.

reading, respectively); thus, analyses did not account for variance attributable to educational-level characteristics (Luke, 2004).

The approach used to test the linear regression models was as follows. First, we compared a model in which performance was indicated by the Cito-VAS subtest scores only, hereafter Cito-Only Model, to a model in which a latent performance score was indicated by both the Cito-VAS subtest scores and the English course grades, hereafter Gold Standard Model, under planned missing data assumptions (Little & Rhemtulla, 2013; Rippe & Merkelbach, 2019). In the latter approach, latent performance scores for individuals with missing data on the Cito-VAS subtests were approximated based on the English course grades and their correlation with the Cito-VAS subscales. These latent scores were not constructed explicitly before being entered into the model; instead, they were "estimated" implicitly within the model itself based on maximum likelihood estimations of the latent variable regression coefficients, accounting for the covariance between the (two) observed outcome scores.

As a consequence of using the latent variable approach, most classical multicollinearity measures could not be computed. Through inspection of all pairwise correlations, no indication of multicollinearity was found. All correlations were .70 or lower, with only one exception: the correlation between Dutch-to-English and English-to-Dutch word translations was .73, meaning 53% of their variance was shared. Variance inflation factors were well below 10, ranging between 2.55 and 3.33.

In the first stage, the two latent performance variants were evaluated on overall model fit only with all possible combinations of predictors. Model parameters were not interpreted. Based on overall model fit and parameter effect size, for parameter interpretation a subset of models was reevaluated as observed-variable-only models using a proportional bootstrap with 1000 samples to obtain standard errors.

For model estimations, we used Lavaan (Rosseel, 2012) version 0.6-4 in R version 3.5.3. Full information maximum likelihood was used to handle missing data. The number of EM iterations for FIML was set at a maximum of 5,000. To determine the fit of each model, the comparative fit index (CFI), normed fit index (NFI), standardized root mean residual (sRMR), and root mean square error of approximation (RMSEA) were inspected. The CFI and NFI should be as high as possible (above .90), while the sRMR and the RMSEA should be as low as possible (below .06). The RMSEA and sRMR can yield contrasting conclusions, as the sRMR is a simple absolute fit index comparing observed and predicted correlations without accounting for complexity, while the RMSEA is based on the non-centrality parameter and can be considered more precise.

Preliminary analyses revealed that there was no intra-class effect of educational level and no differential effect of gender. Therefore, educational level and gender were not accounted for in the regression analyses. Moreover, preliminary analyses showed that the models with the Gold Standard latent performance score yielded a less favorable fit-complexity ratio than the models with the Cito-Only latent performance score, indicated by the Akaike information criterion (AIC). A lower AIC relative to the other model means a better trade-off between the fit and the complexity of the model. The English course grades did not add any unique information to the latent performance score beyond the scores from the two Cito-VAS subtests. Therefore, models with the dependent latent performance score consisting only of the two Cito-VAS subtests were used in the subsequent analyses.

As a first step in the regression analysis, standard assumptions were checkedfor the subsample of students with Cito-VAS scores (N = 63). No violations on multicollinearity, normality, and homoscedasticity were found, although correlations among the predictors were high, ranging from $r = .51$ to .78 (see Table 8), suggesting caution in interpretation. s

In Stage 2, after all combinations of predictors had been compared in the Stage 1 analyses, a selection of models were reevaluated as observed-variable-only models using bootstrapped standard errors with 1,000 runs. The models were selected based on their fit. An overview of the seven models selected is presented in Table 9, with their respective fit indices provided in Table 10. The AIC allows only for comparing Models 2 through 7 to Model 1, because these models are nested. The model with all four predictors (Model 1) had the best fit in terms of complexity trade-off, as indicated by the lowest AIC value (AIC = 1,103.00; see Table 10). Even though the model consisting of all four CBM measures as predictors is the most complex, it outperformed simpler one- and two-predictor models in terms of the balance between fit of the model and complexity. The two-predictor models (Models 6 and 7) were not favored over single predictor models. For Model 6, the AIC was high. For Model 7, although the AIC was low, the RMSEA was unfavorable (.14). Therefore, neither model qualified for further interpretation. Among the simpler single-predictor models, Model 5 (Reading Aloud) resulted in the smallest difference with Model 1 in terms of AIC value (1148.72 vs. 1103.00, respectively).

Further model evaluation was based on both absolute and comparative fit using the NFI, CFI, RMSEA and sRMR values (see Table 10). As shown, all models had high values (equal to or closely approaching 1.0) on the NFI and CFI, and low values (approaching 0) on the RMSEA and sRMR, indicating good fit with little error. The all-predictor Model 1 showed somewhat poorer values on some of the indices (NFI = .97, sRMR = .01). The single-predictor models (Models 2 to 5) showed the best fit (NFI and CFI = 1.00, RMSEA and sRMR < .001).

In the above models, contributions of both the English vocabulary subtest and the English reading subtest to the latent performance score were significant ($\beta = .76$ to .83, $z = 5.67$ to 15.58, $p < .001$ and $\beta = .89$ to .97, $z = 7.48$ to 17.93, $p < .001$, respectively), with one exception: In Model 3, the contribution of the English vocabulary subtest to the latent performance score was significant ($\beta = .78$, $z = 5.67$, $p < .001$), while the English reading subtest was not significant ($\beta = .95$, $z = 1.31$, $p = .192$). This discrepancy can be explained by the fact that the models describing the same amount of variance do not necessarily describe the same part of the variance in the outcome.

In Stage 3, Models 1, 2, and 5 were selected as final models. For these models, the contribution of the predictors within the model was evaluated. The estimated regression coefficients are displayed in Table 11. In the all-predictor model (Model 1), scores from maze selection ($\beta = 0.59$, $z = 3.99$, $p < .001$) and English-to-

Table 8
Correlations Among the CBM Measures

Measure	Reading aloud WRC	Maze selection MCCI	English-to-Dutch WTC	Dutch-to-English WTC
Reading Aloud WRC	-	.70	.51	.74
Maze selection MCCI		-	.67	.71
English-to-Dutch WTC			-	.78
Dutch-to-English WTC				-

Note. All correlations significant at $p < .001$.

Note. These are correlations for the subsample of students with Cito-VAS scores ($N = 63$). WRC = words read correct; MCCI = maze choices correct minus incorrect; WTC = words translated correctly.

Dutch word translation (β = -0.34, z = -2.49, p = .012) contributed significantly to the prediction of the latent performance score (see Table 11). Removing the overlapping contribution of the four CBM measures, maze selection had the strongest unique contribution to the prediction of the latent English performance score, explaining 21.9% of the remaining total variance. The unique contribution of English-to-Dutch word translation was in the negative direction, and explained 7.3% of the variance. Thus, after accounting for the overlap between the CBM measures in the prediction of the latent English performance score, maze selection had the strongest contribution in the positive direction whereas English-to-Dutch had the next strongest contribution, but in the negative direction.

In the single-predictor model with maze selection (Model 2), scores from maze selection contributed significantly to the prediction of the latent performance score (β = .71, z = 10.41, p < .001). Maze selection explained 29.7% of the total variance in the latent performance score when the other CBM measures were not accounted for. In the single-predictor model with reading aloud (Model 5), scores from reading aloud contributed significantly to the prediction of the latent performance score (β = .66, z = 6.32, p < .001). Finally, reading aloud explained 23.6% of the total variance in the latent performance score when the other CBM measures were not accounted for.

Discussion

The goal of this study was to examine the technical adequacy of scores from four CBM measures – reading aloud, maze selection, Dutch-to-English, and English-to-Dutch word translation – as potential indicators of FL learning, replicating and extending an earlier stusdy by Chung and Espin (2013). In general, our results provided the greatest support for maze selection and reading aloud scores as general indicators of FL performance.

The first research question addressed the alternate-form reliability and validity of scores from the CBM measures. Reliability coefficients were high, with all but one coefficient falling between .82 and .87. The coefficients for maze selection and for English-to-Dutch word translation were higher than those found by Chung and Espin (2013), where coefficients were below r = .80. Differences may be due to the fact that in the present study, the same word translation measure was administered across educational levels, whereas Chung and Espin used a different form of the measure for lower and higher educational levels. Variability in scores was greater in the present study (*SDs* = 5.71 to 9.92; see Table 1) than in the Chung and Espin study (*SDs* = 2.64 to 7.24).

The effects of different scoring procedures on alternate-form reliability were also examined. For reading aloud and maze selection, reliability was not affected by scoring procedure (correct vs. correct minus incorrect or with vs. without a guessing rule). For the word translation tasks, consistent with the findings of Chung and Espin (2013), higher reliabilities were found for correct than for correct minus incorrect scores.

A select number of scores were carried forward for validity analysis: WRC for reading aloud, MCCI for maze selection, and WTC for the word translation tasks. The patterns of correlations differed across criterion measure and across educational level. For Cito-VAS, correlations could be computed across educational level. These correlations ranged from r = .31 to .65, a range similar to that reported by Chung and Espin (2013; range r = .37 to .79). Correlations with the Cito-VAS were higher for reading aloud and maze selection (r =.56 to .65) than for word translation (r = .31 to .52). For English course grades, correlations had to be computed within educational level. The patterns of results differed by educational level: (a) at lower educational levels, stronger correlations

Table 9
Final Models With the Latent Performance Score From Cito-VAS English Vocabulary and English Reading Subtests as Dependent Variable

Model	Predictors
Model 1	Maze MCCI + English-Dutch WTC + Dutch-English WTC + Reading Aloud WRC
Model 2	Maze MCCI
Model 3	English-Dutch WTC
Model 4	Dutch-English WTC
Model 5	Reading Aloud WRC
Model 6	Maze MCCI + English-Dutch WTC
Model 7	Maze MCCI + Reading Aloud WRC

Table 10
Model Fit and Complexity Trade-Off for the Selected Models

Model	*N*	AIC	NFI	CFI	RMSEA	sRMR
1	63	1103.00	.97	1.00	< .001	.01
2	69	1210.06	1.00	1.00	< .001	< .001
3	68	1225.43	1.00	1.00	< .001	< .001
4	69	1227.99	1.00	1.00	< .001	< .001
5	65	1148.72	1.00	1.00	< .001	< .001
6	68	1188.82	1.00	1.00	< .001	.01
7	64	1125.33	.98	.99	.14	.01

Table 11
Coefficients From Final Models on the Latent Performance Score

Predictors	β	*SE*	*z*	*p*	95% CI	Total variance
Model 1 (*N* = 63)						
Maze MCCI	0.59	0.15	3.99	< .001	[0.30, 0.88]	.219
English-Dutch WTC	-0.34	0.14	-2.49	.012	[-0.61, -0.07]	.073
Dutch-English WTC	0.26	0.17	1.58	.115	[-0.06, 0.59]	-
Reading Aloud WRC	0.27	0.16	1.71	.089	[-0.04, 0.58]	-
Model 2 (*N* = 69)						
Maze MCCI	0.71	0.07	10.41	< .001	[0.58, 0.85]	.297
Model 5 (*N* = 65)						
Reading Aloud WRC	0.66	0.10	6.32	< .001	[0.45, 0.86]	.236

Note. CI = confidence interval. β = standardized estimate.

were found for reading aloud and maze selection than for word translation; (b) at higher educational levels, stronger correlations were found for Dutch-to-English word translation and maze selection than for the reading aloud. These results may reflect the importance of English vocabulary knowledge at more advanced levels of English-language learning and the greater sensitivity of a reading-aloud measure for beginning learners.

Consistent across all educational levels was the finding that scores on English-to-Dutch word translation tended to result in lower correlations with the criterion variables. Similarly, Chung and Espin (2013) found low correlations between English-to-Dutch word-translation and Cito-VAS scores (although not with English course grades). The lower validity coefficients for English-to-Dutch word-translation scores might be related to the fact that students had to spell the Dutch words correctly; thus, their scores on the task reflected both English-language knowledge and Dutch spelling ability. Although students also had to spell the English words correctly, perhaps if they knew what the English word was, they also knew how to spell it. A second, more likely, explanation might be the lower variability in English-Dutch translation scores leading to an attenuation in correlations. For example, standard deviations for English-to-Dutch translation were smaller than for Dutch-to-English translation.

Even though the English-to-Dutch translation produced the smallest validity coefficients, it may be prudent to not yet discard the measure as a potential CBM FL measure. English-to-Dutch translation requires recognition rather than production, and thus might serve as a good measure for students who are just beginning to learn English.

The second research question examined whether a combination of measures predicted FL proficiency in English better than a single measure. The sample size was relatively small for this analysis (*N* = 63), and the predictors correlated with each other; thus, the results should be considered suggestive. Findings showed that

a combination of measures did not predict FL proficiency in English better than a single measure. Although all models had adequate fit, the single-predictor models had the best fit. Thus, a single measure contributed more strongly to the prediction of the latent English performance score than a combination of four or two measures. The single-predictor models with either the reading-aloud or maze-selection measure performed the best. Without the overlap with the other CBM measures, maze selection alone accounted for 29.7% of the variance in the latent performance score, and reading aloud alone for 23.6% of the variance.

These findings thus suggest that scores from maze selection are valid indicators of general FL proficiency, a finding that is in line with the results of the simple correlational analyses. Scores from maze selection accounted for nearly 30% of the variance in the latent performance score constructed from scores on the Cito-VAS vocabulary and reading subtests. Given that maze selection takes only 2 minutes to administer, whereas the Cito-VAS subtests together take 100 minutes, 30% is a substantial amount. The slight advantage of the maze selection over reading aloud may be due to the fact that maze selection requires understanding of the text passage and recognition of the words used as choices to fluently progress through the text, perhaps making it a more robust FL indicator than reading aloud. Alternately, both the maze and Cito-VAS reflect silent FL reading, whereas reading aloud also reflects speaking skills.

The differences in model performance were small. The combination of all four measures – although slightly worse than the single-predictor models – yielded good model fit indices as well. In the model with all four CBM measures, maze selection and English-to-Dutch word translation were found to make significant unique contributions to the prediction of the latent performance score. Accounting for the overlapping contribution of the four CBM measures, maze selection still made a significant unique contribution to the prediction of the latent performance score, explaining 21.9% of the remaining variance. Apparently, after accounting for the common contribution of the CBM measures, the maze-selection task measures an additional, different aspect of the construct than the other measures. English-to-Dutch word translation also made a significant unique contribution after accounting for the overlap between the four measures (7.3% of the variance), but in a negative direction. This negative unique contribution, in combination with the lower validity coefficients for English-to-Dutch translation, suggests that the measure demands skills other than FL proficiency, such as spelling in the native language.

Our results diverge from those of Chung and Espin (2013), who found that a combination of maze selection and English-to-Dutch word translation resulted in better prediction than either measure alone. The results from the present regression analyses were based on a larger sample combining all educational levels, and used a latent FL performance score. Thus, although still suggestive, they provide a basis for somewhat firmer conclusions.

It was surprising that the two measures that represented the construct of FL reading showed the highest correlations with the criterion variables and the strongest contributions as single predictors to the prediction of the latent performance score, as opposed to the measures that represented the construct of FL vocabulary knowledge. Because scores from both Cito-VAS English reading and vocabulary subtests contributed to the latent performance score, one might have expected a combination of CBM measures representing reading and vocabulary to best predict student performance. Perhaps vocabulary knowledge is an integral part of reading in the FL. That is, beginning learners need a sufficient level of vocabulary knowledge in order to read a text in the FL (Wallace, 2007). Scores on CBM reading tasks may reflect not only FL reading proficiency but also vocabulary knowledge.

In sum, the results from the regression analyses indicate that a combination of measures does not predict FL proficiency better than a single measure. Practically speaking, this is "good news" in the sense that screening and CBM progress monitoring with a single measure is less time consuming and more feasible in the classroom than using a combination of measures. Determining which single measure to use may depend on practical considerations, however. Maze selection can be administered in a group setting, whereas reading aloud must be administered individually. Thus, although maze selection is more efficient, teachers still may prefer to administer reading aloud because it provides additional information related to the students' ability to pronounce words in the foreign language.

Limitations

One limitation of the present study relates to the criterion measures used. Although course grades and the Cito-VAS have social validity in the sense that both are used to make decisions about students' promotion to the next grade, technical adequacy data on the measures were limited. Although the Cito-VAS is the most widely used standardized achievement test in Dutch secondary education, reliability and validity data were available only for a previous version of the test, and that

version did not include the vocabulary subtest (technical adequacy for the reading subtest was good; see Method section). Although course grades are probably the most commonly used indicator of performance in secondary education and have a large impact on the student's school career, course grades are largely based on teacher judgment and have a restricted range. Nevertheless, the use of both grades and standardized test scores allowed for a convergence of evidence.

A second limitation of the study relates to the sample. First, analyses involving grades had to be conducted within educational level, thereby reducing sample sizes for these analyses. Second, it was not possible to examine whether results varied by language background because native-language information was available for only 40% of students. Finally, students in the university-preparation levels were overrepresented (47.7%) and students in vocational levels underrepresented (24.1%) compared to reported national levels (19% and 55%, respectively; Dutch Inspectorate of Education, 2018). Replication of the study with a larger, more representative sample, therefore, is in order.

Implications

The results of this study have implications for the use of CBM measures in FL instruction. If the results were to be replicated with a larger and more diverse sample, it would provide support for the use of CBM maze and reading aloud as screening measures to identify students who are likely to be at risk for FL learning difficulties. Such students could be provided with additional support and instruction before they begin to fail. In addition, if future Stage 2 progress-monitoring research supports the technical adequacy of scores, the measures could be used to monitor the progress of students with severe and persistent FL learning difficulties and to evaluate the effects of specialized, individualized interventions on that progress.

The increasing need for all students to learn English in our globalized society underscores the need for related screening and progress measures. This need is further underscored by the extent to which some students struggle to learn a foreign language. For example, recall that the percentage of students with dyslexia in The Netherlands increases from 7.5% in sixth grade to 13% in ninth grade. This increase may be related to the increase in language requirements at the secondary-educational level. All Dutch secondary students must learn both Dutch and English. At higher educational levels, students are required to learn up to five FLs (English, French, German, Latin and Greek). A label of dyslexia allows for accommodations in FL learning.

Conclusion and Future Research

In conclusion, our findings support the reliability and validity of (Dutch Inspectorate of Education, 2019) scores from reading-aloud and maze selection measures (and potentially word-translation measures) as potential CBM indicators of English-language learning. Future Stage 2 research must examine the reliability and validity of the growth rates produced by scores from these measures. An important aspect of this work will be to establish the equivalence of alternate forms of the measures. This may prove to be a challenge for reading-aloud and English-to-Dutch word-translation measures, where significant mean differences in scores were found between the alternate forms. Future research also must examine whether teachers' implementation of CBM progress monitoring in FL results in improved instruction and, ultimately, in improved learning for students who struggle to learn a foreign language.

References

Baker, S. K., & Good, R. (1995). Curriculum-based measurement of English reading with bilingual Hispanic students: A validation study with second grade students. *School Psychology Review, 24,* 561-578. doi:10.1080/02796015.1995.12085788

Blomert, L. (2005). *Dyslexie in Nederland* [Dyslexia in The Netherlands]. Retrieved from https://www.boomtestonderwijs.nl/media/14/boek_dyslexie_in_nederland.pdf

Campus Explorer. (2019). *College language requirements.* Retrieved from https://www.campusexplorer.com/college-advice-tips/16292AF6/College-Language-Requirements/

Campbell, H., Espin, C. A., & McMaster, K. (2013). The technical adequacy of curriculum- based writing measures with English learners. *Reading and Writing, 26,* 431-452. doi:10.1007/s11145-012-9375-6

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:10.1177/15345084134897

Cito. (2015). *Cito Volgsysteem Voortgezet Onderwijs. Toets 2* [Cito monitoring system for secondary education. Test 2]. Cito.

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

De Jong, P. F., De Bree, E. H., Henneman, K., Kleijnen, R., Loykens, E. H. M., Rolak, M., Struiksma, A. J. C, Verhoeven, L., & Wijnen, F. N. K. (2016). *Dyslexie: Diagnostiek en behandeling. Brochure van de Stichting Dyslexie Nederland* [Dyslexia: Diagnostics and treatment. Brochure of the Foundation Dyslexia the Netherlands.] Retrieved from http://www.stichtingdyslexienederland.nl/publicaties/brochures-sdn

Dempster, A. P., Laird, N. M., & Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society: Series B (Methodological), 39*, 1-22. doi:10.1111/j.2517-6161.1977.tb01600.x

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

Deno, S. L., & Marston, D. (1987). *Standard reading passages: Measures for screening and progress monitoring.* Children's Educational Services.

DiFino, S. M., & Lombardino, L. J. (2004). Language learning disabilities: The ultimate foreign language challenge. *Foreign Language Annals, 37,* 390-400. doi:10.1111/j.1944-9720.2004.tb02697.x

Domínguez de Ramírez, R., & Shapiro, E. (2006). Cross-language relationship between Spanish and English oral reading fluency among Spanish-speaking English Language Learners in bilingual education classrooms. *Psychology in the Schools, 44,* 795-806. doi:10.1002/pits.20266

Domínguez de Ramírez, R., & Shapiro, E. (2007). Curriculum-based measurement and the evaluation of reading skills of Spanish-speaking English Language Learners in bilingual education classrooms. *School Psychology Review, 35,* 356-369. doi:10.1080/02796015.2006.12087972

Dutch Education Council. (2008). *Vreemde talen in het onderwijs* [Foreign languages in education, Advisory report]. Retrieved from https://onderwijsraad.archief-web.eu/?timestamp=20190923033712&url=https%3A%2F%2Fwww.onderwijsraad.nl%2Fpublicaties%2Farchief%2Fitem21#archive

Dutch Inspectorate of Education. (2018). *Rapport De Staat van het Onderwijs 2018: Onderwijsverslag over 2016/2017* [Report: The State of Education 2018: Education report about 2016/2017, Report]. Retrieved from https://www.onderwijsinspectie.nl/documenten/rapporten/2018/04/11/rapport-de-staat-van-het-onderwijs

Dutch Inspectorate of Education. (2019). *Dyslexieverklaringen. Verschillen tussen scholen nader bekeken* [Dyslexia statements. A closer look at differences between schools, Report]. Retrieved from https://www.onderwijsinspectie.nl/documenten/themarapporten/2019/04/10/dyslexieverklaringen-verschillen-tussen-scholen-nader-bekeken

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). Springer. doi:10.1007/978-1-4939-2803-3_13

Eurostat. (2017, February 23). *Foreign language learning: 60% of lower secondary level pupils studied more than one foreign language in 2015* [Press release]. Retrieved from http://ec.europa.eu/eurostat/web/products-press-releases/-/3-23022017-AP

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

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

Ganschow, L., Sparks, R. L., & Javorsky, J. (1998). Foreign language learning difficulties: An historical perspective. *Journal of Learning Disabilities, 31,* 248-258. doi:10.1177/002221949803100304

Graham, J. W., Hofer, S. M., & MacKinnon, D. P. (1996). Maximizing the usefulness of data obtained with planned missing value patterns: An application of maximum likelihood procedures. *Multivariate Behavioral Research, 31*(2), 197-218. doi:10.1207/s15327906mbr3102_3

Grove, A. (2020, September 30). *Foreign language requirement for college admission.* Retrieved from https://www.thoughtco.com/foreign-language-requirement-college-admissions-788842

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,* 47-76. doi:10.1177/026553228900600106

Little, R. J. (1988). A test of missing completely at random for multivariate data with missing values. *Journal of the American Statistical Association*, *83*, 1198-1202. doi:10.2307/2290157

Little, T. D., & Rhemtulla, M. (2013). Planned missing data designs for developmental researchers. *Child Development Perspectives, 7,* 199-204. doi:10.1111/cdep.12043

Looney, D., & Lusin, N. (2018, February). Enrollments in languages other than English in United States Institutions of higher education, Summer 2016 and Fall 2016: Preliminary report. *Modern Language Association of America Web Publication.* Retrieved from https://www.mla.org/content/download/ 83540/2197676/2016-Enrollments-Short-Report.pdf.

Luke, D. (2004). *Multilevel modeling.* Sage.

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

Porter, D. (1976). Modified cloze procedure: A more valid reading comprehension test. *English Language Teaching Journal, 30*(2), 151-155. doi:10.2307/329673

Rippe, R.C.A., & Merkelbach, I. (2019). *Planned missing data in early literacy interventions: A replication study with an additional gold standard.* Manuscript submitted for publication.

Rosseel, Y. (2012). Lavaan: An R package for structural equation modeling. *Journal of Statistical Software, 48*(2), 1-36. doi:10.18637/jss.v048.i02

Sandberg, K. L., & Reschly, A. L. (2011). English learners: Challenges in assessment and the promise of curriculum-based measurement. *Remedial and Special Education*, *32*, 144-154. doi:10.1177/0741932510361260

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*, 42-57. Retrieved from http://www.internationaljournalofspecialed.com

Sparks, R. L. (2006). Is there a "disability" for learning a foreign language? *Journal of Learning Disabilities, 39,* 544-557. doi:10.1177/00222194060390060601

Sparks, R. L. (2008). Evidence-based accommodation decision making at the postsecondary level: Review of the evidence for foreign language learning. *Learning Disabilities Research & Practice, 23,* 180-183. doi:10.111/j.1540-5826.2008.00276.x

Sparks, R. L. (2009). If you don't know where you're going, you'll wind up somewhere else: The case of "foreign language learning disability." *Foreign Language Annals, 42,* 7-26. doi:10.111/j.1944-9720.2009.01005.x

Sparks, R. L. (2016). Myths about foreign language learning and learning disabilities. *Foreign Language Annals*, *49*, 252-270. doi:10.1111/flan.12196

Sparks, R. L., Patton, J., Ganschow, L., Humbach, N., & Javorsky, J. (2006). Native language predictors of foreign language proficiency and foreign language aptitude. *Annals of Dyslexia, 56,* 129-160. doi:10.1007/s11881-0006-2

Sparks, R. L., Schneider, E., & Ganschow, L. (2002). Teaching foreign (second) language to at-risk learners. In J. A. Hammadou-Sullivan (Ed.), *Literacy and the second language learner* (pp. 55-84). Retrieved from http://www.infoagepub.com/

Van Til, A., & Van Boxtel, H. (2015). *Wetenschappelijke verantwoording Toets 0 t/m 3, tweede generatie* [Scientific justification Test 0 to 3, second generation]. Cito. Retrieved from http://www.cito.nl/

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

Wallace, C. (2007). Vocabulary: The key to teaching English language learners to read. *Reading Improvement*, *44*, 189-194. Retrieved from http://www.projectinnovation.biz/reading_improvement

Wight, M.C.S. (2015). Students with learning disabilities in the foreign language learning environment and the practice of exemption. *Foreign Language Annals, 48,* 39-55. doi:10.1111/flan.12122