

Curriculum-Based Measurement in the Content Areas: Validity of Vocabulary-Matching as an Indicator of Performance in Social Studies

Christine A. Espin, Todd W. Busch, and Jongho Shin

University of Minnesota

Ron Kruschwitz

North St. Paul, Oakdale, Maplewood School District

In this study, we examined the reliability and validity of two curriculum-based measures as indicators of performance in a content-area classroom. Participants were 58 students in a 7th-grade social studies class. CBM measures were student- and administrator-read vocabulary-matching probes. Criterion measures were knowledge pre- and posttests, the social studies subtest of the Iowa Test of Basic Skills, and student grades. Results revealed moderate alternate-form reliability for both vocabulary-matching measures. Reliability of the measures was increased by combining scores across two testing sessions. Correlations between the predictor and criterion variables were moderate to moderately strong, with the exception of those between vocabulary-matching and student grades. Observed scores for students with LD were lower than for students without LD on both student- and administrator-read vocabulary-matching measures. Few differences in reliability and validity coefficients were found between the student- and administrator-read measures. Results are discussed in terms of the use of CBM as a system for monitoring performance and designing interventions for students with learning disabilities in content-area classrooms.

SCHOOL FAILURE AND SCHOOL DROPOUT FOR STUDENTS WITH LEARNING DISABILITIES

Students with learning disabilities (LD) at the secondary-school level are faced with a considerable challenge. They must move beyond the basic learning of reading and writing skills to implement their reading and writing skills to amass content-area knowledge (Alley & Deshler, 1979). For many students with LD this challenge proves to be overwhelming, as reflected by their high failure rates (deBettencourt, Zigmond, & Thornton, 1989; Wagner, 1990). These failure rates are especially noticeable in the general education setting (Wagner, 1990) where students with LD spend more than half of their school day with most of their classes in core academic areas (Lovitt, Plavins, & Cushing, 1999; Wagner, 1990).

Concern over the failure rates of students with LD is intensified when the relation between school failure and school dropout is considered. In a comprehensive study of secondary-school students with LD, Wagner (1990) found that the probability of dropping out of school for a student with a failing grade was 16% compared to only 4% for a student without a failing grade.¹ Similarly, deBettencourt

et al. (1989) found that one-third of students with LD who had repeated a grade prior to their 10th-grade year dropped out of school.

The relation between school failure and school dropout is often viewed within the framework of student alienation from the school community. As factors such as school failure, absenteeism, and discipline problems increase, students' social bonding with the school decreases, eventually leading to a permanent disassociation from the school environment (Miller, Leinhardt, & Zigmond, 1988; Newmann, 1981; Sinclair, Christenson, Evelo, & Hurley, 1998; Wehlage, 1983).

Efforts designed to prevent dropout often focus on reducing students' sense of alienation from and increasing their sense of belonging to the school community. Such dropout prevention efforts have been successful in terms of keeping students in school, but they may not go far enough. For example, in an evaluation of a dropout prevention program called "Check and Connect," Sinclair et al. (1998) reported that students participating in the program had higher rates of school enrollment, better attendance, and earned more credits toward graduation than nonparticipating students. However, the performance of the dropout prevention participants remained substandard. A large percentage of the students were not on track to graduate in four years, their rate of assignment completion was relatively low, and teacher ratings of their academic competence placed them below the 20th percentile of a normative sample. Similarly, Miller et al. (1988) conducted a case study of a school known for its low dropout

Requests for reprints should be sent to Christine A. Espin, University of Minnesota, 250B Burton Hall, 178 Pillsbury Dr. SE, Minneapolis, MN 55455.

¹This percentage reflects probabilities after other factors, including demographics and school characteristics, are held constant.

rate. Results revealed that, although accommodations made for students in the school may have contributed to the low dropout rate, the accommodations also may have decreased academic standards and, consequently, student learning.

The Sinclair et al. (1998) and Miller et al. (1988) studies illustrate the importance of focusing not just on keeping students in school, but also on ensuring that the students are learning while they are in school. Yet tracking student learning, especially in general education classes where students with LD experience their greatest difficulty, is not as simple as it first appears.

ASSESSING LEARNING FOR SECONDARY-SCHOOL STUDENTS WITH LD

Perhaps the most common method for tracking student learning at the secondary-school level is the use of teacher grades. Reliance on grades is especially important for special education teachers who often gauge a student's success in general education based on the grade received by the student. However, reliance on grades is problematic for three reasons. First, it is not clear to what extent grades are valid indicators of student learning (see, e.g., Olson, 1989). Grades often represent a multitude of factors other than learning, including homework completion, class participation, attendance, and behavior (see, e.g., Miller et al., 1988), and the extent to which each factor is weighted in a grading system varies from teacher to teacher and from school to school. Thus, it is possible for a student to receive a passing grade while having learned very little. Second, modifications of classroom requirements, and even of the grading system itself, may cloud the meaning assigned to grades (Rojewski, Pollard, & Meers, 1991). For example, a grade of "C" assigned to two students in a class may represent different levels of proficiency based on the type and level of accommodations given to each student. Third, grades are summative rather than formative in nature. Grades are assigned at the end of a chapter or unit, at which time it is too late for the teacher to intervene if the student has not learned the material.

Another common method for assessing student learning at the secondary-school level is the use of high-stakes testing. Although there are general issues related to the use of high-stakes testing (see, e.g., Haertel, 1999; Linn, 2000; McDonnell, McLaughlin, & Morison, 1997; Thurlow & Johnson, 2000), specific concerns arise when considering the use of the tests for students with disabilities. One of these concerns relates to the questionable technical characteristics of the tests for students with disabilities. High-stakes tests often are designed to assess performance at high, "world-class" levels (Linn, 2000). Thus, although the tests may adequately identify students who do and do not meet predetermined standards, the tests are unlikely to provide reliable information about the progress or performance of students at the lowest end of the achievement continuum (Linn, 2000; McDonnell et al., 1997). In addition, the reliability of the tests for students with disabilities is likely to be affected by the small samples of students in each disability category. The fewer the students included in the sample, the higher the sampling error, the more unstable or unreliable the results of the test

(McDonnell et al., 1997). Finally, the effects of various accommodations on the validity of test scores for students with disabilities often are unknown (McDonnell et al., 1997).

In addition to the technical difficulties associated with the use of high-stakes testing, the information provided by the tests does not aid in individual decision making. Because the tests usually are given only once a year, they do not provide ongoing feedback to teachers about what students are learning during the school year. Moreover, as asserted by McDonnell et al. (1997), many high-stakes tests are not designed to be used for the type of individual decision making that occurs in special education:

The assessments at the core of standards-based reform have various functions, but, in one way or another, they are most often used to determine whether groups of students have reached acceptable levels of educational achievement, as embodied in explicit performance standards. They are generally not used for many of the traditional purposes governing assessments of students with disabilities. For example, the new assessments are generally not used for making decisions about individual students, including diagnosing disabilities, monitoring short-term progress toward IEP goals, or making placement decisions. (p. 166)

In conclusion, the tools that secondary-school special education teachers have at their disposal for assessing student learning are inadequate. Both grades and high-stakes tests fall short of providing teachers with direct and ongoing information related to the performance and progress of and the effects of instruction on an individual student. If teachers wish to obtain such individualized information, they must have access to an alternative assessment system. One such system is Curriculum-Based Measurement (CBM).

USE OF CURRICULUM-BASED MEASUREMENT TO ASSESS STUDENT LEARNING

CBM is a systematic procedure for data collection and decision making in special education (Deno, 1985). When using CBM, teachers probe students' performance on a weekly or biweekly basis and graph the results. The graphs present a picture of students' progress. Teachers inspect the graphs to determine whether student performance is improving. If performance is not improving, teachers change their instruction. Unlike grades and high-stakes tests, CBM procedures are designed to promote implementation of interventions before students begin to fail. Moreover, CBM data reflect student learning independent of work and behavior habits and are designed to be used for decision making at an individual level.

The majority of work in CBM has been conducted at the elementary-school level in the areas of reading, written expression, and spelling (see Shinn, 1989). Recently, CBM has been extended to the secondary-school level (Espin & Tindal, 1998). In the area of content learning, the development of CBM performance measures has taken two different approaches: (1) identification of the critical skills needed to

understand and use content-area information (e.g., Tindal & Nolet, 1995), and (2) identification of generalized indicators of performance used to measure growth over time (e.g., Espin & Foegen, 1996). In the present study, we focused on the second approach to CBM—establishing generalized indicators of performance in content-area learning.

DEVELOPMENT OF GENERALIZED INDICATORS IN THE CONTENT AREAS

The first step in developing generalized indicators of performance in CBM is to identify simple and direct measures that are easy to administer, time efficient, can be repeatedly administered, and are reliable and valid with respect to the general academic area of interest. For example, in the area of reading, the number of words read correctly from text in one minute is used as an indicator of students' general reading proficiency. This measure is used because it correlates highly with other measures of reading proficiency (Deno, Mirkin, & Chiang, 1982; Fuchs, Fuchs, & Maxwell, 1988), is easy to administer, time efficient, and parallel forms can be administered on a repeated basis (Deno, 1985).

The initial work conducted on establishing generalized CBM measures for secondary-school students in the content areas focused on three potential indicators of student performance: reading aloud from text, maze completion,² and vocabulary matching. In a study of 10th-grade students, Espin and Deno (1993) examined the relation between reading aloud from text and student performance on a content-area study task. Their results revealed low-moderate correlations between the two measures in both English ($r = 0.37$) and science ($r = 0.37$). In a follow-up analysis, Espin and Deno (1995) found vocabulary-matching to be slightly better than reading aloud as an indicator of study-task performance. Correlations between the vocabulary-matching and the study tasks ranged from $r = 0.40$ to 0.44 . Further, in a regression analysis, the vocabulary-matching task accounted for a larger proportion of the variance in the study task than did reading aloud from text.

The Espin and Deno studies were limited by the nature of the criterion measure used. The study task required students to read through a text to answer questions. This task reflected only a small portion of what students are required to do in their content-area classes. To address this limitation, Espin and Foegen (1996) conducted a study with middle-school students in which they examined the validity of CBM measures with respect to three different criterion measures: comprehension, acquisition, and retention of text information. The CBM measures examined in the study were reading aloud from text, vocabulary-matching, and maze selection. Results revealed moderately-strong correlations between the CBM and criterion measures, with correlations ranging from $r = 0.52$ to 0.65 . Similar to the Espin and Deno (1995) results, a regression analysis revealed that the vocabulary-matching

task accounted for the largest proportion of variance in the criterion measures.

The Espin and Deno (1993; 1995) and Espin and Foegen (1996) studies provided moderate support for the use of CBM measures as indicators of content-area performance, and revealed a slight advantage for vocabulary-matching over reading aloud and maze selection as an indicator of performance. A limitation of the studies was that they were not conducted within the context of an actual content-area classroom; rather, the investigations made use of researcher-made materials and learning tasks. As a consequence, "learning" in these studies was examined only as it applied to a few selected text passages, not as it applied to learning within a classroom. If CBM measures are to be used in content-area classrooms to monitor student performance and to make instructional decisions, the validity of the measures within an actual classroom must be examined.

PURPOSE OF THE STUDY

The purpose of our study was to examine the validity and reliability of CBM measures within a social studies classroom. Because of the support found for it in previous research, we focused on the vocabulary-matching measure. We extended the previous research by making use of classroom materials to construct the CBM and criterion measures, and by examining the technical adequacy of the CBM measures with respect to student performance within the classroom. In addition, we examined the effects of reading on the validity and reliability of vocabulary-matching measures. In previous research, students have read the terms and definitions for the vocabulary-matching probe. A measure that does not require reading may better reflect students' content-area knowledge and thus be a better indicator of content-area performance. Thus, in this study, we compared two different types of vocabulary-matching measures: a student-read form where students read the words and definitions themselves, and an administrator-read form where the words and definitions were read to the students.

Four research questions were addressed in the study:

1. What is the alternate-form reliability of student- and administrator-read vocabulary-matching measures?
2. Does the alternate-form reliability differ for the two types of measures?
3. What is the criterion-related validity of student- and administrator-read vocabulary-matching measures?
4. Does the criterion-related validity differ for the two types of measures?

The first step in the development of measures targeted at children with disabilities is to demonstrate technical adequacy with the general population. Thus, our study included participants across a range of student performance levels, allowing us to examine the extent to which the measures reflected a rank-ordering of student performance. Similarity in the rank-ordering of students on the CBM and criterion measures would lend support to the validity of the measures. Once criterion-related validity is established, the use

²A maze task consists of a passage where every seventh word is deleted and replaced with a multiple choice item consisting of three choices. One choice is correct and the other two are distracters.

of the measures for reflecting growth over time and the influence of implementation on achievement for students with and without disabilities can be examined.

This study was a part of a longitudinal study in which we tracked student performance over the period of a school year with the two vocabulary-matching measures. In this paper, we present only data related to the reliability and criterion-related validity of the two vocabulary-matching measures.³

METHOD

Participants

Participants in the study were 58 7th-grade students (32 male and 26 female) from a suburban middle school located in a large metropolitan area. Twenty-eight percent of the students received free or reduced lunches. Participants were from two 7th-grade social studies classes taught by a teacher with 30 years of teaching experience. Fifty-five of the students were Caucasian, one was African-American, and two were Asian-American. All students spoke English as their first language. The mean age for the participants was 13.6 years. Data from one participant were dropped because he relocated part way through the study.

Five of the students in the sample were receiving services in special education for learning disabilities. The state definition for LD required an ability-achievement discrepancy, evidence of an information-processing deficit, and a history of underachievement. All five students were receiving services in reading and written expression. One student was also receiving service in mathematics. Mean percentile scores for the students with LD on the Iowa Test of Basic Skills (ITBS; Hoover, Hieronymus, Frisbie, & Dunbar, 1992), form K, Level 12 were 30.4 (range 3 to 46) on the Reading Vocabulary subtest, 26.8 (range of 2 to 49) on the Reading Total subtest, and 13.2 (range 1 to 20) on the Social Studies subtest.

Predictor and Criterion Variables

Predictor Variables

Predictor variables were two vocabulary-matching measures. Probes were developed from terms selected from the classroom textbook (*Introduction to the Social Sciences*; Bostingl, 1991), teacher notes, and teacher lectures. Forty-nine terms and definitions were collected from each of three subject areas (sociology, psychology, and geography), resulting in a total of 147 terms used to generate the probes. Terms were randomly selected with replacement to form 22 probes, each with 22 terms and 20 definitions. The two additional terms on each probe served as distracters. Definitions were taken from the textbook or from the teacher's materials, but were

TABLE 1
Sample Items from a Student-Read Vocabulary Probe

5 Classical conditioning	1. A resource that is important for a country's industries or national security
2 Empathy	2. Feeling as another person does; putting yourself in someone else's place
4 First World	3. According to Freud, a part of the personality concerned with right and wrong
6 Socialization	4. The wealthy, industrialized nations
1 Strategic resource	5. Pairing of a neutral stimulus with an unconditioned stimulus so they produce the same response
3 Superego	6. Process through which an individual learns the rules of society

modified if necessary so that each definition consisted of 15 or fewer words.

Two types of vocabulary-matching probes were developed for the study: student-read and administrator-read probes. For the student-read probes, vocabulary terms were placed vertically in alphabetical order on the left side of the page, and definitions were placed in random order on the right side of the page. Students were given five minutes to read the terms and match each with its definition. For the administrator-read probes, vocabulary terms were listed in alphabetical order vertically on the left side of the page, but no definitions were provided on the probe. The test administrator first read the entire list of terms to the students, then read the definitions one at a time at 15-second intervals. Students marked the term corresponding to each definition. Total administration time was five minutes. (See Table 1 for a sample of items selected from a student-read probe.)

Criterion Variables

The criterion variables in the study included a knowledge test, students' grades in the social studies class, and student performance on the social studies subtest of the ITBS, a standardized achievement test.

The knowledge test was created from the social studies curriculum used by the teacher. The test consisted of 36 questions addressing the three areas covered during the course of the study: sociology, psychology, and geography. Questions were generated from each subject area using the classroom textbook and teacher-made worksheets and tests. Questions were in a multiple-choice format with three distracters and one correct answer. Two types of questions were included on the knowledge test: factual and applied. Factual questions were those that required students to identify a term when the definition was given or to make simple one-to-one relations between names, terms, events, and so forth. Applied questions were those that required the students to apply knowledge to correctly answer a question. Samples of each type of question are presented in Table 2.

To ensure the accuracy of the question classifications, a graduate student not involved in the study and the social studies teacher from the class in which the study took place independently read and classified each question as

³As a part of the longitudinal study, we collected data to examine the validity and reliability of the growth rates produced by repeated measures on parallel forms of the vocabulary probes. These results are not reported here.

TABLE 2
Examples of Factual and Applied Questions from Knowledge
Pre- and Posttests

<i>Question Type</i>	<i>Examples</i>
Factual	The process by which a member learns the rules of his/her group is called a. socialization b. community c. role play d. mobility
	The Earth's halfway point from the North Pole to the South Pole is called the a. Prime Meridian b. equator c. rotation d. globe
Applied	Jose comes from a working class home. He marries Juanita who is wealthy and he moves into an upper class neighborhood. Jose's change in status is an example of a. mobility b. sanctions c. mores d. primary group
	England is a wealthy country. It has thousands of factories and industries. Many people own their own homes and businesses. England is an example of a a. First-World Nation b. Second-World Nation c. Third-World Nation d. developing nation

factual or applied. Percent agreement was calculated by dividing agreements by agreements plus disagreements. The percent agreement between the researcher's and the independent rater's classifications was 92%. Disagreements were discussed and revisions made to questions where necessary.

Twelve questions (three factual and nine applied) were generated for each subject area. An emphasis was placed on applied questions to ensure that the relation between the vocabulary probes and the knowledge test was not solely a function of the similarities between vocabulary matching and factual questions. Each content area was represented on the test in a block of three questions to ensure that students would have exposure to questions from all subject areas even if they did not finish the test. The order of the questions and the inclusion of factual and applied questions within each block was randomly determined. Reliability of the knowledge test as determined by Cronbach's alpha was 0.74.

Students' grades were calculated on a 13-point scale, with 13 given for an "A+," 12 for an "A," 11 for an "A-," 10 for a "B+," and so forth. A score of 0 was assigned for a grade of "F." Students in the school who did not pass a class were required to make up the class in a four-week after-school session. Upon completion of this session, the students received a passing grade of "P" for the class. A grade of "P" was assigned a value of 1. Course grades were based

on the following factors: 25% homework, 25% quizzes, 25% unit tests, and 25% current events (students were required to read, summarize, and give their opinions on newspaper articles).

Students completed the ITBS in the spring of the year prior to the beginning of the study. Participants took Form K, Level 12 of the test. The test consisted of 42 questions, covering history, geography, economics, political science, sociology/anthropology, and related social sciences (e.g., ethics, human values, etc.).

The content validity of the ITBS was established through the use of curriculum guides, textbooks, and research to generate item groups (Salvia & Ysseldyke, 1998). The items were reviewed for content fit and item bias by experts in the field and selected items were then tested on 100,000 students from 30 different states plus Guam and the Virgin Islands. The final sample of items was selected on the basis of the performance of the sample group (Salvia & Ysseldyke, 1998). Internal consistency of the ITBS, as reported in the technical manual, ranged from 0.61 to 0.93 for the various levels of the test.

Procedure

Knowledge tests were administered by a graduate student during a 46-minute class session at the beginning and end of the study. All students completed the pre- and post-knowledge tests in the allotted times.

Vocabulary-matching probes were administered by a graduate student once per week over the course of five months from January to May. With three exceptions (due to a field trip and school holidays) data were collected on consecutive weeks. During each data collection session, students completed both vocabulary-matching probes. The order in which the two probes were administered was alternated each week. Standardized instructions were read to the students before administration of each probe.

Students' grades were obtained from the teacher's grade book. The study took place over a period of three grading periods. The mean grade across the three periods was used for all analyses. Scores on the ITBS were obtained from students' school records. Standard scores were used for all analyses.

The knowledge pre- and posttests and the vocabulary probes were scored by the graduate student who administered the tests. For the vocabulary-matching probes, the number of correct matches was scored. For the knowledge pre- and posttests, the number of correct answers was scored. Accuracy of scoring was checked by an independent scorer who scored 20 each (approximately one-third) of the knowledge pre- and posttests, and the administrator- and student-read vocabulary probes. Percent accuracy was calculated by dividing agreements by agreements plus disagreements. Accuracy rates were 95% for the knowledge pretests, 90% for the knowledge posttests, 96% for the vocabulary-matching student-read probes, and 96% for the vocabulary-matching administrator-read probes. All disagreements were discussed and corrections made where necessary.

TABLE 3
Means and Standard Deviations for Administrator- and Student-Read Vocabulary-Matching Measures

	<i>n</i>	<i>Administrator-read</i>		<i>Student-read</i>	
		<i>M</i>	<i>SD</i>	<i>M</i>	<i>SD</i>
Probe 1	53	8.41	3.83	5.23	3.41
Probe 2	54	8.75	4.33	6.02	3.48
Probe 3	53	6.25	3.61	7.11	5.35
Probe 4	44	10.80	4.52	6.00	4.09
Probe 5	48	8.62	4.23	7.04	4.13
Probe 6	53	10.64	4.86	9.83	5.24
Probe 7	49	8.62	4.87	9.35	4.99
Probe 8	50	8.88	4.51	8.84	5.26
Probe 9	50	10.24	4.41	9.78	5.14
Probe 10	52	9.47	3.80	14.10	5.38
Probe 11	51	10.88	4.20	9.71	5.67

RESULTS

Our first and second research questions addressed the alternate-form reliability of the vocabulary-matching measures and the differences in reliability between the measures. Alternate-form reliability was calculated by examining correlations between adjacent probes, that is, week 1 with week 2, week 2 with week 3, and so forth. Means and standard deviations for student- and administrator-read vocabulary-matching probes are reported in Table 3. Alternate-form reliability coefficients for adjacent probes and the mean across all adjacent probes are reported in Table 4.

For the administrator-read probes, alternate-form reliabilities ranged from $r = 0.58$ to 0.87 , and for the student-read probes, from $r = 0.63$ to 0.81 . Mean reliability coefficients for both types of probe was $r = 0.70$, revealing no differences between the administrator- and student-read vocabulary measures ($z = 0$).

We found the alternate-form reliability of both types of vocabulary measures to be somewhat low. We were interested in determining whether the stability of the measures would increase if scores were combined across adjacent probes (i.e.,

TABLE 4
Alternate-Form Reliability for Administrator- and Student-Read Vocabulary-Matching Measures: Single Measures

<i>Adjacent measures</i>	<i>Administrator-read</i>	<i>Student-read</i>
1 with 2	0.58	0.67
2 with 3	0.65	0.63
3 with 4	0.68	0.71
4 with 5	0.71	0.64
5 with 6	0.73	0.66
6 with 7	0.75	0.80
7 with 8	0.87	0.81
8 with 9	0.77	0.77
9 with 10	0.66	0.71
10 with 11	0.64	0.60
Mean	0.70	0.70

Note. All coefficients statistically significant at $p < 0.01$ level.

TABLE 5
Alternate-Form Reliability for Administrator- and Student-Read Vocabulary-Matching Measures: Combined Measures

<i>Adjacent pairs of measures</i>	<i>Administrator-read</i>	<i>Student-read</i>
1/2 with 3/4	0.70	0.76
3/4 with 5/6	0.73	0.83
5/6 with 7/8	0.85	0.87
7/8 with 9/10	0.83	0.88
Mean	0.78	0.84

Note. All correlations statistically significant at $p < 0.01$.

1 with 2, 3 with 4, etc.). As shown in Table 5, combining scores across adjacent probes resulted in higher alternate-form reliability. For the administrator-read probes, reliability coefficients ranged from 0.70 to 0.85, with a mean of 0.78. For the student-read probes, coefficients ranged from 0.76 to 0.88, with a mean of 0.84. Although the obtained coefficients for the student-read probes were larger than for the administrator-read probes, these differences were not statistically significantly ($z = .87$).

Our third research question addressed the criterion-related validity of the vocabulary-matching measures with respect to other measures of general proficiency in the content areas, including the knowledge pretest and posttest, performance on a standardized achievement test in the area of social studies, and students' quarter grades in the social studies class. In addition, we conducted preliminary analyses to examine differences in performance between students with and without LD.

To obtain a more stable estimate of student performance on the vocabulary-matching probes, mean scores on the first three probes (referred to as "pretest probes") and the last three probes (referred to as "posttest probes") were calculated. Three probes were used to ensure that each student had a minimum of two probes contributing to his or her score. (Students may have been absent for one of the three probes.)

TABLE 6
Means and Standard Deviations for Vocabulary-Matching and Criterion Measures

	<i>M</i>	<i>SD</i>
Vocabulary-matching measures		
Student-read		
Pretest (probes 1-3)	6.10	3.73
Posttest (probes 9-11)	11.35	5.08
Administrator-read		
Pretest (probes 1-3)	7.72	3.30
Posttest (probes 9-11)	10.27	3.70
Knowledge pretest ^a	20.27	5.07
Knowledge posttest ^a	24.86	5.62
Iowa Test of Basic Skills ^b	218.72	32.27
Mean quarter grades ^c	9.38	3.24

^aNumber correct out of a possible 36.

^bScore reported is a standard score.

^cGrades are calculated on a scale of 0 to 13; 0 = F, 1 = P, 2 = D-, 3 = D, 4 = D+, 5 = C-, 6 = C, 7 = C+, 8 = B-, 9 = B, 10 = B+, 11 = A-, 12 = A, 13 = A+.

TABLE 7
Correlations Between Administrator- and Student-Read
Vocabulary-Matching Measures and Criterion Measures

	<i>Vocabulary-matching measures</i>			
	<i>Pretest</i>		<i>Posttest</i>	
	<i>Student-read</i>	<i>Administrator-read</i>	<i>Student-read</i>	<i>Administrator-read</i>
Knowledge pretest	0.60**	0.59**	0.66**	0.73**
Knowledge posttest	0.66**	0.67**	0.81**	0.84**
Iowa Test of Basic Skills-Social Studies subtest	0.56**	0.63**	0.64**	0.76**
Quarter grades	0.28*	0.27*	0.51**	0.34*

*Correlations significant at $p < 0.05$.

**Correlations significant at $p < 0.01$.

Means and standard deviations for the combined vocabulary-matching probes and the criterion measures are reported in Table 6. Correlations between the predictor and criterion variables are reported in Table 7.

Results revealed moderately strong to strong correlations between the vocabulary-matching probes and the criterion variables, with the exception of the correlations between vocabulary-matching and students' quarter grades. Correlations between the vocabulary-matching measures and the knowledge pre- and posttest ranged from $r = 0.59$ to 0.84 , and between the vocabulary-matching measures and the ITBS from $r = 0.56$ to 0.76 .

Correlations between the vocabulary-matching and knowledge *posttests* ($r = 0.81$ and 0.84) were significantly stronger than between the vocabulary-matching and knowledge *pretests* ($r = 0.60$ and 0.59 ; $z = 3.23$, $p < 0.05$ and 3.65 , $p < 0.05$ for student- and administrator-read probes respectively). Correlations between the vocabulary-matching *pretests* and the knowledge *posttests* were moderately strong ($r = 0.66$ and 0.67 for student- and administrator-read probes respectively), indicating that the vocabulary-matching probes were fairly accurate in predicting students' future performance.

Coefficients between the vocabulary-matching probes and the social studies subtest of the ITBS were moderately strong and stable ($r = 0.56$ to 0.76) across both pre- and posttest vocabulary-matching probes. However, the relation between the vocabulary measures and quarter grades was only low to moderate ($r = 0.27$ to 0.51).

Our fourth research question addressed differences in validity between administrator- and student-read vocabulary probes. We did not include student grades in this analysis due to the low correlations between the vocabulary measures and grades. Results of our analyses revealed no significant differences in the correlations between the two types of vocabulary-matching probes and the knowledge tests ($t(49) = 0.14$ and $t(47) = 0.74$ for pretests and posttests respectively). Similarly, no differences in correlations between

TABLE 8
Comparison of Students With and Without LD on
Vocabulary-Matching Measures

	<i>Vocabulary-matching measures</i>							
	<i>Student-read measures</i>				<i>Administrator-read measures</i>			
	<i>Pretest</i>		<i>Posttest</i>		<i>Pretest</i>		<i>Posttest</i>	
	<i>M</i>	<i>SD</i>	<i>M</i>	<i>SD</i>	<i>M</i>	<i>SD</i>	<i>M</i>	<i>SD</i>
LD	2.0	1.3	6.1	5.4	4.0	2.9	5.5	3.3
No LD	6.5	3.7	11.9	4.8	8.1	3.1	10.7	3.4

the vocabulary-matching pretests and the ITBS were evident ($t(46) = 1.00$). There was a significant difference in the correlations between the vocabulary-matching posttests and the ITBS, with stronger correlations found for the administrator-read probe than for the student-read probe ($t(46) = 2.25$, $p < 0.05$). These results indicate that, with one exception, reading did not exert an influence on the criterion-related validity of the vocabulary-matching measures.

Our primary method for addressing the criterion-related validity of the vocabulary-matching measures was to examine correlations between the vocabulary measures and the knowledge tests. However, additional support for our findings would be garnered if we also found that students with and without LD performed differently on the measures. That is to say, if the vocabulary-matching measures were valid indicators of student performance, we would expect that students with LD would score lower on the measures than students without LD. Our LD sample was too small ($n = 5$) to systematically test the differences and draw conclusions; however, we were able to examine the obtained pattern of differences between the two groups to examine this issue on a preliminary basis.

Means and standard deviations on the student- and administrator-read vocabulary-matching measures for the students with and without LD are presented in Table 8. The students with LD in our sample obtained lower scores than students without LD on both student- and administrator-read probes. This pattern was evident for both pre- and posttests.

DISCUSSION

Results of this study lend support to the reliability and validity of vocabulary-matching measures as indicators of students' performance in the social studies content area. We addressed four questions in the study:

1. What is the alternate-form reliability of student- and administrator-read vocabulary-matching measures?
2. Does the alternate-form reliability differ for the two types of measures?
3. What is the criterion-related validity of student- and administrator-read vocabulary-matching measures?
4. Does the criterion-related validity differ for the two types of measures?

Alternate-Form Reliability

Results of our research with respect to the alternate-form reliability of the measures reveal that the stability of a single vocabulary-matching measure was somewhat low. It was necessary to combine two measures to obtain a more acceptable level of alternate-form reliability. Practically speaking, the need to combine measures implies that teachers must give two probes and calculate a total or mean score before graphing a point. Because this procedure increases the measurement time needed to obtain a single data point, there are a few factors to consider before recommending combining measures.

First, it is possible that reliabilities were lowered somewhat by the week interval that occurred between testing situations. Alternate-form reliability usually is calculated for two alternate forms given at approximately the same time. Because previous studies have not examined alternate-form reliability for the vocabulary-matching measure (e.g., Espin and Foegen (1996) examined only test-retest reliability), it is not possible to estimate the effect of the week interval on the reliability coefficients.

Second, it is possible that the stability of the measures was affected by students' familiarity with or attention to the task. As can be observed in Table 3, reliability coefficients in the middle of the study were somewhat larger than those at the beginning of the study (where students were perhaps becoming familiar with the task) and at the end of the study (where students' attention to the task was perhaps waning). The practical implication of this pattern of results is that it may be possible to increase the stability of the vocabulary-matching measures by having students complete two or three practice probes before beginning data collection. In addition, teachers may need to implement a system to keep students motivated to perform their best on the CBM tasks throughout the school year. For example, teachers may have students graph and inspect their data throughout the year so that students feel a sense of responsibility for their own progress.

Lastly, rather than combining scores across two probes, it may be possible to effect the same increase in reliability by increasing the length of the probe and the length of time given to work on the probe. The advantage of a longer probe is that teachers can obtain a single data point for each testing session. The disadvantage is that a longer probe requires more class time. In addition, a longer probe may prove to be frustrating for students, especially for those who struggle academically. Further research is needed to examine the most practical and "user-friendly" method for obtaining reliable measures across time.

Criterion-Related Validity

Our results with respect to the validity of the vocabulary-matching measures were positive. Validity coefficients between the vocabulary-matching pretests and both knowledge pre- and posttests were moderately strong to strong, as were the correlations with the ITBS. In addition, the students with LD in our sample scored lower than the students without LD on the vocabulary-matching measures. These re-

sults indicate that both of the vocabulary-matching measures were good indicators of students' performance in their social studies classroom, as well as good indicators of students' general social studies knowledge.

Correlations between the vocabulary-matching and knowledge *posttests* were significantly stronger than for the *pretests*, indicating that the vocabulary-matching measures were stronger indicators of students' course knowledge at the end of the year than at the beginning of the year. We suspect that these differences were due in part to an increase in variability from pre- to posttest on the vocabulary-matching measures (see Table 6). The increased variability in students' scores, and subsequent increase in validity coefficients, lends support to the sensitivity of the vocabulary measures as indicators of students' general performance. That is to say, we would expect that as the year progresses, students who learn more will post greater increases on the vocabulary-matching measures than students who learn less. Such differences would serve to increase variability in the vocabulary-matching scores, leading to stronger correlations for the posttests than for the pretests.

Although correlations for the posttests were stronger than for the pretests, it is interesting to note that the correlations between the vocabulary-matching *pretests* and the knowledge *posttests* were still quite respectable ($r = 0.63$ and 0.62 for student- and administrator-read probes respectively), indicating that the vocabulary-matching probes were fairly accurate in predicting students' future performance. These results imply that the vocabulary-matching measures may prove useful as screening measures to determine whether students with LD are likely to succeed and learn in their general education classrooms. Such screening measures could be used to make decisions regarding placement and inclusion of students into general education. In addition, the measures could be used to determine the students who are most likely to need help once placed into general education content-area classrooms.

The correlations between the vocabulary-matching probes and the social studies subtest of the ITBS supported the validity of the vocabulary-matching measure as a general indicator of social studies knowledge. These correlations were moderately strong and stable across both pre- and posttest vocabulary-matching probes.

Our confidence in the validity of the vocabulary-matching measures is tempered somewhat by the unexpectedly low correlations with students' quarter grades. We suspect that the lowered correlations associated with grades may be due in part to the restricted range of the grades (most students obtained average grades of between a C and an A+), and the fact that diverse factors contributed to students' grades, including homework and current events assignments. While we may reflect upon the reasons for the low correlations, the results imply that the vocabulary-matching measures cannot be used by teachers to predict the grades students will receive in their content-area classes. It is possible that the measures might be useful for predicting whether students with LD will pass or fail a content-area class. We did not have enough students with LD in our sample to examine this issue.

Results with respect to differences in the reliability and validity of the two vocabulary-matching measures revealed

that both administrator- and student-read measures were valid and reliable indicators of general social studies performance. A comparison of the strength of the validity coefficients resulted in only one significant difference, indicating that, in general, the requirement to read the probe had little effect on the technical adequacy of the measures.

Although the posttest administrator-read vocabulary-matching probe was found to be significantly better at predicting performance on the ITBS than the student-read probe, we do not believe that this single difference warrants a recommendation for use of the administrator-read measure. First, although only a small difference, somewhat lower reliability coefficients were associated with the administrator-read measure. Second, the student-read measure offers practical advantages. Students can work independently on the student-read probe while the teacher delivers instruction to other students. In addition, the student-read measure presents the possibility of students conducting their own monitoring. Students can complete a probe, score it, graph the results, and then discuss options for interventions with the teacher.

A final observation related to our data concerns our results as they relate to the results of previous studies. The validity coefficients obtained between the vocabulary-matching measures and the criterion variables in our study are similar to those found by Espin and Foegen (1996), but stronger than those found by Espin and Deno (1995). We suspect that the differences might be related to the age of the participants in the studies.⁴ The participants in both our study and the Espin and Foegen study were at the middle-school level, whereas the participants in the Espin and Deno study were at the high-school level. It is possible that the sensitivity of CBM measures decreases as students get older, although it is difficult to determine this across diverse studies that make use of different measures. A more direct examination of the effect of age on the validity and reliability of CBM measures is in order.

Limitations

One of the limitations of our study was that we selected students from the classroom of only one teacher. It is possible that our results were influenced by a close correspondence between the format of the vocabulary-matching measures and the teacher's style of instruction. This would be especially true if the teacher emphasized factual learning. To address this limitation, we conducted informal interviews with and observations of the teacher. We found that the teacher in our study used a variety of teaching methods to instruct his students, including many discovery learning activities. He made use of group-learning activities as well as more traditional lecture-type activities. His primary goal was to

⁴The differences may also be related to the structure of the vocabulary probes. In the Espin and Deno study, the vocabulary probe was a 10-item probe and students had 10 minutes to work. The probe used in the Espin and Foegen study was more similar to the probe in the current study: a 20-item probe where students had five minutes to work. However, given the fact that age differences were also evident in the correlations generated for the reading-aloud measures between the Espin and Deno and Espin and Foegen studies, we think that age may be the significant factor.

teach students the concepts associated with a particular unit of study, not just the facts.

An example may perhaps best serve to illustrate the instructional approach taken by our teacher. During a psychology unit, our teacher had students read articles on the hereditary and environmental influences on behavior. Students also viewed films and heard lectures on the developmental stages of childhood. Following these activities, students were required to generate materials designed to teach a young child a specific skill. The students tried out their materials on children whose parents had volunteered to bring them to school for a day. Following this "real-life" experience, the students discussed whether or not the materials they had developed were successful in teaching the desired skill, and to what extent the innate abilities of the child (i.e., his or her stage of development) influenced student learning versus to environmental factors (i.e., the students' own teaching). Although just one example, this scenario illustrates the type of experiential and conceptual learning emphasized by the teacher in our study.

A second limitation to our study was related to our examination of the vocabulary-matching measures as static measures. Because CBM measures are designed to be used as repeated measures of growth over time, it is important to examine the validity and reliability of the measures as growth measures. A related limitation is that this study did not focus on the use of CBM for ongoing progress monitoring and the effects of such progress monitoring on teacher and student instructional decision making. These are questions for future research.

Conclusion

The results of this study lend support to the use of either a student- or administrator-read vocabulary-matching measure as an indicator of student knowledge in a social studies classroom. Vocabulary-matching measures might prove helpful to teachers in making decisions regarding student placement in general education content-area classrooms. If a student is not learning in a particular content-area classroom as indicated by the vocabulary-matching probes, the teacher could consider placement in a different class or a different setting. Further, teachers could use CBM data to determine the need for interventions for selected students. For example, teachers may be surprised to learn that a student is obtaining a passing grade in class because of work-study habits, but is not learning the material. Such a student could be targeted for special interventions or accommodations. The use of valid and reliable vocabulary-matching measures as progress measures, and the influence of CBM use on teacher decision making and student achievement in the content areas, has yet to be examined.

NOTE

The research reported here was funded in part by the Guy Bond Foundation, University of Minnesota. We wish to thank the teachers, administration, and students of the Maplewood

schools for their participation in the study. We wish to thank Erica Lembke for comments on an earlier version of this paper, and Dana Frederick for assistance in data coding. In addition, we wish to acknowledge the Netherlands Institute for Advanced Study in the Humanities and Social Sciences for its support in the preparation of this manuscript.

REFERENCES

- Alley, G. R., & Deshler, D. D. (1979). *Teaching the learning disabled adolescent: Strategies and methods*. Denver, CO: Love.
- deBettencourt, L. U., Zigmond, N., & Thornton, H. (1989). Follow-up of postsecondary-age rural learning disabled graduates and dropouts. *Exceptional Children, 56*, 40–49.
- Bostingl, J. J. (1991). *Introduction to the social sciences*. Needham Heights, MA: Prentice Hall.
- Deno, S. L. (1985). Curriculum-based measurement: The emerging alternative. *Exceptional Children, 49*, 36–45.
- Deno, S. L., Mirkin, P. K., & Chiang, B. (1982). Identifying valid measures of reading. *Exceptional Children, 49*, 36–45.
- Espin, C. A., & Deno, S. L. (1995). Curriculum-based measures for secondary students: Utility and task specificity of text-based reading and vocabulary measures for predicting performance on content-area tasks. *Diagnostique, 20*, 121–142.
- Espin, C. A., & Deno, S. L. (1993). Performance in reading from content-area text as an indicator of achievement. *Remedial and Special Education, 14*(6), 47–59.
- Espin, C. A., & Foegen, A. (1996). Validity of three general outcome measures for predicting secondary students' performance on content-area tasks. *Exceptional Children, 62*, 497–514.
- Espin, C. A., & Tindal, G. (1998). Curriculum-based measurement for secondary students. In M. Shinn (Ed.), *Advanced applications of curriculum-based measurement*. New York: Guilford.
- Fuchs, L. S., Fuchs, D., & Maxwell, L. (1988). The validity of informal reading comprehension measures. *Remedial and Special Education, 9*, 21–28.
- Haertel, E. H. (1999). Validity arguments for high-stakes testing: In search of evidence. *Educational Measurement: Issues and Practice, 18*, 5–9.
- Hoover, H. D., Hieronymus, A. N., Frisbie, D. A., & Dunbar, S. B. (1992). *Iowa test of basic skills*. Chicago, IL: Riverside Publishing Company.
- Linn, R. L. (2000). Assessments and accountability. *Educational Researcher, 29*, 4–16.
- Lovitt, T. C., Plavins, M., & Cushing, S. (1999). What do pupils with disabilities have to say about their experience in high school? *Remedial and Special Education, 20*, 67–76, 83.
- McDonnell, L., McLaughlin, M., & Morison, P. (Eds.) (1997). *Educating one & all: Students with disabilities and standards-based reform*. Washington, DC: National Academy.
- Miller, S. E., Leinhardt, G., & Zigmond, N. (1988). Influencing engagement through accommodations: An ethnographic study of at-risk students. *American Educational Research Journal, 25*, 465–487.
- Newmann, F. N. (1981). Reducing student alienation in high schools: Implications of theory. *Harvard Educational Review, 51*, 546–564.
- Olson, G. H. (1989). *On the validity of performance grades: The relationship between teacher-assigned grades and standard measures of subject matter acquisition*. ERIC Document Reproduction Service No. ED 307 290.
- Rojewski, J. W., Pollard, R. R., & Meers, G. D. (1991). Grading mainstreamed special needs students: Determining practices and attitudes of secondary vocational educators using a qualitative approach. *Remedial and Special Education, 12*, 7–15, 28.
- Salvia, J., & Ysseldyke, J. (1998). *Assessment* (7th ed.). Boston, MA: Houghton-Mifflin.
- Shinn, M. R. (Ed.). (1989). *Curriculum-based measurement: Assessing special children*. New York: Guilford.
- Sinclair, M. F., Christenson, S. L., Evelo, D. L., & Hurley, C. M. (1998). Dropout prevention for youth with disabilities: Efficacy of a sustained school engagement procedure. *Exceptional Children, 65*, 7–21.
- Thurlow, M. L., & Johnson, D. R. (2000). High-stakes testing of students with disabilities. *Journal of Teacher Education, 51*, 289–298.
- Tindal, G., & Nolet, V. (1995). Curriculum-based measurement in middle and high schools: Critical thinking skills in content areas. *Focus on Exceptional Children, 27*(7), 1–22.
- Wagner, M. (1990). The school programs and school performance of secondary students classified as learning disabled: Findings from the National Longitudinal Transition Study of Special Education Students. Paper presented at the annual meeting of the American Educational Research Association, Boston, MA.
- Wehlage, G. G. (1983). The marginal high school student: Defining the problem and searching for policy. *Children and Youth Services Review, 5*, 321–342.