UNIVERSITY OF CALIFORNIA
RIVERSIDE

Screening for Reading Problems in Middle School: Maze and STAR Reading Assessment

A Thesis submitted in partial satisfaction
of the requirements for the degree of

Master of Arts

in

Education

by

Lisa Christine Kowalko

March 2014

Thesis Committee:
Dr. Michael Vanderwood, Chairperson
Dr. Cathleen Geraghty
Dr. Gregory Palardy

The Thesis of Lisa Christine Kowalko is approved:

_____

_____

_____

Committee Chairperson

University of California, Riverside

ABSTRACT OF THE THESIS

Screening for Reading Problems in Middle School: Maze and STAR Reading Assessment

by

Lisa Christine Kowalko

Master of Arts, Graduate Program in Education
University of California, Riverside, March 2014
Dr. Michael Vanderwood, Chairperson

This study examined the relationship between the curriculum-based measurement Maze task and the Standardized Test for the Assessment of Reading (STAR-R) and California Standards Test (CST) scores. Participants included 1,479 sixth, seventh, and eighth grade students across three public middle schools in an urban school district in southern California. The results indicated a significant correlation between Maze scores in the winter and in the spring, but failed to find a significant correlation between fall Maze screening scores and CST standard scores. Similarly, the second inquiry examined the magnitude of the relationship between STAR-R scores and the scores on the CST. This analysis revealed that STAR-R standard scores are significantly correlated to CST standard scores at all three screening points. Additional analysis showed significant differences in the above relationships between grade levels with a general trend of increased correlation values from Grade 6 to Grade 8 for the majority of screening points. Lastly, a diagnostic accuracy analysis and a receiver operating characteristic curve analysis revealed low to moderate levels of reliability and sensitivity for both measures in predicting CST scaled score classification.

Table of Contents

List of Tables

List of Figures

Reading is a skill that is integral to academic success. The development of reading skills serves as a major foundation for all school based learning. Students who fail to learn to read proficiently will have difficulty comprehending grade-level textbooks and are limited in their opportunities for academic and occupational success (Sum, Khatiwada & McLaughlin, 2009). The middle school years are an especially critical period of reading development. During this time, students acquire the ability to understand a variety of texts, to use sophisticated comprehensions and study strategies, and develop lifelong reading preferences (National Middle School Association, 2001). However, recent statistics provided on the 2009 National Assessment of Education Progress (NAEP) Reading Test indicated that approximately 26 percent of eighth grade students performed below the Basic level of reading proficiency (NCES, 2009). With such high stakes riding on student reading achievement and statistics suggesting a large portion of students continue to experience below basic proficiency levels, it is vital that school systems be able to identify students who are at risk for reading failure. Currently, there are no state regulations specifying which reading screening measures should be used to achieve this goal, leaving individual schools to make this decision for themselves.

One of the most popular reading measures used in middle schools to establish student reading levels is the Curriculum Based Measurement Maze task (Maze) (Shinn & Shinn, 2002). Over the past two decades, researchers have found results supporting the utility of this classic, paper and pencil style exam to assess reading comprehension (Fuchs & Fuchs 1990; 1992; Jenkin & Jewel, 1993). As schools have begun to have increased access to technology, however, there is a growing use of computers to deliver

assessments. Many administrators are choosing these new, computerized reading assessment software packages as universal screeners, because programs are typically able to provide student and group level results in less time and use fewer staff hours than traditional testing methods. The Standardized Test for the Assessment of Reading (STAR-R) is one such computerized reading achievement that has begun to gain prevalence for assessing middle school students.

Although the results of initial research conducted by the publishers of STAR-R have indicated high test-retest reliability and moderate to high predictive validity coefficients (Renaissance Learning, 2001; Renaissance Learning, 2010), currently only one independent peer-reviewed research article has been published exploring its utility. The goal of this paper is to compare the predictive validity and diagnostic accuracy of STAR-R and the Maze task as they relate to scores on the California Standards Test (CST), a high-stakes assessment of student progress towards state reading standards.

AimsWeb Curriculum Based Measurement – Maze Task

In 1985, Deno proposed curriculum-based measurement (CBM) as a set of methods for assessing academic competence in reading. Unlike traditional assessments that measure mastery of skills, CBM indexes overall student proficiency, and the format and difficulty of the tests remain constant over the year. The tests represent reading curriculum that the student should master by the end of the year (Deno, 1985). In addition to being an indicator of overall reading achievement, CBM is sensitive to the effects of small adjustments made in instructional programming so that teachers can use the results to inform instruction (Deno, Mirkin, & Chiang, 1982). When CBM is

administered to an entire school population, normative standards can be developed and used for decision-making about individual students, and district-level analyses can then be used to predict performance on high-stakes assessments (Silberglitt, Burns, Madyun, & Lail, 2006).

The most widely used and validated CBM measure is the oral reading fluency task (ORF). ORF is a read-aloud task where the passages are calibrated for the target level of reading for each grade level. Student performance is measured by having students read a passage aloud for one minute. Words self-corrected within three seconds are scored as accurate. Words omitted or substituted are scored as errors. The number of correct words per minute from the passage is the oral reading fluency score.

Over the last several decades, many studies have focused on the reliability and validity of this measure. Results from reliability studies have found ORF to be reliable, with alternate form reliability coefficients ranging from .80 to .95 (Good & Kaminski, 2003; Espin & Deno, 1994; Fuchs & Deno, 1994). Moreover, results from multiple studies have confirmed ORF to be an efficient predictor of elementary-school students' scores on traditional tests of reading achievement, with correlations generally ranging from .60 to .90 (Fuchs & Deno, 1991; Fuchs, Fuchs, Hosp, & Jenkins, 2001; Fuchs, Fuchs, & Maxwell, 1988; Jenkins, Fuchs, Espin, van den Broek, & Deno, 2001; Good & Jefferson, 1998).

In spite of this solid evidence base, there is also literature that identify threats to the validity and reliability of oral reading fluency measures. Firstly, several research studies have concluded that the reliability and criterion-related validity of CBM ORF

measures decline for older students (Jenkins & Jewell, 1993; Silberglitt et al., 2006; Wayman et al., 2007). Furthermore, in 2002, two research papers reported that ORF lacks face validity with many teachers, who do not believe that oral reading alone could accurately represent all areas related to reading success, especially in the area of comprehension (Shinn, Good, Knutson, Tilly & Collins, 1992; Fuchs, Fuchs, Hamlett, & Ferguson, 1992). This opinion persists among many practitioners despite a strong research base that has found passage reading to be highly correlated with measures of comprehension (Deno, 1985; Deno, Mirkin & Chiang, 1982; Fuchs, Fuchs, & Maxswell, 1988; Hintze, Conte, Shapiro, & Basile, 1997; Shinn, Good, Knutson, Tilly, & Collins, 1992

A similar threat to the validity of this assessment was identified by Hamilton and Shinn in 2003 who described "word callers", or students who purportedly can read fluently but do not comprehend what they read. If this is the case for some students, ORF scores would not correlate to reading achievement and therefore be invalid. The researchers investigated the idea of "word callers" by comparing thirty-three teacher identified "word callers" with their typical peers on oral reading fluency, Maze, and the Passage Comprehension subtest of the Woodcock Reading Mastery Test. Results from a multivariate analysis of variance showed that those students identified as word callers did not score significantly differently from other low scoring students on ORF and comprehension measures (Hamilton & Shinn, 2003).

In 1992, Fuchs and Fuchs proposed the Maze task as part of their CBM assessment tools as a viable alternative for screening and progress monitoring students in

grades three through eight. Maze is a multiple-choice cloze task that students complete while reading silently. The first sentence of the word passage is left intact. Thereafter, every 7th word is replaced with three words inside parenthesis. One of the words is the exact one from the original passage. The score is the number of correct word choices the student selects in 1-3 minutes. The Maze task has many advantages over traditional ORF screening measures, including that it can be administered in a group setting, is adaptable to computer administration, and has more face validity for many teachers because the students must have some level of reading comprehension to make correct Maze selections (Fuchs, Fuchs, Hamlet, & Ferguson, 1992). Overall, the Maze task has been found to be easier to administer, less time consuming to score, and more sensitive to middle school student growth than alternatives (Fuchs & Fuchs, 1992).

Although standardized and popularized by Fuchs and Fuchs (1992) as part of CBM, the process associated with the Maze task is not new to education. The idea of a passage deletion task was introduced by Kingston and Weaver in 1970. The researchers concluded that this task was a better predictor of standard reading test scores than were readiness and basal reader instruments. In 1973, Guthrie explored the validity and reliability of this technique and found that the correlation between the Maze task and the *Gate-MacGintie* comprehension subtest for a group including normal and poor readers was found to be .82 and reliability coefficients for a single maze was found to be greater than .90. Further results from research by Guthrie, Seifert, Burnham, and Caplan (1974) indicated that Maze reliably identified students with and without disabilities. They also

found that the Maze was sensitive to change and could be used to monitor progress in reading comprehension (Guthrie et al., 1974).

Additional published studies addressed the use of the Maze assessments with students who are not proficient English speakers. Two studies of the untimed Maze task have suggested that this task is an appropriate measure of reading comprehension for this population. Hinofotis and Snow (1980) assessed incoming foreign college students with language backgrounds in Arabic, Japanese, Farsi, French, Spanish, Chinese, and Vietnamese using both the Maze and an English language placement test. A moderate correlation of .63 was found between the measures. In the second study, students in Grades three through five who spoke with native Australian dialects were assessed using the Maze tasks and the *Gates-MacGinitie* vocabulary and comprehension subtests. The results revealed a moderate to high correlation of .68 to .83 for vocabulary and .69 to .76 for Comprehension (Baldouf & Propst, 1978).

One additional study examined the use of the timed Maze task with English language learners. Wiley and Deno (1995) examined oral reading and Maze measures with both English language learners and native English speakers in grades three and five at an urban elementary school. The results indicated that correlations between the Maze task and the *Minnesota Comprehensive Assessment* were moderately strong for all students in both grades. These findings suggest that the Maze task is an appropriate measure to use when assessing the English reading comprehension of English language learners.

The second wave of research on the Maze task focused on correlating scores on the Maze measure with scores on the oral reading fluency measure and other well-established reading measures. Because of this, studies in the 1990's set time limits for the task to parallel oral reading measures and better standardize the procedure (Deno, Maruyama, Espin & Cohen, 1990). Later studies confirmed that timed Maze scores are more valid and less negatively skewed than untimed scores (Parker, Hasbrouck & Tindal, 1992; Shin, Deno, & Espin, 2000).

Espin, Deno, Maruyama, and Cohen (1989) reported on the technical adequacy of a Maze measure that was part of a group-administered screening instrument called the *Basic Academic Skills Samples* (BASS; Deno, Maruyama, Espin & Cohen, 1989). This subtest consisted of three 1 minute Maze selection tasks that were at the first to second grade level. The BASS was administered to more than 2,000 elementary school students across 31 schools. The subsequent analysis revealed that the correlation between the number of correct words read orally and the number of correct maze replacements was from .77 to .85 in the third through fifth grades. Data from the entire sample revealed a stable pattern of increase in Maze scores from Grade one to six, as well as from winter to spring within each grade.

In light of these positive findings, Fuchs and Fuchs (1990) conducted preliminary studies which established correlations of .83 between scores on Maze and scores on reading-aloud measures and of .77 between scores on Maze and the Reading Comprehension subtest of the Stanford Achievement Test (SAT; Fuchs & Fuchs, 1990). In 1993, Jenkins and Jewell extended this work by examining the relation between Maze

scores and the oral reading rate, scores on the Gates MacGinite Reading Tests, and the Metropolitan Achievement Tests for first grade students. The results were promising, with correlations between .80 and .89 with oral reading fluency, .85 with the Gates MacGinitie, and .80 with the Metropolitan Achievement Test (Jenkin & Jewells, 1993).

In 1992, Fuchs and Fuchs published a study that extended the research on Maze selection in order to find a CBM reading measure that might be suitable for data collection via the computer and would have a greater acceptance for teachers than had been reported for read-aloud measures. The Maze selection in this study was a 2.5 minute measure administered twice weekly for 18 weeks via computer. Technical adequacy and level of teacher acceptance were compared for several alternative CBM measures including question answering, story recall, cloze, and Maze selection. The results of this study found that the Maze task was sensitive to change in performance over time. Additionally, unlike other measures, the Maze tasks had a relatively small ratio of slope to standard error of estimate, which makes it easier to detect growth on a graph. Moreover, teachers rated their satisfaction with Maze highly. It was in this paper that CBM maze was proposed as a viable alternative for screening and progress monitoring (Fuchs & Fuchs, 1992).

Since this article was published, use of the Maze has become increasingly popular in schools and a number of researchers have continued to investigate its usage. These researchers have provided evidence that Maze yields reliable and valid scores that are sensitive to growth and can differentiate poor readers from typical readers at both the classroom and the district level (Brown-Chidsey, Davis, & Maya, 2003; Yeo, 2010).

Moreover, Shin, Deno, and Espin (2000) reported that Maze scores provided reliable estimates of student growth and showed group growth as well as individual differences. They also established that repeated administration of Maze measures yields alternate-form reliability coefficients in the .80s (Shin, Deno, & Espin, 2000). In 2010, Yeo synthesized 27 studies regarding the relationship between ORF, Maze, and statewide achievement tests in reading. Results of this multi-level meta-analysis indicated a large correlation coefficient of .69. Additionally, both CBM measures were able to identify English language learners and students with disabilities versus their typically developing peers.

Additional studies have revealed that the predictive validity of Maze measures is not significantly different from read-aloud measures when correlated with state accountability tests (Shin et al, 2000; Silberglitt et al., 2006; Wiley & Deno, 1995). Specifically, Wiley & Deno (1995) found that both Maze and ORF measures are predictive of performance on the *Minnesota Comprehensive Assessment* in reading. However, the Maze task was a better predictor of performance on the *Minnesota Comprehensive Assessment* than ORF assessments for fifth grade students. Silberglitt and colleagues (2006) also examined the relationship between the Maze task and the *Minnesota Comprehensive Assessment* reading subtest and found that the Maze task was able to predict performance on the state standards test. In 2000, Shinn, Deno, and Espin examined the validity of the Maze task by looking at the relationship between students' growth rates estimated on repeated Maze scores and student performance on the reading

subtest of the California Achievement Tests. The results revealed a significant positive relationship between the two measures.

Few researchers have explored the validity of the Maze task across grade levels. An initial study by Jenkin and Jewels (1993) determined that Maze scores have greater stability across grades two through six when compared to read-aloud scores which show decreasing validity after grade four. Similarly, in 2000, an analysis by Shinn, Deno, and Espin revealed Maze to be superior to oral reading fluency in sensitivity to student growth over time for eighth grade students. Finally, Silberglitt et al. (2006) found that the magnitude of the relationship between oral reading fluency scores and state accountability tests decreases from a strong relationship in grade 3 to a moderate relationship in grade 8, while maze scores maintained a moderate relationship across grades 7 and 8.

Standardized Test for Assessment of Reading (STAR-R)

Another increasingly popular option for screening and progress monitoring in schools is the STAR Reading test. STAR Reading (STAR-R) is a computer adaptive, norm-referenced reading test developed by Renaissance Learning in 2001 as a measure used for screening and progress monitoring students' progress in reading. Computer-adapted testing is a form of computer-based testing that is based on item response theory, which adapts to the examinee's ability level. Computer-adapted testing successively selects questions so as to maximize the precision of the exam based on what is known about the examinee from previous questions (Weiss & Kingburry, 1984).

STAR-R has many advantages over traditional CBM measures. Firstly, STAR-R is a computer-based assessment and can be administered to groups of students within 10-15 minutes, making it more time efficient for teachers to administer. A second advantage is the data from the STAR-R assessment is readily available to teachers and administrators through the STAR database within 15 minutes of the assessment being completed. This nearly instant feedback allows teachers to make immediate changes in instruction or provide intervention to students identified as at-risk.

Preliminary studies conducted by Renaissance Learning (2001) have shown promising results for the reliability and validity of this test. The reliability of STAR-R was established through testing three types of reliability – test-retest reliability, split-test reliability, and generic reliability – on 29,169 students in grades one through twelve. Results indicated an overall reliability of .95 for all grades, with a split-test reliability of .92 and a test-retest coefficient of .91 (Renaissance Learning, 2001). These results have not been confirmed by independent research.

Renaissance Learning also conducted additional analysis to determine the validity of the STAR-R assessment. In 2010, Renaissance Learning gathered data from approximately 12,000 students in schools using the STAR-R assessment. This data included test results from several large-scale state assessments including the California Achievement Test (CAT), the Comprehensive Test of Basic Skills (CTBS), the Iowa Test of Basic Skills (ITBS), the Metropolitan Achievement Test (MAT), and the Stanford Achievement Test – in addition to the students' STAR-R scaled score results. Overall, the predictive validity coefficients ranged from .68 to .82 in grades one to six with an

overall average of .79. For grades seven to twelve, the predictive validity coefficients ranged from .81 to .85, with an overall average of .82 (Renaissance Learning, 2010). To date, two independent researchers have conducted validity studies regarding STAR-R. Bennicoff-Nan (2002) found with-in grade correlations averaging .82 when comparing STAR-R to the Stanford 9 and .80 with the California Standards Test. Yoes (1999) found within-grade correlations with the Degrees of Reading Power Program averaging .81.

Current Study

Presently, limited research exists that can establish the predictive validity of the Maze task. Predictive validity is an indication of how well performance on a criterion measure is predicted by performance on a screening measure and is a key component in determining the quality of a screening measure (Salvia, Ysseldyke, & Bolt, 2009). Additionally, no previous studies have focused exclusively on the utility of these measures across the middle school years. The purpose of this study is to examine the strength of the relationship between the Maze task and STAR-R screening measures and the California Standards state accountability test, as well as if the strength of these relationships change as a function of grade. Additional research goals are to compare the predictive validity and diagnostic accuracy of these assessments. This study addresses the following research questions (a) What is the strength of the relationship between the Maze task and state accountability test scores? (b) What is the strength of the relationship between the STAR reading measure and state accountability test scores? (c) To what extent do the STAR-R and the Maze task differ in their ability to predict CST scores? (d) To what extent do grade level differences exist in the relationship between STAR-R and

maze measures and the CST state accountability test scores? (e) Which measure shows greater accuracy in identifying students classified as at-risk or not at-risk for reading failure on the CST?

Method

Setting and Participants

In the 2009, three middle schools in a public urban school district in southern California screened their sixth, seventh, and eighth grade students for reading failure using both Star-R and Maze at three time points throughout the year – the fall (July through November), winter (December through March), and spring (April through June). The results of this screening were gathered and analyzed. In total, data from 4,147 students were available; however, when students who did not have data for all three screening points were excluded this number was reduced to 1,479. Of these, 320 students were in Grade 6, 914 in Grade 7, and 245 in Grade 8.

A further examination of student demographics revealed that approximately 67 percent of students identified as Latino/Hispanic, 17 percent were African American, 13 percent Caucasian, 2 percent Asian and 1 percent of students identified as Native American/Alaskan. Fifteen percent of the sample was classified as English language learners. Additionally, 65.8 percent of students were eligible for free or reduced lunch programs.

Materials and Procedure

As part of the curricula of these California middle schools, all students were administered a screening instrument at least three times a year in order to predict which

students were unlikely to reach proficiency on the state's criterion-referenced test, which is used to report adequate yearly progress to the U.S. Department of Education as a condition of the No Child Left Behind policy (2001). The staff at the participating schools chose to use both STAR-R and AIMSweb CBM-Maze measures to screen for student failure.

AIMSweb Maze. Teachers in these schools screened the reading ability of all students three times a year, in the fall, winter, and spring, using AIMSweb Maze CBM. This measure consists of one passage created by deleting every seventh word of a reading passage and replacing it with a choice of three words, only one of which is correct. Students were asked to read the passage silently and circle the correct words when they came to a choice. Maze passages were selected by grade level. The students read one three-minute passage, and the score is determined by the number of word choices made correctly.

In the middle school grades, maze tasks are moderate to highly reliable across grades with alternate form reliabilities that range from .70 to .91 (Brown-Chidsey, Davis, & Maya, 2003; Shinn, Deno, & Espin 2000; Ticha, Espin & Wayman, 2009). Additionally, convergent validity coefficients ranged from .48 to .79 for middle school and high school students while criterion validity coefficients ranged from .32 to .88 (Espin & Foegen, 1996; Fuchs, Fuchs, & Maxwell, 1988; Silberglitt, Burns, Madyun & Lail; Ticha et al., 2009).

Standardized Test for the Assessment of Reading. The Standardized Test for the Assessment of Reading (STAR-R) is a nationally normed, computer-adaptive test for

reading achievement. It was administrated to all middle school students via computer three times over the school year (fall, winter, and spring) in order to monitor students' progress toward state standards. Computer-adaptive tests continually adjust the difficulty level of each student's assessment by choosing test questions based on the responses to previous questions. The process of customizing item difficulty to match a student's skill level was achieved through item response theory. Item response theory is able to place student performance and item difficulty on the same scale and offers a means to estimate the probability that the student will answer a test item correctly. Previous research studies have shown that assessment based on computer-adaptive testing is more efficient than conventional tests (Lord, 1980; McBride & Martin, 1983; Agdonini & Harris, 2010). Initial studies conducted by Renaissance Learning indicate a concurrent validity for middle schools students ranging from .72 to .75 while predictive validity ranges from .81 to .82 (Renaissance Learning, 2008).

California Standards Test. The California Standards Test (CST) is a standardized, curriculum-based test that is administered to all California students in grades two through eleven. The CSTs are designed to match the state's academic content standards for each grade. All students take this between mid-March and mid-June of every year. At the middle school level, the CSTs cover the topics of mathematics and English/language arts. For the purposes of this study, only the English Language Arts score was analyzed. The CST English-Language Arts is a criterion-referenced assessment intended to measure selected California English-Language Arts content standards. This portion of the test requires students to perform comprehension tasks such

as identifying the main idea and supporting details of a passage, demonstrating techniques for learning new vocabulary, understanding ideas not explicitly stated, making predictions based on information, and drawing conclusions based on information in the passage (California Department of Education, 2011).

Test items for middle school students are multiple-choice with the exception of an additional writing component. The scores on the writing task and the multiple-choice questions are combined to create an overall score and accompanying performance level. These scores range from 150 to 600 points which correspond to five levels of performance: far below basic, below basic, basic, proficient, and advanced (California Department of Education, 2011). Test-retest reliability estimates provided by the California Standards Tests Technical Report ranged from 92-.94. Validity scores relating the scores on the CST to the CAT/6 Survey Reading Language and the CST were found to have a strong alignment to the California content standards (California Department of Education, 2009).

The CST was administered in the spring of 2010 by trained classroom teachers and/or by other trained school personnel. School testing coordinators followed the California Education test administration guidelines when administrating the CST (California Department of Education, 2009).

Results

The purpose this study was to analyze the relationship between the curriculum based measurement Maze task, STAR-R assessments, and the California Standards state accountability test to examine the strength of these relationships at three points

throughout the year. The study addressed the following research questions: (a) What is the strength of the relationship between the Maze task and state accountability test scores? (b) What is the strength of the relationship between the STAR reading measure and state accountability test scores? (c) To what extent do
STAR-R and the Maze task differ in their ability to predict CST scores? (d) To what extent do grade level differences exist in the relationship between STAR-R and Maze measures and the CST state accountability test scores? (e) Which measure shows greater accuracy in identifying students classified as at-risk or not at-risk for reading failure on the CST?

Maze scores and STAR-R standard scores were correlated with California Standards Test (CST) standard scores for the fall, winter, and spring screening points in order to determine the strength of the relationships between CBM-Maze scores, STAR-R scores, and scores on the CST. All correlations were determined using Pearson product-moment statistics. The results indicated a significant correlation between Maze scores and CST scores in the fall (r = .36, p = .00), winter (r = .38, p =.00), and in the spring (r = .41, p = .00). Similarly, results from a second Person correlation revealed that STAR-R standard scores are significantly correlated to CST standard scores at all three screening points (r = .48, p = .00; r = .47, p = .00; r = .39, p = .00). A summary of these results can be found in Table 2.

Table 1
Descriptive Statistics of Maze and STAR-R Scores

			Range	
Variable	M	SD	Min	Max
Maze - Fall	22.96	9.13	2	48
Maze - Winter	26.41	9.68	3	52
Maze-Spring	31.33	14.16	2	261
Star-R Fall	575.71	213.59	77	1340
Star-R Winter	596.29	237.60	93	1342
Star-R Spring	618.60	239.85	0	1340
California Standards Test	336.91	51.32	214	508

Note. The aimsweb maze task is from Shinn (2002) and the Standardized Test of Assessment in Reading (Star-R) is from Renaissance Learning (2008)

Table 2
Correlations Between Maze, STAR-R, and CST scores

Measure	Benchmark	Pearson Correlation	Sig (2 tailed)
Maze	Fall	.36	.00
	Winter	.38	.00
	Spring	.41	.00
STAR-R	Fall	.48	.00
	Winter	.47	.00
	Spring	.39	.00

Differences in correlation magnitude between STAR-R and CST scores and CBM-Maze and CST scores were examined using a Steiger's *Z* analysis. This type of analysis is able to reveal whether two correlations obtained from two different models have different strengths (Steiger, 1980). First, a Fisher's *Z* transformation was used to determine the significance of the correlation coefficient *z* for each pair of measures. Then, this statistic was used in the Steiger's *Z* test to compare the correlation coefficients at the 5% significance level. Table 3 displays the Steiger's *Z* coefficient between STAR-R−CST and Maze−CST scores for fall, winter, and spring. STAR-R scores were a better predictor of CST scores during the fall and winter, whereas there was no statistically significant difference between Maze scores and STAR-R scores as predictors of CST scores during the spring. Additional information regarding the means and standard deviations for the study variables are also reported in Table 1.

Table 3
Steiger's Z values comparing the strengths of the correlations between screening measures and the CST

Measure	CST		
	Fall	Winter	Spring
Maze	.36	.38	.41
STAR-R	.48	.47	.39
Fisher's Z Test	3.27*	2.96*	-0.49

*$p<.05$

Also of interest is whether grade level differences exist in the relationship between STAR-R and Maze measures and the CST state accountability test scores. A Fisher's *r*-to-*Z* transformation with an alpha level of .01 was used to compare the correlation coefficients between grades. As seen in Table 4, the results for the Maze task indicated that the strength of the correlation between Maze screening scores and CST scores increased from grade six through grade eight. The coefficient for sixth graders accounted for 15% of the variance. This rises to 30% for seventh graders and 66% for students in eighth grade. These results demonstrate that Maze continues to account for a substantial amount of variance in student performance in early middle school, and that the overall value of this predictor increases significantly as middle school progresses.

Table 4
Fisher's z Transformation Result Comparing Coefficients Between Maze, Star-R, and CST Score for Grade Levels

	Grade	Grade 6	Grade 7	Grade 8
Maze	Grade 6	-	2.58*	2.86*
	Grade 7		-	2.11*
	Grade 8			-
Star-R	Grade 6	-	2.84*	3.71*
	Grade 7		-	2.87*
	Grade 8			-

* $p < .05$

Similar results were found for the STAR-R screening measure. Data from this analysis revealed that the relationship between the STAR-R measure and CST scores differ significantly between all grades, with the relationship decreasing from grade 6 to grade 8. The coefficient for sixth graders accounted for 6% of the variance. Seventh grade scores accounted for 26% of the variance in CST scores and this increased to 81% for eighth grade students.

A diagnostic accuracy analysis was also conducted. To examine the diagnostic accuracy of STAR-R and Maze, an initial analysis was performed using STAR-R and Maze as predictor variables and the CST as the criterion measure. The goal of this analysis was to assess the sensitivity, specificity, and predictive power of STAR-R and Maze over a range of cut-score values. Sensitivity refers to the likelihood that STAR-R or Maze will accurately identify those students who have been identified by the CST as being below the current standard. Similarly, false negatives denote the likelihood that STAR-R or Maze will fail to accurately identify students who have been identified on the CST as being below proficient. False negatives and sensitivity statistics add to 100%.

Conversely, specificity refers to the likelihood that STAR-R or Maze will accurately identify those students who have been identified by the CST as achieving state standards in reading. Correspondingly, false positives refer to the likelihood that STAR-R or Maze will fail to correctly identify those students who have been identified on the CST as being below proficient in overall reading skills. False negatives and sensitivity total to 100% (Meng, Rosenthal, & Rubin, 1992).

In the first portion of this analysis, the scaled score of the STAR-R was used to predict the scaled score of the CST and, conversely, CST scores were also used to predict scores on the STAR-R. A similar analysis was conducted using the Maze total score to predict scores on the CST and using the CST to predict scores on the Maze. All analyses were conducted using author suggested cut-scores of below 560 for the STAR-R assessment, 349 for the CST, and a score of 17 or below for the Maze in the fall, 18 or below for Maze in the winter, and 20 or below for Maze in the spring.

Results of the analysis indicated that the STAR-reading assessment is adequately sensitive to scores on the California Standards Tests for the fall, winter, and spring time points investigated, with sensitivity levels ranging from .69 to .76 (see Table 6). Similarly, the specificity of STAR-R was also moderate for all three time points, with specificity levels ranging from .61 to .69. Furthermore, the probability that STAR-R will accurately predict which students are likely to score below proficient was adequate for all time periods, ranging from .76 to .77. The statistic for negative predictive power was also moderate for all three time points (.61, .61, and .53). Overall, the ability of STAR-R to accurately predict a student's correct diagnostic classification using the suggested CST cut-scores was 35.4% for STAR-R during the fall screening period, 36.2% for winter, and 28.7% for spring. This was accompanied by low levels of Phi and Kappa coefficients. The results of the diagnostic accuracy analysis are displayed in Table 6.

A parallel diagnostic accuracy analysis was conducted regarding Aimsweb CMB-Maze and the CST. This is portrayed in Table 6. Sensitivity statistics indicated low levels of sensitivity for all time points (.38, .27, and .20). On the other hand, maze

showed high levels of specificity, ranging from .90 to .97. Maze also shows high levels of positive predictive power (.85, .86, and .91). Negative predictive power, however, was only moderate, ranging between .49 and .46. The overall accuracy with which maze correctly identified a students' diagnostic classification on the CST was less than chance with low Phi and Kappa coefficients.

Table 5
Performance of STAR Over a Range of Cut –Scores Using CST Composite Scores as the Criteria

STAR	Sensitivity	Specificity	PPP	NPP	CC	Phi	Kappa
Fall	.76	.68	.76	.61	.35*	.38*	.38*
Winter	.75	.69	.78	.60	.36*	.39*	.39*
Spring	.69	.61	.77	.53	.29*	.30*	.29*
Maze							
Fall	.38	.90	.85	.49	.30*	.31*	.25*
Winter	.27	.93	.86	.47	.25*	.26*	.18*
Spring	.20	.97	.91	.46	.24*	.25*	.15*

*p<.001

Results in Table 6 and Figure 1 represent the ROC curves for the STAR-R and Maze measures in predicting total scaled scores on the CST. Results indicated that STAR-R and Maze both resulted in moderate to adequate levels of sensitivity and specificity (.75 or higher; Swets, 1996). These statistics indicate that both STAR-R and

Maze reading assessments are able to predict CST group placement at a rate significantly greater than chance.

Table 6
Results of ROC analysis for Star-R and Maze

Test Result Variables	Area	Std. Error	Sig.	95% confidence interval	
				Lower	Upper
STAR-R	.75	.02	.00	.71	.78
Maze	.77	.02	.00	.73	.80

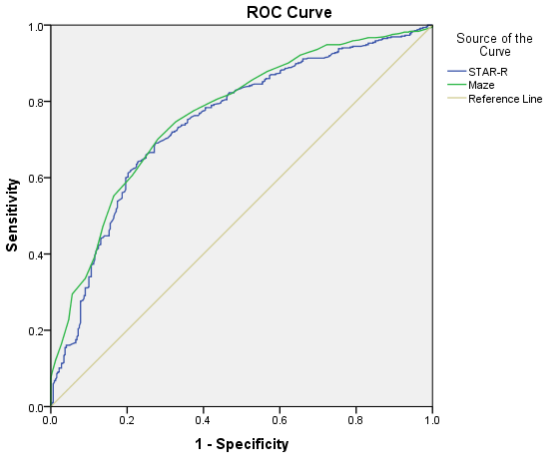

Figure 1. ROC curves for STAR and Maze predicting CST composite scores.

Discussion

Relationship between the Maze task and the CST

The first set of results yielded evidence that the CBM Maze task is predictive of students' performance on the CST when used as a screening measure for reading achievement. These findings are consistent with the results from previous research literature which has shown a significant relationship between Maze and other well-established reading measures and state accountability tests such as the Gates MacGinitie Reading Test and the Metropolitan Achievement Test (Shin et al., 2000; Siberglitt et al., 2006; Wiley & Deno, 1995).

A more unique aspect of this research design explored the difference in predictive validity coefficients between the Maze and the CST across grades. Currently, no peer-reviewed articles have explored this relationship in middle school students. Results revealed that the strength of the correlation increased from sixth grade to eighth grade, suggesting that Maze is most appropriate for use with eighth grade students. Future research in this area should explore the use of Maze as a reading screener for high school students in order to see if the predictive accuracy of Maze continues to increase with grade level.

Relationship between STAR-R and the CST

Additional Pearson correlations and Fisher's Z analyses between the STAR-R reading assessment and the California Standards Test (CST) produced results similar to those seen in the Maze task. STAR-R scores were found to be significantly predictive of CST scores at all three time points, with the strength of the correlation increasing from

sixth to eighth grade. Although there is a paucity of peer-reviewed research on the STAR-R reading measure, these results are comparable to initial findings that established STAR-R as being predictive of reading success in children (Renaissance Learning, 2000; 2010). Future research in this area should examine the utility of STAR-R for identifying reading failure in high school students, in order explore this pattern of increasing predictive validity and help to identify populations of students that may be able to benefit from the features of this exam.

Comparing the Maze Task and STAR-R

With significant relationships found between STAR-R and the CST, and the Maze task and the CST, the question then arises: to what extent is there a difference between the ability of the STAR-R assessment to predict CST scores and the Maze task to predict CST scores? Additional analyses comparing the STAR-R and Maze task predictive abilities indicated that there are statistically significant differences between the Maze task and the STAR-R measure in their ability to accurately predict CST scores at each time point. This is an important finding because school districts usually choose one mode of screening students in reading and want to use a measure that would most accurately predict future scores. These results suggest that STAR-R would predict CST scores better than the Maze during the fall and winter, whereas there is no evidence that either is a significantly better predictor of CST scores than the other during the spring.

Diagnostic Accuracy Analysis

Although scaled scores between STAR-R and the CST were significantly correlated, the diagnostic accuracy analysis revealed that STAR-R shows only moderate

to adequate levels of specificity and sensitivity. Furthermore, use of the STAR-R only led to approximately 30 to 40 percent of students being correctly classified based on their CST scores, which is no greater than chance. Maze results indicated low levels of sensitivity and high levels of specificity. This indicates that the use of these cut-score leads to a few positives-negatives and higher levels of false negatives. This is concerning because it suggests that the Maze task maybe under-identifying students at risk for failure on the CST. Further research in this area should seek to determine if either the Maze task or the STAR-R assessment are able to correctly classify students identified as being below proficient in reading on state accountability tests.

Limitations

Limitations of this study include that given that the STAR Reading assessment is a measure that was developed recently (2001), a solid peer-reviewed research base has been established regarding the validity and reliability of this evaluation system. Additionally, the sample is from a largely Hispanic urban population and therefore these findings may not generalize fully to schools in rural areas. The results of this study, however, are relevant for large urban school districts or communities with diverse populations. Moreover, data was not available on students who did not have data for all three screening points, therefore students who changed schools or were not available for testing were excluded from analysis.

Conclusions

In summary, the present study adds to the research base of studies that have established Maze task probes and the STAR-R reading assessment as valid screening

assessments for reading skills. The fact that this study produced mixed results with the diagnostic accuracy of both the STAR-R and Maze assessment when compared with the CST suggests that future research should focus on the utility of these measures in a high school setting, as well as on the diagnostic accuracy of STAR-R and Maze to standardized state assessments. Significant levels of correlation between both STAR-R, Maze, and the CST indicate that these measures have a strong relationship even with inadequate ability to correctly classify students.

References

Baldouf, R. B., & Propst, I. K. (1978). Matching and multiple-choice cloze tests. *The Journal of Educational Research, 72,* 321–326.

Brown-Chidsey, R., Davis, L., & Maya, C. (2003). Sources of variance in curriculum-based measures of silent reading. *Psychology in the Schools, 40,* 363-377.

California Department of Education Assessment and Accountability Division. (2010). *California Standards Tests Technical Report: Spring 2010 administration*. Retrieved from http://www.cde.ca.gov/ta/tg/sr/documents/csttechrpt2010.pdf

Carver, R. P. (1992). What do standardized tests of reading comprehension measure in terms of efficiency, accuracy, and rate? *Reading Research Quarterly, 27,* 347-359.

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

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

Deno, S., Maruyama, G., Espin, C., & Cohen, C. (1990). Educating students with mild disabilities in general education classrooms: Minnesota alternatives. *Exceptional Children, 57,* 150-161.

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

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. (1994). 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, & Foegen, A. (1996). Validity of general outcome measures for predicting secondary student's performance on content-area tasks. *Journal of Exceptional Children, 62,* 497-514.

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

Fuchs, L. S., & Fuchs, D. (1990). The role of skills analysis in curriculum-based measurement in math. *School Psychology Review, 19*(1), 6-22.

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

Fuchs, L. S., Fuchs, D., & Maxwell, L. (1988). The validity of informal reading comprehension measures. *Remedial and Special Education, 9*, 20-28.

Fuchs, L. S., Fuchs, D., Hamlett, C. L., & Ferguson, C. (1992). Effects of expert system consultation within curriculum-based measurement using a reading maze task. *Exceptional Children, 58,* 436–450.

Guthrie, J. T. (1973). Reading comprehension and syntactic responses in good and poor readers. *Journal of Educational Psychology, 65,* 294-299.

Guthrie, J. T., Seifert, M., Burnham, N. A., & Caplan, R. I. (1974). The maze technique to assess, monitor reading comprehension. *The Reading Teacher, 28,* 161-168.

Hamilton, C., & Shinn, M. R. (2003). Characteristics of word callers: An investigation of the accuracy of teachers' judgments of reading comprehension and oral reading skills. *School Psychology Review, 32*, 228–240.

Hinofotis, F. B., & Snow, B. G. (1980). An alternative cloze testing procedure: Multiple choice format. In J.W. Oller & K. Perkins (Eds.), *Research in language testing* (pp. 129-133). Rowley, MA: Newbury House.

Hintze, J. M., Shapiro, E. S., Conte, K. L., & Basile, I. M. (1997). Oral reading fluency and authentic reading materials: Criterion validity of the technical features of CBM survey-level assessment. *School Psychology Review, 26,* 535-553.

Ikeda, M. J., Neessen, E., & Witt, J. C. (2008). Best practices in universal screening. In A. Thomas & J. Grimes (Eds.), *Best practices in school psychology* (5th ed., Vol. 2, pp. 103–114). Bethesda, MD: National Association of School Psychologists.

Kingston, A. J., & Weaver, W. W. (1970). Feasibility of cloze techniques for teaching and evaluating culturally disadvantaged beginning readers. *The Journal of Social Psychology, 82,* 205-214.

Jenkins, J., Fuchs, L., Espin, C., van den Broek, P., & Deno, S. (2000). Effects of task format and performance dimension on word reading measures: Criterion validity,

sensitivity to impairment, and context facilitation. *Journal of Educational Psychology, 95*, 719-729.

Jenkins, J.R., & Jewell, M. (1993). Examining the validity of two measures for formative teaching: Read aloud and maze. *Exceptional Children, 59,* 421-432.

McGlinchey, M. T., & Hixon, M. D. (2004). Using curriculum-based measurement to predict performance on state assessments in reading. *School Psychology Review, 33,* 193–203.

Meng, X. L., Rosenthal, R., & Rubin, D. B. (1992). Comparing correlated correlation coefficients. *Psychological Bulletin*, *111*: 172-175.

No Child Left Behind Act of 2001, Pub. L. No. 107-110.

Parker, R., Hasbrouck, J. E., & Tindal, G. (1992). The maze as a classroom-based reading measure: Construction methods, reliability, and validity. *The Journal of Special Education, 26,* 195-218.

Salvia, J., Ysseldyke, J. E., & Bolt, S. (2009). *Assessment: In special and inclusive education* (11th ed.). New York: Wadsworth.

Shin, J., Deno, S. L., & Espin, C. (2000). Technical adequacy of the maze task for curriculum-based measurement of reading growth. *The Journal of Special Education, 34,* 164-172.

Shinn, M. R., Good, R. H., Knutson, N., Tilly, W. D., & Collins, V. L. (1992). Curriculum-based measurement of oral reading fluency: A confirmatory factor analysis of its relation to reading. *School Psychology Review, 21,* 459-479.

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

Stage, S. A., & Jacobsen, M. D. (2001). Predicting student success on a state-mandated performance-based assessment using oral reading fluency. *School Psychology Review, 30,* 407-419.

Renaissance Learning. (2001). *Understanding reliability and validity (Version 2.2)*. Wisconsin Rapids, WI: Advantage Learning Systems.

Renaissance Learning. (2010). *The foundation of the STAR Assessments*. Wisconsin Rapids, WI: Advantage Learning Systems. http://doc.renlearn.com/KMNet/R001480701GCFBB9.pd

Steiger, J. H. (1980). Test for comparing elements of a correlation matrix. *Psychological Bulletin, 87*(2), 245-251.

Swets, J. A. (1996). *Signal detection theory and ROC analysis in psychology and diagnostics.* NJ: Lawrence Erlbaum.

Sum, A., Katiwada, I., & McLaughlin, J. (2009). The consequences of dropping out of high school: Jobless and jailing for high school dropouts and the high cost for taxpayers. Center for Labor Market Studies Publications. Paper 23 http://hdl.handle.net/2047/d20000596

Ticha, R., Espin, C. A., & Wayman, M. M. (2009). Reading progress monitoring for secondary students: Reliability, validity, and sensitivity of read-aloud and maze-selection measures. *Learning Disabilities Research & Practice, 24,* 132-142.

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

Wiley, H. I., & Deno, S. L. (2005). Read-aloud and maze measures as predictors of success for English learners on a state standards assessment. *Remedial and Special Education, 26,* 207-214.

Yeo, S. (2010). Predicting performance on statewide achievement tests using curriculum-based measurement in reading: A multi-level analysis. *Remedial and Special Education, 25,* 170-190.