A Comparison of Progress Monitoring Scores and End-of-Grade Achievement

Bob Algozzine
Chuang Wang
Anatoli Boukhtiarov
University of North Carolina at Charlotte

Abstract

In this study, we addressed the need for research demonstrating the extent to which performance on widely-used progress monitoring measures related to end-of-grade achievement on statewide assessments. Specifically, we evaluated the usefulness of STAR Reading and Scholastic Reading Inventory-Interactive for predicting performance on the Florida Comprehensive Assessment Test. We found that scores obtained from regular use of these measures were statistically significantly related to overall end-of-grade achievement markers. We discuss our findings in the context of other similar research.

The No Child Left Behind (NCLB) Act (U.S. Department of Education, 2004) has raised academic standards for all children. Most professionals agree that there has been a correlated increase in the amount of time students spend in assessment-related activities, especially those linked to high stakes testing and similar education reforms (Ananda & Rabinowitz, 2001; Brown & Coughlin, 2007; Simpson, LaCava, & Graner, 2004; Sibley, Biwer, & Hesch, 2001). NCLB also directs that schools use frequent classroom-based assessments to keep track of the progress students are making in reading and other academic areas (Perie, Marion, & Gong, 2007; Schilling, Carlisle, Scott, & Zeng, 2007). Gathering yearly student performance data on local, state, and national indicators has been a part of America's educational accountability efforts for some time and most states require participation of all students in reading and math assessments during elementary, middle, and high school years (Simpson et al., 2004; Thurlow & Thompson, 1999; Thurlow & Wiley, 2004). Frequent academic progress monitoring has achieved a new level of prominence as a critical feature of Response-to-Intervention (RTI) practices promising to reform the numbers and types of children receiving special education and outcomes for all children (Fuchs & Deshler, 2007). The assumption is that frequently reviewing performance will help teachers identify students who are at-risk and not making adequate progress so that they have a basis for devising suitable plans for instruction. This belief and emerging system is grounded in the existence and use of valid and reliable predictors of students' progress toward the goal of achieving grade-level reading skills (Buck, Torgesen, & Schatschneider, n.d.; Brown & Coughlin, 2007; Fuchs & Deshler, 2007; Hintze, Callahan, Matthews, Williams, & Tobin, 2002; Hintze, Ryan, & Stoner, 2003; Hixson & McGlinchey, 2004; Perie et al., 2007; Roehrig, Petscher, Nettles, Hudson, & Torgesen, 2008).

Literacy skills are fundamental to successful academic performance and frequent assessment and monitoring of them is the foundation for response-to-intervention practices that inform teachers about their students' instructional needs (Fuchs & Fuchs, 2006). In this regard, Coyne and Harn (2006) point out that knowledge of literacy performance directs improved outcomes by providing important

answers that support data-based decision making to improve instruction. In recent years, researchers and other education professionals have expressed concern about the importance this placed on high stakes achievement testing (Wixson & Carlisle, 2005) and there is continuing concern that infrequent, general, end-of-the-year assessments are not useful for making instructional decisions at the classroom level (Roehrig et al., 2008).

Research shows, and virtually all educators agree, that academic improvement requires practice to reinforce skills being learned and continuous monitoring of progress to ensure appropriate areas are targeted for instruction (Fuchs, 1989; Scott & Weishaar, 2003). Unfortunately, the role of practice and progress monitoring is often overlooked and misunderstood. Just setting aside time for student practice is not enough. Similarly, checking performance several times a year provides insufficient evidence for improving skills requiring more frequent attention. Practice must be personalized to each student's individual ability level and immediately followed by informed feedback to ensure a high rate of engagement and success. It must also provide progress monitoring evidence for teachers and other professionals to use to improve instruction and outcomes.

Progress-monitoring assessments must meet several requirements (Brown & Coughlin, 2007; Deno, 1992, 1997, 2003; Deno & Mirkin, 1977; Fuchs & Fuchs, 1999; Perie et al., 2007; Roehrig et al., 2008). First, the content used for keeping track of progress must be representative of the academic performance expected of students at the end of the school year. The measures must also be free of floor or ceiling effects and sensitive to change over a short period of time, over repeated measurements as students gain more skills. The assessment also must be authentic and have adequate technical characteristics (i.e., validity and reliability). Finally, the outcomes must accurately predict improvements on more generalized assessment measures, such as standardized tests. Progress-monitoring tools that meet or exceed these requirements provide valid and reliable data.

Teachers use progress-monitoring to make decisions about an individual student's instructional needs. For example, based on a student's performance, a teacher may increase the amount and type of instruction, slow the pace of it, or change methods of teaching completely. The use of progress-monitoring instruments has been demonstrated to improve student outcomes in academic areas and has been a widely-accepted evidence-based practice in special education for many years (e.g., Fuchs, 2004; Fuchs & Fuchs, 1986; Madelaine & Wheldall, 2004; Safer & Fleischman, 2005). Relationships between progress monitoring measures and end-of-grade outcomes have also been reported for elementary school students across several demographic groups and statewide assessments (e.g., Barger, 2003; Buck & Torgesen, n.d.; Buck, Torgesen, & Schatschneider, n.d.; Roehrig et al., 2008; Vander Meer, Lentz, & Stollar, 2005; Wanzek, Roberts, Linan-Thompson, Vaughn, Woodruff, & Murray, 2010; Wilson, 2005). This extant knowledge base is grounded in studies illustrating the predictive value of early literacy skills (e.g., oral reading fluency) for success in third grade. In this research, we investigated similarities and differences in performance as well as relationships between scores and predictive accuracy of two widely-used progress monitoring assessments and a state-wide high stakes reading test for students in grades six, seven, and eight. Specifically, we addressed on three research questions with implications for improving summative and

formative assessment practices for at-risk students in middle school:

1. To what extent is performance for sixth, seventh, and eighth grade students on different measures of progress monitoring and end-of-grade reading achievement similar for different demographic groups?
2. To what extent is performance for sixth, seventh, and eighth grade students on different measures of progress monitoring and end-of-grade reading achievement related?
3. To what extent does performance for sixth, seventh, and eighth grade students on different measures of progress monitoring similarly predict performance on end-of-grade reading achievement?

Method

Renaissance Learning offers a computer-adaptive test of general reading ability (STAR Reading) that has good reliability and validity as evidenced primarily by its technical characteristics and correlation with other tests (Renaissance Learning, Inc., 2000, 2006a, b). Scholastic offers a reading comprehension test (Scholastic Reading Inventory-Interactive: SRI-I) that assesses students' reading levels, tracks students' reading growth over time, and helps guide instruction according to students' needs (Scholastic, 2001a, b, 2006). The focus of this project was an analysis of the relationships between scores on STAR Reading, SRI-I, and the Florida Comprehensive Assessment Test (FCAT: Florida Department of Education, 2002; n.d.). Our work addressed the need for research examining the use of interim assessments for improving classroom practice answered and restarted by Roehrig et al. (2008) and others (Council of Chief State School Officers, n.d.; Perie et al., 2007).

Participants

The sample included a total of 1,077 students with complete assessment information. Of the participants, 514 (48%) were female and 563 (52%) were male. Slightly more than half of the students (53%) were African-American; Caucasian students were the second largest group (23%), and Hispanic students were the third largest group (19%); and, there were 29 (3%) Asian and 18 (2%) multi-racial students in the sample. A total of 580 (54%) were eligible for free or reduced price lunch program. Statistically similar distributions were evident across grades for gender (X^2 = 0.45, df = 2, p > .05), ethnicity (X^2 = 12.00, df = 8, p > .05), and free lunch status (X^2 = 2.98, df = 2, p > .05). Other information about the participants is summarized in Table 1.

Procedures

In early February, all students were administered the *STAR Reading Version 2.0* (STAR Reading: Renaissance Learning, Inc., 2000, 2006a, b) test in a three-week period. The majority of students were tested in a single week. Anyone who was absent or missed the first assessment was followed up during the next two weeks. All students available took the test and there were no special criteria for including or excluding them. In mid-February, all available students were administered the Scholastic Reading Inventory-Interactive (SRI-I: Scholastic, Inc. 2001a, b, 2006) over a two-week period. All students remained in their classroom for all tests and school personnel used laptop carts to complete the STAR Reading and SRI-I assessments. All students took Florida Comprehensive Assessment Test (FCAT: Florida Department of Education, 2002) in

May. Scores for the three reading measures were described and compared to address our research questions.

Measures

STAR Reading is a norm-referenced and criterion-referenced computer adaptive test that is available for students in grades 1-12; we used scaled scores for sixth, seventh, and eighth grade students in this research. The test is timed and usually takes less than 10 minutes to complete. Reading levels are provided relative to national norms which allow teachers to quickly determine appropriate student instructional level. Students who are offered the test are expected to have reading vocabulary of at least 100 words, which corresponds to the beginning reading skills level or above. The test consists of 25 items of multiple choices for all grades. Students of grades 1-2 are offered all 25 items of vocabulary-in-context, whereas students of grades 3 -12 are offered 20 items of vocabulary-in-context and five authentic text passages. The test is computer-adaptive; that is, if a student answers one item correctly then the next item will be of increasing difficulty. Conversely, if the student misses the right answer, then the next item will be of lesser difficulty. The STAR Reading 2.x and higher has 1,159 vocabulary-in-context items and 250 authentic text passage items. This makes it possible to use the test as a diagnostic tool to measure students' progress and administer the test to the same group of students five times a year without repeating the items.

According to the STAR Reading Technical Manual (Renaissance Learning, 2006b), each vocabulary-in-context item is a complete sentence that requires students to actually interpret meaning to identify the correct answer. The vocabulary-in-context section is also used to determine the initial difficulty level of authentic text passages. The test provides grade equivalent, normal curve equivalent, and scaled scores. Additionally, it provides information about the zone of proximal development which indicates the lowest and highest range a student can read. The test software can also generate reports for teachers and parents.

Salvia, Ysseldyke, and Bolt (2006) reported that the test-retest reliabilities of STAR Reading varied from .85 to .95 for scaled scores, and from .79 to .91 for instructional reading level. A total of 34,446 students were tested twice with the interval of about five days between the first and the second test. The validity was established by correlating STAR Reading to other standardized tests. It was found that STAR Reading scores correlate closely to the scores of other reading measures such as: California Achievement Test, Comprehensive Test of Basic Skills, Degrees of Reading Power, Gates-MacGinitie, Iowa Test of Basic Skills, Metropolitan Achievement Test, and Stanford Achievement Test. Some custom-built state tests were also used. They include such states as: Connecticut, Texas, Indiana, Tennessee, Kentucky, North Carolina, and New York (Salvia et al., 2006).

SRI-I is a computer-adaptive test that is designed to assess student's reading comprehension level with texts of increasing difficulty (Scholastic, Inc., 2006). It usually lasts 20-30 minutes. The test ends after enough questions have been answered to compute a Lexile score for the student; we used these scores for sixth, seventh, and eighth grade students in this research. Students can print and view their Recommended Reading reports. The test uses authentic written materials and usually consists of 20-25 questions but no more than 30. The test bank contains more than 4,500 questions, which allows creating a unique test each time. The test measures such reading comprehension skills as referring to

details in the passage, drawing conclusions, and making comparisons and generalizations.

SRI-I is administered to K-12 students and uses the Lexile Framework for Reading (Knutson, 2006; Schnick & Knickelbine, 2000; Scholastic, n.d.). The Lexile measure is criterion-referenced and indicates the reading level of a particular student and that student's reading growth. The Lexile Scale for SRI scores ranges from 0 to 1,700. Comparing to Grade Equivalent, Grade Levels and Lexile Levels can be represented as follows: Grade 1-200 to 400; Grade 2-300 to 600; Grade 3-500 to 800; Grade 4-600 to 900; Grade 5-700 to 1,000; Grade 6-800 to 1,050; Grade 7-850 to 1,100; Grade 8-900 to 1,150; Grade 9-1,000 to 1,200; Grade 10-1,010 to 1,205; Grade 11-1,050 to 1,210; and, Grade 12-1,075 to 1,275. The Lexile measure was associated with other standardized tests such as Stanford 9 (SAT9), the North Carolina End-of-Grade Test, Stanford Diagnostic Reading Test (SDRT). SRI-I has been administered to more than three million students of all grades over the last five years.

SRI-I test-retest reliability was .89 (Renaissance Learning Inc., 2000). Knutson (2006) reported test-retest correlations for grades 3-10 students ranging from .81 to .85. The test was administered first in the fall and then in the spring. It was also administered to second graders in the spring and then to third graders in the fall. Correlation in this case went down to .78. SRI-I criterion-related validity was determined by correlating both spring and fall SRI-I scores to the spring 2002 FCAT-SSS Reading scores. The fall-to-spring correlations for grades 3 through 10 were in the range of .71 to .76, whereas spring-to-spring correlations ranged from .75 to .82.

The FCAT in Reading consists of two parts: criterion-referenced tests (CRT) assessing selected benchmarks in reading from the Sunshine State Standards (SSS) and norm-referenced tests (NRT) in reading assessing individual student performance in regards to national standards. Multiple choice items are used for grades 3 through 10. Additional short response items are administered at grades 4, 8, and 10. For each grade, the reading scores range from 100 to 500 points. According to the 2004 assessment, internal consistency reliability on the reading test varied from .87 to .91. Criterion-related validity for the same year was determined by correlating FCAT-SSS reading scores with the FCAT-NRT (Stanford-9) scores. According to the 2004 assessment, the correlation between the two tests was in the range of .80 to .84. National percentile rank scores are also available for FCAT. We used current grade FCAT scaled scores in reading in our analyses.

Design and Data Analysis

We used a cross-sectional design to document similarities and differences within and between group performances and relationships between them on two progress monitoring assessments and end-of-grade achievement scores. Multivariate analysis of variance (MANOVA), multiple linear regression, Pearson Correlation, and predictive discriminant analysis were employed in statistical analyses.

Results

We report three types of outcomes. First, we provide descriptive and inferential findings to illustrate levels of performance across the measures between groups of students at each grade level participating in the study. Second, we report simple correlations across measures within grades. Third, we describe predictive analyses of relationships between progress monitoring assessments and statewide achievement test performance.

Descriptive Comparisons

Means and standard deviations for performance of sixth, seventh, and eighth grade students on reading measures across different demographic groups are in Table 1. Since FCAT scores are not comparable across grades, we completed a series of MANOVAs to document the extent reading performance was statistically similar for sixth, seventh, and eighth grade students on different measures of progress monitoring and end-of-grade achievement for different demographic groups.

Table 1
Means and Standard Deviations across Comparison Group

				Test					
				STAR		SRI-I		FCAT	
Grade	Group	Subgroup	n	Mean	SD	Mean	SD	Mean	SD
Six	Gender	Female	201	601.57	260.89	829.49	257.11	300.30	58.48
		Male	209	625.67	264.60	809.85	282.90	295.98	67.74
	Ethnicity	Asian	8	935.38	246.61	1095.13	211.70	383.75	66.21
		African American	233	558.28	237.20	772.02	259.93	288.42	58.90
		Hispanic	76	580.04	217.78	792.96	257.77	288.83	67.53
		Multi-racial	10	606.90	197.10	933.20	204.20	321.00	35.55
		Caucasian	83	770.70	295.76	936.70	272.85	322.73	61.37
	Lunch	FRL	224	588.89	228.63	802.79	245.88	292.53	58.68
		Non-FRL	186	643.93	296.56	839.58	296.71	304.81	68.06
	Total		410	613.86	262.74	819.48	270.41	298.10	63.33
Seven	Gender	Female	155	711.48	276.64	898.98	245.20	311.00	65.39
		Male	176	754.56	287.40	928.26	244.39	307.14	73.51
	Ethnicity	Asian	14	955.79	255.57	1095.50	148.12	347.57	41.52
		African American	166	643.29	241.53	865.36	229.55	293.11	70.11
		Hispanic	59	696.86	304.62	888.85	311.98	299.20	68.51
		Multi-racial	4	913.00	258.46	1077.50	139.78	345.25	35.68
		Caucasian	88	888.03	22.27	988.36	203.85	337.56	63.48
	Lunch	FRL	165	685.83	250.34	881.15	238.29	297.96	73.84
		Non-FRL	166	783.23	305.12	948.15	247.44	319.99	63.72
	Total		331	734.38	282.81	914.55	244.84	308.95	69.75
Eight	Gender	Female	158	779.95	298.26	953.72	264.60	312.78	48.03
		Male	178	804.10	293.82	912.21	273.55	299.12	58.83
	Ethnicity	Asian	7	1123.43	133.56	1187.14	119.31	358.00	11.86
		African American	175	677.93	244.97	848.77	247.58	284.91	49.46
		Hispanic	67	784.09	280.73	955.76	241.97	315.69	42.58
		Multi-racial	4	1041.50	318.46	1014.50	102.32	337.25	52.34
		Caucasian	83	1001.94	276.84	1061.72	283.78	334.92	44.18
	Lunch	FRL	190	747.91	290.17	910.52	262.81	298.32	52.59
		Non-FRL	146	851.10	293.66	959.33	277.05	314.95	49.20
	Total		336	792.74	295.72	931.73	269.77	305.54	51.56

Gender. Box's test of the assumption of equality of covariance matrices across gender was non-significant for sixth (M = 10.02, p > .01), seventh (M =

3.55, $p > .01$), and eighth ($M = 29.90$, $p < .01$) grades. Using Pillai' trace, there was a non-significant difference for gender on STAR Reading, SRI-I, and FCAT performance in sixth grade, $V = .02$, $F(3, 406) = 2.43$, $p > .01$, and seventh grade, $V = .02$, $F(3, 327) = 1.77$, $p > .01$. A statistically significant gender effect was indicated for eighth grade, $V = .06$, $F(3, 332) = 6.68$, $p < .01$; however, univariate follow-up tests revealed non-significant differences between eighth grade girls and boys for STAR Reading, $F(1, 334) = 0.56$, $p > .01$, SRI-I, $F(1, 334) = 1.64$, $p > .01$, and for FCAT, $F(1, 334) = 5.97$, $p > .01$.

Ethnicity. Box's test of the assumption of equality of covariance matrices across ethnicity was non-significant for sixth ($M = 27.25$, $p > .01$) and significant for seventh ($M = 55.95$, $p < .01$) and non-significant for eighth ($M = 47.29$, $p > .01$) grade. Using Pillai' trace, there was a significant difference for ethnicity on STAR Reading, SRI-I, and FCAT performance for sixth grade, $V = .16$, $F(12, 1215) = 1.61$, $p < .01$, for seventh grade, $V = .18$, $F(12, 978) = 5.28$, $p < .01$, and for eighth grade, $V = .28$, $F(12, 993) = 8.60$, $p < .01$.

Univariate follow-up tests revealed differences in STAR Reading scores, $F(4, 405) = 15.16$, $p < .01$, for sixth grade students from different ethnic backgrounds; scores for students from Asian ($M = 935.38$), Caucasian ($M = 770.70$), and multi-racial ($M = 606.90$) ethnic backgrounds were statistically different from their peers from Hispanic ($M = 580.04$) and African American ($M = 558.28$) ethnic backgrounds. Univariate follow-up tests revealed differences in SRI-I scores, $F(4, 405) = 10.47$, $p < .01$, for sixth grade students from different ethnic backgrounds; scores for students from Asian ($M = 73.13$), Caucasian ($M = 57.72$), and multi-racial ($M = 55.90$) ethnic backgrounds were statistically different from their peers from Hispanic ($M = 44.05$) and African American ($M = 42.66$) ethnic backgrounds. Univariate follow-up tests revealed differences in FCAT end-of-grade achievement scores, $F(4, 405) = 9.65$, $p < .01$, for sixth grade students from different ethnic backgrounds; scores for students from Asian ($M = 383.75$), Caucasian ($M = 322.73$), and multi-racial ($M = 321.00$) ethnic backgrounds were statistically different from their peers from Hispanic ($M = 288.83$) and African American ($M = 288.42$) ethnic backgrounds.

Univariate follow-up tests revealed differences in STAR Reading scores, $F(4, 326) = 16.09$, $p < .01$, for seventh grade students from different ethnic backgrounds; scores for students from Asian ($M = 955.79$), multi-racial ($M = 913.00$), and Caucasian ($M = 888.03$) ethnic backgrounds were statistically different from their peers from Hispanic ($M = 696.86$) and African American ($M = 643.29$) ethnic backgrounds. Univariate follow-up tests revealed differences in SRI-I scores, $F(4, 326) = 7.12$, $p < .01$, for seventh grade students from different ethnic backgrounds; scores for students from Asian ($M = 65.29$), multi-racial ($M = 63.00$), and Caucasian ($M = 55.13$) ethnic backgrounds were statistically different from their peers from Hispanic ($M = 47.32$) and African American ($M = 43.54$) ethnic backgrounds. Univariate follow-up tests revealed differences in FCAT scores, $F(4, 326) = 8.12$, $p < .01$, for seventh grade students from different ethnic backgrounds; scores for students from Asian ($M = 347.57$), multi-racial ($M = 345.25$), and Caucasian ($M = 337.53$) ethnic backgrounds were statistically different from their peers from Hispanic ($M = 299.20$) and African American ($M = 293.11$) ethnic backgrounds.

Univariate follow-up tests revealed differences in STAR Reading scores, $F(4, 331) = 25.77$, $p < .01$, for eighth grade students from different ethnic backgrounds;

scores for students from Asian (M = 1123.43), multi-racial (M = 1042.50), and Caucasian (M = 1000.94) ethnic backgrounds were statistically different from their peers from Hispanic (M = 784.09) and African American (M = 677.93) ethnic backgrounds. Univariate follow-up tests revealed differences in SRI-I scores, $F(4, 331) = 14.83$, $p < .01$, for eighth grade students from different ethnic backgrounds; scores for students from Asian (M = 69.57), Caucasian (M = 58.99), and multi-racial (M = 52.00) ethnic backgrounds were statistically different from their peers from Hispanic (M = 47.84) and African American (M = 38.80) ethnic backgrounds. Univariate follow-up tests revealed differences in FCAT scores, $F(4, 331) = 20.42$, $p < .01$, for eighth grade students from different ethnic backgrounds; scores for students from Asian (M = 358.00), multi-racial (M = 7.25), and Caucasian (M = 334.92) ethnic backgrounds were statistically different from their peers from Hispanic (M = 315.69) and African American (M = 284.91) ethnic backgrounds.

Federal lunch status. Box's test of the assumption of equality of covariance matrices across federal lunch status was significant for sixth (M = 21.58, $p < .01$) and non-significant for seventh (M = 15.73, $p > .01$) and eighth (M = 5.34, $p > .01$) grades. Using Pillai' trace, there was a non-significant difference for federal free lunch status on STAR Reading, SRI-I, and FCAT performance in sixth grade, $V = .01$, $F(3, 406) = 1.61$, $p > .01$, and a significant difference for seventh grade, $V = .04$, $F(3, 327) = 3.90$, $p < .01$ and eighth grade, $V = .04$, $F(3, 332) = 6.68$, $p < .01$. Univariate follow-up tests revealed significant differences for STAR Reading, $F(1, 329) = 10.09$, $p < .01$, between seventh grade students receiving free or reduced lunch (M = 685.83) and their peers not receiving free or reduced lunch (M = 783.23), for SRI-I, $F(1, 329) = 7.33$, $p < .01$, between seventh grade students receiving free or reduced lunch (M = 45.22) and their peers not receiving free or reduced lunch (M = 51.70), and for FCAT end-of-grade achievement, $F(1, 329) = 8.44$, $p < .01$, between seventh grade students receiving free or reduced lunch (M = 297.96) and their peers not receiving free or reduced lunch (M = 319.99). Univariate follow-up tests revealed significant differences for STAR Reading, $F(1, 334) = 10.33$, $p < .01$, between eighth grade students receiving free or reduced lunch (M = 747.91) and their peers not receiving free or reduced lunch (M = 851.10), for SRI-I, $F(1, 334) = 3.53$, $p < .01$, between eighth grade students receiving free or reduced lunch (M = 44.32) and their peers not receiving free or reduced lunch (M = 49.08), and for FCAT end-of-grade achievement, $F(1, 334) = 8.78$, $p < .01$, between eighth grade students receiving free or reduced lunch (M = 298.32) and their peers not receiving free or reduced lunch (M = 314.95).

Correlation Comparisons

Correlation coefficients for STAR Reading, SRI-I, and FCAT scores for sixth, seventh, and eighth grade students are in Table 2. Relationships were stronger for Grade 6 students than for their peers in Grade 7 or Grade 8. For Grade 6 students, the correlation coefficients were similar with each of the tests explaining about 57% of the variance of the others. For Grade 7 students, (1) STAR Reading explains about 50% of the variance of SRI-I and 53% of the variance of FCAT; (2) SRI-I explains about 50% of the variance of STAR Reading and 49% of the variance in FCAT; and (3) FCAT explains about 49% of the variance of SRI-I and 53% of the variance of STAR Reading. For Grade 8 students, (1) STAR Reading explains about 41% of the variance of SRI-I and 54% of the variance of FCAT;

(2) SRI-I explains about 41% of the variance of STAR Reading and 48% of the variance in FCAT; and (3) FCAT explains about 48% of the variance of SRI-I and 54% of the variance of STAR Reading.

Table 2
Correlation Coefficients for SRI-I, STAR Reading, and FCAT Scaled Scores across Grades

	STAR Reading	*SRI-I*	*FCAT*
Grade 6 (*n* = 410)			
STAR Reading	--	.76	.75
SRI-I		--	.76
FCAT			--
Grade 7 (*n* = 331)			
STAR Reading	--	.73	.61
SRI-I		--	.58
FCAT			--
Grade 8 (*n* = 336)			
STAR Reading	--	.67	.71
SRI-I		--	.68
FCAT			--

Note. All correlation coefficients are statistically significant at $p < .01$.

Concerned with the differences among the subgroups with regard to students' social economic status and ethnicity (Roehrig et al., 2008), we correlated the scaled scores for STAR Reading, SRI-I, and FCAT end-of-grade achievement for each group of participants by school lunch status and ethnicity. The correlation coefficients ranged from .707 to .754 for regular-lunch students; ranged from .708 to .729 for free/reduced price lunch students; ranged from .675 to .704 for African American; ranged from .732 to .761 for Hispanic; and ranged from .680 to .714 for Caucasian. The relationship between STAR Reading, SRI-I, and FCAT were found to be invariant across groups classified by student lunch status and ethnicity using the *t* test suggested by (Bruning & Kintz, 1997).

Predictive Comparisons

Criterion-referenced reading scaled scores were used in multiple linear regressions to provide additional estimates and predictors of FCAT scores. Standardized coefficients as well as partial correlation coefficients of the independent variables were compared to determine the best predictor of FCAT scores. Since STAR Reading and SRI-I are highly correlated with each other, multicollinearity was examined before each variable was entered into the regression. The variance inflation factor (VIF) was 2.047 and the tolerance value was 0.488. According to Myers (1990), multicollinearity would be of a concern if the VIF is larger than 10. According to Lynch (2003), multicollinearity is a problem when the sample size was small and the model had considerable error. In addition, Lynch (2003) further pointed out three classic symptoms of multicollinearity: (1) significant *F* without significant *t*-ratios, (2) wildly changing estimates when an additional/collinear variable was included in a model, and (3) the estimates of the coefficients were unreasonable. None of these occurred in the current data; therefore,

multicollinearity was not concerned in the following analyses.

We were interested in what variables were good predictors of FCAT scores. Student demographic information (gender, ethnicity, free/reduced price lunch status) as well as their performance on STAR and SRI-I were entered into the regression as independent variables with a stepwise regression procedure. Variables representing student demographic information were recoded into dichotomous variables: gender was coded as 0 and 1 where 1 represents male and 0 represents female. Student eligibility for free/reduced price lunch program was coded as 0 and 1 where 1 represents eligible for this program and 0 represents not eligible for this program. Student ethnicity was coded into three variables African American (1) or not (0), Caucasian (1) or not (0), and Hispanic (1) or not (0). Asians and multi-racial students were not included because the sample size for the interaction effects with these groups of students is extremely small (sometimes less than 1).

The interaction effect between gender and the variable representing African American ethnicity or not was the only statistically significant interaction noted: $t = -4.46$, $p < .001$. As a result, the model was tested for male and female students separately (Table 3) and STAR Reading was entered into the model first (Model 1). In Model 2, both STAR Reading and SRI-I were included. In Model 3, all three predictors (STAR, SRI-I, and African American ethnicity) were entered into the model. We used this hierarchical approach to examine the amount of variance that was explained by each variable while taking into account the variance already explained in previous models (e.g., how much additional variance can be explained by SRI-I when considering the variance that has been accounted for by STAR). For male students, $R^2 = .50$ for Model 1 when STAR was the only significant predictor; $R^2 = .57$ for Model 2 when SRI-I was also a significant predictor; $R^2 = .58$ for Model 3 when all three variables (STAR, SRI-I, and African American or not) are significant predictors. The change of R^2 was .50 for Model 1, .07 for Model 2, and .01 for Model 3. Each of the change of R^2 was statistically significantly different from zero. For female students, $R^2 = .57$ for Model 1 when STAR was the only significant predictor; $R^2 = .66$ for Model 2 when SRI-I was also a significant predictor. The change of R^2 was .57 for Model 1 and .09 for Model 2. Each of the change of R^2 was also statistically significantly different from zero.

Table 3
Stepwise Estimates of Coefficients for the Multiple Regressions on FCAT

		B	*SE B*	*β*	*t*	*p*	*Partial*	*Part*
Male								
Model 1	STAR	0.148	0.006	0.708	23.64	<.001	.708	.708
Model 2	STAR	0.092	0.008	0.439	11.041	<.001	.424	.307
	SRI-I	0.083	0.009	0.376	9.451	<.001	.372	.263
Model 3	STAR	0.087	0.008	0.415	10.207	<.001	.398	.283
	SRI-I	0.084	0.009	0.377	9.524	<.001	.375	.264
	Black	-8.902	3.541	-0.073	-2.514	.012	-.106	-.070
Female								
Model 1	STAR	0.144	0.006	0.753	25.865	<.001	.753	.753
Model 2	STAR	0.084	0.007	0.439	11.830	<.001	.464	.306
	SRI-I	0.092	0.008	0.438	11.800	<.001	.463	.305

Note. The adjusted R^2 for the final model is .573 for male students and .659 for female students.

The final model fits quite well, $F(3, 555) = 250.11$ for male students and $F(2, 509) = 494.80$ for female students. The adjusted R^2 value was .57 for male students and .66 for female students, suggesting that the percentage of the variance of FCAT that could be explained by the predictors was 57% for male students and 66% for female students. Differences were noted between male and female students: African American male students had a statistically significantly lower performance on FCAT in comparison to non-African American male students ($t = -2.51$, $p = .01$); however, this difference was not statistically significant for female students ($t = 0.10$, $p = .92$). All other variables representing student demographic information were excluded because they did not meet the inclusion criterion: Probability of *F*-to-enter is less than or equal to .05.

The estimates of the standardized coefficients are interpreted for male and female students, respectively. For male students, a unit increase in STAR Reading scores would result in 0.42 unit of increase in FCAT scores after controlling for SRI-I and student ethnicity of African American or not whereas a unit increase in SRI-I scores would result in 0.38 unit of increase in FCAT scores after controlling for STAR Reading and student ethnicity of African American or not. The partial correlation coefficients indicated that a unit increase in STAR Reading scores would result in 0.40 unit of increase in FCAT scores after removing the linear effect of SRI-I scores on both FCAT and STAR Reading. Similarly, a unit increase in SRI-I scores would result in 0.38 unit of increase in FCAT scores after removing the linear effect of STAR Reading scores on both FCAT and SRI-I. The part correlation coefficients indicated that a unit increase in STAR Reading scores would result in 0.28 unit of increase in FCAT scores after removing the linear effect of SRI-I scores on STAR Reading only.

Similarly, a unit increase in SRI-I scores would result in 0.26 unit of increase in FCAT scores after removing the linear effect of STAR Reading scores on SRI-I only. For female students, a unit increase in STAR Reading scores would result in 0.44 unit of increase in FCAT scores after controlling for SRI-I whereas a unit increase in SRI-I scores would result in 0.44 unit of increase in FCAT scores after controlling for STAR Reading. The partial correlation coefficients indicated that a unit increase in STAR Reading scores would result in 0.46 unit of increase in FCAT scores after removing the linear effect of SRI-I scores on both FCAT and STAR Reading. Similarly, a unit increase in SRI-I scores would result in 0.46 unit of increase in FCAT scores after removing the linear effect of STAR Reading scores on both FCAT and SRI-I. The part correlation coefficients indicated that a unit increase in STAR Reading scores would result in 0.31 unit of increase in FCAT scores after removing the linear effect of SRI-I scores on STAR Reading only. Similarly, a unit increase in SRI-I scores would result in 0.31 unit of increase in FCAT scores after removing the linear effect of STAR Reading scores on SRI-I only. All these coefficient estimates suggested that both STAR Reading and SRI-I assessments are good predictors of FCAT.

Finally, to assess the accuracy of prediction, predictive discriminant analysis (PDA) was employed to measure how well STAR and SRI-I predicted the students' performance on FCAT based upon FCAT achievement levels. Participants were put into two groups according to the technical report of Florida Center for reading Research (Buck & Torgesen, n.d.): adequate (Levels 3-5) and inadequate (Levels 1-2). True positive (TP), true negative (TN), false positive (FP), and false negative (FN) were counted based upon the results of PDA. TP refers to students who did not master the skill and were predicted as not having mastered the skill. TN refers to students who mastered the skill and were predicted as having mastered the skill. FP refers to students who had mastered the skill but were predicted as not having mastered the skill. FN refers to students who did not master the skill but were predicted as having mastered the skill. Hit rate, sensitivity, and specificity indices were calculated for each PDA to reflect the accuracy of PDA. Hit rate provides an overall indication of how well STAR and SRI-I predicted students' performance on FCAT, sensitivity reflects how well STAR and SRI-I identified students who did not master the skills measured by FCAT, and specificity suggests how well STAR and SRI-I identified students who mastered the skills measured by FCAT. The formulas to calculate these indices were as follows:

$$\text{Hit rate} = \frac{TP + TN}{N};\ \text{Sensitivity} = \frac{TP}{TP + FN};\ \text{and Specificity} = \frac{TN}{TN + FP}\ \text{(Hosp \& Fuchs, 2005).}$$

Cut-off scores for adequate or inadequate mastery of skills are suggested by the manuals for each test (Florida Department of Education, 2002; Renaissance Learning, Inc., 2006b; Scholastic, Inc., 2006): FCAT (296 for sixth graders, 300 for seventh graders, and 310 for eighth graders); STAR (638 for sixth graders, 781 for seventh graders, and 878 for eighth graders); SRI-I (800 for sixth graders, 850 for seventh graders, and 900 for eighth graders). Results in Table 4 indicate that both STAR and SRI-I are accurate in predicting students' performance on FCAT (the average of the

hit rate is 76% across the grades). Specifically, STAR and SRI-I are more accurate in predicting students who are considered "adequate" by FCAT (the average of the specificity is 88% across grades) than predicting students who are considered "inadequate" by FCAT (the average of the sensitivity is 70% across grades).

Table 4
Hit rates, Sensitivity, and Specificity Indices for STAR and SRI-I Predicting FCAT Mastery

Reading Skill	TP	FP	TN	FN	Hit Rate (%)	Sensitivity (%)	Specificity (%)
Grade 6 (*n* = 410)							
STAR	187	15	133	75	78	71	90
SRI-I	201	1	87	121	70	62	99
Grade 7 (*n* = 326)							
STAR	113	9	122	82	72	58	93
SRI-I	73	49	169	35	74	68	78
Grade 8 (*n* = 335)							
STAR	160	12	116	47	82	77	91
SRI-I	124	48	141	22	79	85	75

Note. TP = true positive; FN= false negative; TN = true negative; FP = false positive.
Hit rate = (*TP* + *TN*)/*n*; sensitivity = *TP*/(*TP* + *FN*); specificity = *TN*/(*TN* + *FP)*

Discussion

Data-driven accountability has reached a new level under mandates and directives in the NCLB Act of 2004 (U.S. Department of Education, 2004). State and local education agencies are making school personnel test students regularly with screening, diagnostic, progress-monitoring, and high stakes outcome measures (Hasbrouck & Tindal, 2006). As Crawford, Tindal, and Stieber (2001) indicate, the widespread adoption of statewide tests as markers of academic proficiency and an implied link to school quality have made it important that students' academic *progress* be closely monitored for at least four reasons:

> 1. Statewide testing programs often involve a format that is difficult for teachers to replicate at the classroom level.
> 2. Decisions being made are so important that other confirming information is needed to complement the data.
> 3. Teachers need other performance indicators related to statewide tests that are available more frequently so that instructional programs can be improved in a timely fashion.
> 4. Statewide tests may be insensitive to change for low-performing students. (p. 304)

Interest in relationships between progress monitoring measures and high stakes achievement tests has a long and renewing history (Council of Chief State School Officers, n.d.; Coyne & Harn, 2006; Crawford et al., 2001; Deno, Mirkin, Chiang, & Lowry, 1980; Fuchs & Deno, 1981; Linn, 2000; Marston, Deno, & Tindal, 1983; Perie et al., 2007; Roehrig et al., 2008; Schatschneider et al., 2004; Schilling et al., 2007; Sibley et al., 2001; Stecker & Fuchs, 2000).

The purpose of the current study was to examine levels of performance on and relationships between performance on two progress monitoring measures and a statewide end-of-grade achievement test. We reported and compared performances on the STAR Reading, SRI-I, and FCAT end-of-grade achievement across grades, lunch status, and ethnicity. We also examined relationships between and among these measures. Our work adds to extant knowledge by responding to the need for research evaluating the usefulness of monitoring learning progress and predicting high stakes performance as important "social consequences" and by extending prior research on these relationships beyond elementary school students and measures of oral reading fluency (Roehrig et al., p. 362).

Education policy and practice driven by the NCLB Act requires school personnel to disaggregate data on school outcomes by race; and, it is easy to find evidence in government and other reports illustrating significant differences between test scores and other indicators of educational engagement and success for students representing some ethnic minorities and their Caucasian peers (cf. Fang, 2010; Ladson-Billings, 2006; National Center for Education Statistics, 2001, 2007, 2009; Uzzell, Simon, Horwitz, Hyslop, Lewis, & Casserly, 2010; Warikoo & Carter, 2009; Wiggan, 2007). Our comparisons of the performance of sixth, seventh, and eighth grade students on different measures of reading achievement supported these widely-recognized trends. We also documented that two widely-used progress monitoring assessments were good predictors of end-of-grade achievement at three different grade levels and that predictive bias across different demographic groups was minimal. In general, our correlational findings were similar to those of other researchers using different progress monitoring and outcome measures with younger children (cf. Hintze et al., 2002, 2003; Hixson & McGlinchey, 2004; Roehrig et al., 2008).

Implications for Improvement of Practice

We agree with Roehrig et al. (2008) that "it is essential that educators be provided with precise student achievement data and benchmarks if the rigorous grade-level reading standards set forth in accountability policies are to be met by all students" (p. 362). Our findings indicated that STAR Reading and SRI-I were good predictors of end-of-grade achievement in grades 6, 7, and 8, and the usefulness of our findings was evident across different demographic groups. These outcomes were previously unavailable in research on relationships between progress monitoring measures and statewide assessment outcomes completed in elementary schools in several states (cf. Barger, 2003; Buck & Torgesen, n.d.; Buck, Torgesen, & Schatschneider, n.d.; Hintze et al., 2002, 2003; Vander Meer et al., 2005; Wilson, 2005).

Because our finding of similar predictive relationships had limited practical value in making decisions about which progress monitoring measure to use, we conducted a *post hoc* analysis. According to marketing materials provided by Scholastic, the SRI-I usually takes 20-30 minutes to administer. Administration times were available in the data dictionary provided by the participating schools for STAR Reading but not for the SRI-I. We found that STAR Reading took an average of about seven minutes to administer. Most administrators and teachers believe that despite the value of regularly monitoring student progress there is too much time spent testing or preparing for tests. In our research, a test that required about eight minutes per child to administer did the same job predicting end-of-grade

performance compared with another test that developers believe would take three times as long. Accepting conservative estimates of test administrations every other month, using STAR Reading for progress monitoring would free over 20 minutes per student for critical instructional skills many teachers believe are trumped by testing; and, saving would be greater in schools doing monthly or more frequent progress monitoring.

Conclusion

Although our findings of consistent achievement differences across some demographic groups of middle school students and strong predictive relationships for two different progress monitoring measures and statewide end-of-grade achievement provide support for prior research as well for continued use in school-based decision making, we agree with Roehrig et al. (2008), Wiggan (2007), and Warikoo and Carter (2009) that more research is needed. Areas of clear extension for our study include investigations of similarities and differences in performance as well as relationships between scores on widely-used progress monitoring assessments and high stakes state-wide tests used in elementary schools. Research building on our finding of the potential differential benefits of progress monitoring systems with similar predictive capacity are also warranted and recommended.

Each year millions of students are at risk for serious and continued failure in school and many fail to make acceptable progress, especially when compared to their peers across different demographic groups (National Center for Education Statistics, 2001, 2007, 2009; U.S. Department of Education, 2006). We documented that the persistent and consistent differences evident for reading achievement in elementary school continue into middle school for many students. We also found that progress monitoring measures administered during the school year predicted end-of-year performances very well. We believe that the value of documenting achievement and predictions of it is not in the magnitude of the differences or relationships that are revealed but in deriving direction for change from them. The best predictor of achievement in elementary school is prior performance in elementary school (Roehrig et al., 2008; Schilling et al., 2007; Wanzek et al., 2010); and, the best predictor of performance in middle school is performance in elementary school (Fang, 2010). Progress monitoring measures administered during sixth, seventh, and eighth grade were found to be strong predictors of end-of-grade achievement. Gaps on two formative measures of achievement for students from Asian and Caucasian ethnic backgrounds and their peers from Hispanic and African American ethnic backgrounds were also evident in summative end-of-grade achievement scores for these students. The implications for improving practice are clear: Continued use of progress monitoring measures such as STAR Reading are powerful tools in efforts to identify students needing assistance to persist and affect high stakes assessments.

References

Ananda, S., & Rabinowitz, S. (2001). *High stakes assessment innovation: A negative correlation?* San Francisco, CA: WestEd. (Ed 462 446).

Barger, J. (2003). *Comparing the DIBELS Oral Reading Fluency indicator and the North Carolina end of grade reading assessment (Technical Report)*. Asheville, NC: North Carolina Teacher Academy. Retrieved from https://dibels.uoregon.edu/techreports/NC_Tech_Report.pdf

Brown, R. S., & Coughlin, E. (2007). *The predictive validity of selected benchmark assessments used in the Mid-Atlantic Region* (Issues & Answers Report, REL 2007–No. 017). Washington, DC: U.S. Department of Education, Institute of Education Sciences, National Center for Education Evaluation and Regional Assistance, Regional Educational Laboratory Mid-Atlantic. Retrieved from http://ies.ed.gov/ncee/edlabs

Bruning, J. L., & Kintz, B. L. (1997). *Computational handbook of statistics.* New York, NY: Longman.

Buck, J., & Torgesen, J. (n.d.). The relationship between performance on a measure of oral reading fluency and performance on the Florida Comprehensive Assessment Test. FCRR Technical Report #1. Tallahassee, FL: Florida Center for Reading Research. Retrieved from http://www.fcrr.org/TechnicalReports/TechnicalReport1.pdf

Buck, J., Torgesen, J., & Schatschneider, C. (n.d.). Predicting FCAT-SSS scores using prior performance on the FCAT-SSS, FCAT-NRT, and SAT9. (FCRR Technical Report #4). Tallahassee, FL: Florida Center for Reading Research. Retrieved from http://www.fcrr.org/TechnicalReports/fcat_predicting.pdf

Council of Chief State School Officers (CCSSO). (n.d.). *CCSSO and Renaissance Learning R & D Consortium Using Interim Assessments to Predict Student Proficiency*. Washington, DC: Author.

Coyne, M. D., & Harn, B. A. (2006). Promoting beginning reading success through meaningful assessment of early literacy skills. *Psychology in the Schools*, *43*, 33-43.

Crawford, L., Tindal, G., & Stieber, S. (2001). Using oral reading rate to predict student performance on statewide achievement tests. *Educational Assessment*, *7*, 303-323.

Deno, S. L. (1992). The nature and development of curriculum-based measurement. *Preventing School Failure, 35*(2), 5-10.

Deno, S. L. (1997). Whether thou goest . . . Perspectives on progress monitoring. In J. W. Lloyd, E. J. Kameenui, & D. Chard (Eds.), *Issues in educating students with disabilities* (pp. 77-99). Mahwah, NJ: Erlbaum.

Deno, S. L. (2003). Developments in curriculum-based measurement. *Journal of Special Education, 37,* 184-192.

Deno, S. L., & Mirkin, R K. (1977). *Data-based program modification: A manual.* Reston, VA: Council for Exceptional Children.

Deno, S. L., Mirkin, P. K., Chiang, B., & Lowry, L. (1980). *Relationships among simple measures of reading and performance on standardized achievement tests* (Res. Rep. No. 20). Minneapolis: University of Minnesota, Institute for Research on Learning Disabilities.

Fang, L. (2010). Are boys left behind? The evolution of the gender achievement gap in Beijing's middle schools. *Economics of Education Review, 29,* 383-399.

Florida Department of Education. (2002). *Florida Comprehensive Assessment Test (FCAT) for Reading and Mathematics: Technical report for test administrations of FCAT 2002*. Tallahassee, FL: Florida Department of Education.

Florida Department of Education. (n.d.). *Florida Comprehensive Assessment Test*. Retrieved from http://fcat.fldoe.org/

Fuchs, D., & Deshler, D. D. (2007). What we need to know about responsiveness to intervention (and shouldn't be afraid to ask). *Learning Disabilities Research and Practice, 22*, 129-136.

Fuchs, L. S. (1989). Evaluating solutions: Monitoring progress and revising intervention plans. In M. Shinn (Ed.), *Curriculum-based measurement: Assessing special children* (pp. 153-181). New York, NY: Guilford.

Fuchs, D., & Fuchs, L. S. (2006). Introduction to Response to Intervention: What why, and how valid is it? *Reading Research Quarterly, 41*, 93-99.

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

Fuchs, L., & Deno, S. (1981). *The relationship between curriculum-based mastery measures and standardized achievement tests in reading* (Res. Rep. No. 57). Minneapolis: University of Minnesota, Institute for Research on Learning Disabilities.

Fuchs, L. S., & Fuchs, D. (1986). Effects of systematic formative evaluation on student achievement: A meta-analysis. *Exceptional Children, 53*, 199-208.

Fuchs, L. S., & Fuchs, D. (1999). Monitoring student progress toward development of reading competence: A review of three forms of classroom-based assessment. *School Psychology Review, 28*, 659-671.

Hasbrouck, J., & Tindal, G. A. (2006). Oral reading fluency norms: A valuable assessment tool for reading teachers. *The Reading Teacher, 59*, 636-644.

Hintze, J. M., Callahan, J. E., III, Matthews, W. J., Williams, S. A. S., & Tobin, K. G. (2002). Oral reading fluency and prediction of reading comprehension in African American and Caucasian elementary school children. *School Psychology Review, 31*, 540-553.

Hintze, J. M., Ryan, A. L., & Stoner, G. (2003). Concurrent validity and diagnostic accuracy of the dynamic indicators of basic early literacy skills and the comprehensive test of phonological awareness. *School Psychology Review, 32*, 541-556.

Hixson, M. D., & McGlinchey, M. T. (2004). The relationship between race, income, and oral reading fluency and performance on two reading comprehension measures. *Journal of Psychoeducational Assessment, 22*, 351-364.

Hosp, M. K., & Fuchs, L. S. (2005). Using CBM as an indicator of decoding, word reading, and comprehension: Do the relations change with grade? *School Psychology Review, 34*, 9-26.

Knutson, K. (2006). *Because you can't wait until spring: using the SRI to improve reading performance*, Scholastic Professional Paper. New York, NY: Scholastic Inc.

Ladson-Billings, G. (2006). From the achievement gap to the education debt: Understanding achievement in U.S. schools. *Educational Researcher, 35*, 3–12.

Linn, R. L. (2000). Assessments and accountability. *Educational Researcher, 29*, 4-16.

Lynch, S. M. (2003). *Multicollinearity.* Retrieved from http://www.princeton.edu/~slynch/SOC_504/multicollinearity.pdf

Madelaine, A. & Wheldall, K. (2004). Curriculum-based measurement of reading: Recent advances. *International Journal of Disability, Development and Education, 51*, 57-82.

Marston, D., Deno, S., & Tindal, G. (1983). *A comparison of standardized achievement tests and direct measurement techniques in measuring pupil progress* (Res. Rep. No. 126). Minneapolis: University of Minnesota, Institute for Research on Learning Disabilities. (ERIC Document Reproduction Service No. ED236198)

Myers, R. (1990). *Classical and modern regression with applications (2nd ed.).* Boston, MA: Allyn & Bacon.

National Center for Education Statistics. (2001). *Education achievement and Black–White inequality.* Washington, DC: Department of Education.

National Center for Education Statistics. (2007). *Status and trends in the education of racial and ethnic minorities.* Washington, DC: Department of Education.

National Center for Education Statistics. (2009). *The Nation's Report Card: Reading 2009*. Washington, DC: Department of Education.

Perie, M., Marion, S., & Gong, B. (2007). *A framework for considering interim assessments*. Dover, NH: National Center for the Improvement of Educational Assessment. Retrieved from http://www.nciea.org/publications/ConsideringInterimAssess_MAP07.pdf

Renaissance Learning, Inc. (2000). *Comparison of the STAR Reading Computer-Adaptive Test and the Scholastic Reading Inventory Test*. Wisconsin Rapids, WI: Author.

Renaissance Learning, Inc. (2006a). *STAR Reading Computer-Adaptive Reading Test and Database: Software manual*. Wisconsin Rapids, WI: Author.

Renaissance Learning, Inc. (2006b). *STAR Reading Computer-Adaptive Reading Test and Database: Technical Manual*. Wisconsin Rapids, WI: Author.

Roehrig, A. D., Petscher, Y., Nettles, S. M., Hudson, R. F., & Torgesen, J. K. (2008). Accuracy of the DIBELS Oral Reading Fluency measure for predicting third grade reading comprehension outcomes. *Journal of School Psychology, 46*, 343-366.

Safer, N. & Fleischman, S. (2005). How student progress monitoring improves instruction. *Educational Leadership, 62*, 81-83.

Salvia, J., Ysseldyke, J. E., & Bolt, S. (2006) *Assessment in special and inclusive education*. Boston, MA: Houghton Mifflin Company.

Schatschneider, C., Buck, J., Torgesen, J., Wagner, R., Hassler, L., Hecht, S., & Powell-Smith, K. (2004). A multivariate study of individual differences in performance on the reading portion of the Florida Comprehensive Assessment Test: A brief report. Tallahassee, FL: Florida Center for Reading Research.

Schilling, S. G., Carlisle, J. F., Scott, S. E., & Zeng, J. (2007). Are fluency measures accurate predictors of reading achievement? *The Elementary School Journal*, *107*, 429-448.

Schnick, T., & Knickelbine, M. (2000). *The lexile framework: An introduction for educators*. Durham, NC: MetaMetrics.

Scholastic, Inc. (2001a). *Scholastic reading inventory interactive educator's guide*. New York, NY: Author.

Scholastic, Inc. (2001b). *Scholastic Reading Inventory: Technical Guide*. New York, NY: Author.

Scholastic, Inc. (2006). *Scholastic Reading Inventory Educator's Guide: An Overview of SRI Software and Teacher Support*. New York, NY: Author.

Scholastic, Inc. (n.d.). Accuracy matters: Reducing measurement error by targeted. New York, NY: Author. Retrieved from http://teacher.scholastic.com/products/sri/pdfs/SRI_Accuracy_Professional_Paper.pdf

Scott, V. G., & Weishaar, M. K. (2003). Curriculum-based measurement for reading progress. *Intervention in School and Clinic, 38,* 153–158.

Sibley, D., Biwer, D., & Hesch, A. (2001, April). *Establishing curriculum-based measurement oral reading fluency performance standards to predict success on local and state tests of reading achievement.* Paper presented at the Annual Meeting of the National Association of School Psychologists, Washington, DC.

Simpson, R. L., LaCava, P., & Graner, P. (2004). The No Child Left Behind Act: Challenges and implications for educators. *Intervention in School and Clinic, 40,* 67–75.

Stecker, P. M., & Fuchs, L. (2000). Effecting superior achievement using curriculum-based measurement: The importance of individual progress monitoring. *Learning Disabilities Research and Practice, 15,* 128-134.

Thurlow, M. L., & Thompson, S. J. (1999). District and state standards and assessments: Building an inclusive accountability system. *Journal of Special Education Leadership, 12,* 3–10.

Thurlow, M. L., & Wiley, H. I. (2004). *Almost there in public reporting of assessment results for students with disabilities* (Technical Report 39). Minneapolis: University of Minnesota, National Center on Educational Outcomes.

U.S. Department of Education. (2006). Proven methods: Early reading first and reading first. Retrieved from http://www.ed.gov/nclb/methods/reading/readingfirst.html

U.S. Department of Education, Office of the Deputy Secretary. (2004). *No Child Left Behind: A toolkit for teachers.* Washington, DC: Author.

Uzzell, R., Simon, C., Horwitz, A., Hyslop, A., Lewis, S., & Casserly, M. (2010). *Beating the odds: Analysis of student performance on state assessments and NAEP. Results from the 2008-09 School Year*. Washington, DC: Council of the Great City Schools.

Vander Meer, C. D., Lentz, F. E., & Stollar, S. (2005). *The relationship between oral reading fluency and Ohio proficiency testing in reading* (Technical Report). Eugene, OR: University of Oregon. Retrieved from https://dibels.uoregon.edu/techreports/ohio.pdf

Wanzek, J., Roberts, G., Linan-Thompson, S., Vaughn, S., Woodruff, A. L., & Murray, C. S. (2010). Differences in the relationship of oral reading fluency and high-stakes measures of reading comprehension. *Assessment for Effective Intervention, 35,* 67-77

Warikoo, N., & Carter, P. (2009). Cultural explanations for racial and ethnic stratification in academic achievement: A call for new and improved theory. *Review of Educational Research, 79,* 366-394.

Wiggan, G. (2007). Race, school achievement, and educational inequality: Toward a student-based inquiry perspective. *Review of Educational Research, 77,* 310-333.

Wilson, J. (2005). *The relationship of Dynamic Indicators of Basic Early Literacy Skills (DIBELS) oral reading fluency to performance on Arizona Instrument to Measure Standards (AIMS)* (Technical Report). Tempe, AZ: Assessment and Evaluation Department, Tempe School District No.3. Retrieved from https://dibels.uoregon.edu/techreports/arizona.pdf

Wixson, K. K., & Carlisle, J. F. (2005). The influence of large-scale assessment of reading comprehension on classroom practice: A commentary. In S. G. Paris & S. A. Stahl (Eds.), *Children's reading comprehension and assessment* (pp. 395-405). Mahwah, NJ: Erlbaum.