The Mathematics Educator
2010, Vol. 20, No. 1, 22-34

Reading Grade Levels and Mathematics Assessment: An Analysis of Texas Mathematics Assessment Items and Their Reading Difficulty[1]

John H. Lamb

Increased reading difficulty of mathematics assessment items has been shown to negatively affect student performance. The advent of high-stakes testing, which has serious ramifications for students' futures and teachers' careers, necessitates analysis of reading difficulty on state assessment items and student performance on those items. Using analysis of covariance, this study analyzed the effects of reading grade level of mathematics assessment items on student performance on the Texas Assessment of Knowledge and Skills. Results indicated that elementary and middle school students performed significantly worse on mathematics assessment items having a reading level above the student grade level. The implications of these results are discussed.

Rubenstein (2000) stated, "[Teachers of mathematics] want [students] to speak the language of mathematics, using standard terms that others recognize and understand" (p. 243). Boero, Couek, and Ferrari (2008) noted that the language of mathematics requires a mastery of one's natural language, both words and structures, in order to incorporate this language within the context of mathematical syntax. They illustrated how teachers of mathematics must go beyond simple classroom discussions in order to promote their students' mastery of the language of mathematics.

Carter and Dean (2006) illustrated how 5th through 11th grade mathematics teachers spend a considerable amount of time teaching reading strategies such as decoding written language into spoken words, understanding and discovering the meaning of new vocabulary, and making connections between the written language and the learner's prior knowledge. Out of 72 mathematics lessons they observed, nearly 70% of the implemented reading strategies addressed vocabulary. According to Lager (2006), "Without a strong command of both everyday language and specialized mathematical language students cannot fully access the mathematics content of the text, lesson, or assessment" (p. 194). Teachers of mathematics are faced with the challenge to not only prepare their students to successfully understand mathematical concepts but to also to prepare students to read and comprehend technically dense, descriptive mathematics problems.

John H. Lamb is an Assistant Professor of Mathematics Education at the University of Texas at Tyler. His research interests include interdisciplinary connections in mathematics, high stakes testing, and effective instructional practices in mathematics education.

Student Difficulties in Reading Related to Mathematics Achievement

Evidence has shown that a student's level of reading proficiency can be a strong indicator of mathematical success (Jiban & Deno, 2007). The correlation between reading and mathematics achievement has been well documented over the last five decades (e.g., Breen, Lehman, & Carlson, 1984; Fuchs, Fuchs, Eaton, Hamlett, & Karns, 2000; Helwig, Rozek-Tedesco, Tindal, Heath, & Almond, 1999; Jerman & Mirman, 1974; Pitts, 1952; Thompson, 1967). Evidence suggests that students without a strong ability to read and with difficulties in mathematics struggle more to be as successful as they could be in mathematics when compared to students only having difficulties in mathematics (Jordan, Hanich, & Kaplan, 2003). This situation is exacerbated for students with limited English proficiency (LEP).

Many students with LEP have accommodations to help account for students' academic difficulties related to their language deficiency. Accommodations may include extended time to complete assessments or

[1] *This paper was presented at the Conference for the Advancement of Mathematics Teaching on July, 2009 in Houston, TX. I would like to thank Dr. Ross Sherman, Mrs. Cindy Sherman, and Dr. Dennis Combs for their advice and review of early drafts of this work.*

assignments and having test questions read aloud. In Oregon, 6th-grade students with low reading ability performed better on mathematical problem solving assessment items when each assessment item was projected on a video monitor while a recorded narration of the written portions of the assessment items was played (Helwig et al., 1999). Fuchs et al. (2000) conducted a quasi-experimental study that randomly assigned students with and without learning disabilities (LD) to treatment accommodations that included extended time, calculator use, having questions read aloud, and encoding (when the teacher writes the student's responses). Fuchs et al. found these accommodations significantly benefited students with LD on their achievement in mathematical problem solving. Ketterlin-Geller, Yovanoff, and Tindall (2007) found that students with lower reading abilities scored better on linguistically and mathematically difficult assessment items when the questions were read aloud. Bolt and Thurlow (2007) found that reading questions aloud to students had a positive effect on 4th-grade student performance on mathematical assessment items with challenging text. For 8th-grade students, however, this positive effect was not evident. As this study demonstrates, there are inconsistencies in research conclusions that both reinforce and challenge the validity of such accommodations (Ketterlin-Geller, Yavonoff, & Tindal, 2007).

Reading Difficulty of Assessment Items Related to Mathematics

When students have difficulties in reading, research has indicated that they also struggle in mathematics. The way mathematics assessment items are written may influence the mathematical achievement of these poor readers. In 1967, Thompson generated two sets of mathematically similar test items written at different reading levels. He found that sixth grade students performed significantly better on the assessment items written at a lower reading level. Jerman and Mirman (1974) correlated several measures of readability with student performance on mathematics assessment items and found that the higher the word count, character count, sentence count, syllable count, word length, and sentence length of an assessment item, the lower the student performance. Similarly, Walker, Zhang, and Surber (2008) found that the reading difficulty of mathematics assessment items significantly lowered student performance. Hence, research conducted in recent decades has indicated mathematics assessment items and how they are written can have an effect on student performance. One may question how connected these two subjects are in terms of their predictive power towards one another.

Predictive Power of Reading Ability on Mathematics Achievement

Research indicates that students with difficulties in reading tend to have lower achievement in mathematics (e.g., Fuchs et al., 2000; Jiban & Deno, 2007; Jordan et al., 2003; Reikerâs, 2006). This supports the research indicating a strong correlation between reading and mathematics achievement (e.g., Breen, Lehman, & Carlson, 1984; Pitts, 1952). Evidence also exists for a connection between higher reading levels of mathematics assessment items and decreased student performance (e.g., Jerman & Mirman, 1974; Thompson, 1967; Walker et al., 2008). With the available data indicating reading affects mathematical performance, several states have implemented quantitative measures that utilize students' reading scores as predictive variables in their mathematics performance. The push for these states to identify predictive models came from the United States Department of Education (USDOE) Growth Model Pilot Program initiated in 2005 (USDOE, 2008). Twenty-three states have submitted applications, and as of January 2009, 15 states, including Texas, have been fully or conditionally approved (USDOE, 2009). The Texas Education Agency (TEA, 2009) established a Texas Projection Measure (TPM) that generated student expectant scores on state mathematics assessments based on reading and mathematics scores from previous end-of-year assessments. The TPM is designed to strengthen TEA's measure of Adequate Yearly Progress (AYP) for the No Child Left Behind (NCLB) law by generating a measure that assesses growth. This growth is evaluated using predictive measures for each child's 5th-, 8th-, and 11th-grade achievement level. The TPM projects a student's performance in 5th-grade reading and mathematics by utilizing 3rd- and 4th-grade reading and mathematics performance results. The TPM predicts 8th-grade performance by using 5th-, 6th-, and 7th-grade results and 11th-grade performance by using 8th-, 9th-, and 10th-grade results. TEA found that nearly 1% of the variance in mathematics was accounted for by the student's previous reading performance and up to 5.2% of the variance in reading was accounted for by the student's previous mathematics performance. TEA generated these results by finding the difference in explained variance with and without the other subject as a predictive variable. TEA showed that this percentage of explained variance was significant and

highlights the likelihood that mathematics and reading abilities are highly intertwined.

The Ohio Department of Education (ODE, 2007), also using a USDOE-approved growth-model with a regression-based value-added design of assessing student growth , found results similar to TEA (2009), identifying strong correlations and covariance between reading and mathematics achievement. In 2003, nearly two years prior to the USDOE's growth-model pilot program, Mississippi proposed and implemented an accountability model that predicted future performance in mathematics based on students' previous English, reading, and mathematics results on state assessments (Mississippi Department of Education, 2008). States across the country have illustrated, both in action and through quantitative measures, that reading and mathematics achievement are connected.

Purpose of This Study

The purpose of this study was to determine whether the reading grade level (RGL) of mathematics assessment items had a significant effect on 3rd- through 11th-grade student performance in the state of Texas. RGL of a mathematics assessment item was operationally defined as the approximate grade level a student was expected to obtain in order to comprehend a reading passage within the item. Based on the available research, I hypothesized that student performance is negatively affected by the RGL of mathematics assessment items on the state mandated Texas Assessment of Knowledge and Skills (TAKS).

To test this hypothesis, I analyzed assessment items and utilized available item analysis data from the 2006 TAKS. Several extraneous variables were identified and used as covariates in order to isolate the effect RGL might have on student performance. The three research questions in this study were:

1. Do students in grades 3 through 11 perform better on mathematics assessment items that are written with a RGL At-or-Below their grade level than on assessment items written with a RGL above their grade level?
2. Do students in specific grade bands perform better on mathematics assessment items that are written with a RGL At-or-Below their grade level than on assessment items written with a RGL above their grade level?
3. Do students regardless of their grade or grade band perform better on mathematics assessment items that are written with a RGL At-or-Below their grade level than on assessment items written with a RGL above their grade level?

Methods

This study used the Spring 2006 TAKS 3rd-11th grade released tests and item analysis from TEA (2008a). The TAKS objective and percentage of Texas students who correctly responded for each assessment item was collected for 438 test items. Of the six TAKS objectives in items for grades 3 through 8, five were aligned with the NCTM (2000) content standards, and the sixth objective was aligned to the NCTM process standard, problem solving. Ten 9th through 11th grade TAKS objectives incorporated algebraic, geometric, and problem solving objectives. Tables 1 and 2 list the total number of TAKS assessment items at each grade level and categorize each grade-level objective.

Table 1

Texas Assessment of Knowledge and Skills Objectives at Grades 3-8

Objective	Number of Items per Grade Level					
	3rd	4th	5th	6th	7th	8th
1—Numbers, operations, and quantitative reasoning	10	11	11	10	10	10
2—Patterns, relationships, and algebraic reasoning	6	7	7	9	10	10
3—Geometry and spatial reasoning	6	6	7	7	7	7
4—Measurement	6	6	7	5	5	5
5—Probability and statistics	4	4	4	6	7	8
6—Mathematical processes and tools	8	8	8	9	9	10
Total Number of Items	40	42	44	46	48	50

Table 2

Texas Assessment of Knowledge and Skills Objectives at Grades 9-11

Objective	Number of Items per Grade Level		
	9th	10th	11th
1—Functional relationships	5	5	5
2—Properties and attributes of functions	5	5	5
3—Linear functions	5	5	5
4—Linear equations and inequalities	5	5	5
5—Quadratic and other nonlinear functions	4	5	5
6—Geometric relationships and spatial reasoning	4	5	7
7—2D and 3D representations	4	5	7
8—Measurement	6	7	7
9—Percents, proportions, probability, and statistics	5	5	5
10—Mathematical processes and tools	9	9	9
Total Number of Items	52	56	60

Reading Grade Level

Prior to conducting this study, I needed to determine if TEA relied on any measure of reading grade level when creating assessment items for the TAKS. In each of the TEA (2008b) TAKS Information Booklets, a passage of text was used to illustrate how experts reviewed and determined the appropriateness of each assessment item prior to field testing. TEA did not specifically state whether experts analyzed the reading level of each mathematics assessment item; however, the use of experts in education, specifically mathematics education, does provide some level of validity that items were written at appropriate reading grade levels. Because TEA did not quantify the reading level of TAKS assessment items, a quantitative measure of the readability of these assessment items needed to be determined for the purposes of this study.

In 1935, Gray and Leary found that there were nearly 228 variables that affected readability. Semantic, syntactic, and stylistic elements accounted for the majority of these variables. Over the last century, numerous methods have been created to determine the readability of a passage. In 1948, Rudolph Flesch created one of the earliest formulas, the Flesch Reading Ease Formula. This formula incorporated the variables of sentence length and syllable count into a calculation of reading difficulty. Working for the United States Navy, Kincaid, Fishurn, Rogers, and Chissom (1975) adapted the Flesch Reading Ease Formula, known as the Flesch-Kincaid Grade Level Formula, to provide an outcome that better predicts a reader's grade-level. Edward Fry (1977) further advanced the science of readability by creating another popular readability formula known as the Fry Graph Readability Formula. He utilized the averages of sentence count and syllable count per 100 words to determine an ordered pair (average sentence count, average syllable count) located within sectors of a coordinate plane corresponding to the reading age of the text. The Flesch Reading Ease Formula, Flesch Kincaid Grade Level Formula, and Fry Graph Readability Formula all utilized some type of average counts based on sentences, words, and syllables that provided strong support for the syntactic structure of a passage of text and an effective means for determining semantic complexity. Determining readability based on these methods did not improve until the use of electronic databases surfaced in recent decades.

Electronic databases have been useful in improving the validity and reliability of calculated semantic complexity in readability formulas. The New Dale-Chall Readability Formula incorporates the variable of sentence length and a calculated percentage of words not found in a database of 3,000 common 4th-grade vocabulary terms (Chall & Dale, 1995). Chall and Dale believed words not found on this list of 3,000 common words were more difficult and thus used a calculated percentage of words not found in this list as a measure to produce higher grade-equivalence scores. Touchstone Applied Science Associates (1999) built upon the work of Chall and Dale (1995) by using a database of common vocabulary terms in their readability formula and by adding a new calculation of the average number of letters per word.

One of the latest and most sophisticated measures of readability is the Lexile Framework © for Reading (referred to as "Lexile") (2008). Lexile measures of readability are also based on semantic complexity and sentence length, but Lexile determines semantic complexity through the calculation of word frequency incorporating a database of nearly 600-million words, whereas other databases have only 5- to 25-million words (Wilson, Archer, & Rayson, 2006). Lexile has incorporated the advances of the past and paired them with an enormous database of words.

> Lexile text measures are based on two well-established predictors of how difficult a text is to comprehend: semantic difficulty (word frequency) and syntactic complexity (sentence length). In order to determine the Lexile measure of a book or article, the text is split into 125-word slices. Each slice is compared to the nearly 600-million word Lexile corpus—taken from a variety of sources and genres—and the words in each sentence are counted. These calculations are put into the Lexile equation. Then, each slice's resulting Lexile measure is applied to the Rasch psychometric model to determine the Lexile measure for the entire text. (Lexile Frequently Asked Questions, 2008)

The Lexile measure was chosen for this study because it is a grade equivalence measure that accounts for both syntactic and semantic elements, relies on an extensive database, and is readily accessible.

The first step in calculating the Lexile (2008) Measure for the TAKS assessment items required each mathematics assessment item to be converted into a text format, thus eliminating any graphs, tables, or other figures not representing standard sentence structure. Next, each assessment item was uploaded into the online Lexile Analyzer to obtain a Lexile Measure ranging from 10L to above 1700L. Each Lexile Measure was then located on the Lexile Map to determine its approximate reading grade level (RGL) ranging from 1 to 17, where scores of 13 through 17 represent post-secondary grade equivalencies. The Lexile Map provides an interval of Lexile Measures that correspond to a particular grade level at which the reader should have at least a 75% comprehension. Many Lexile Measures span more than one grade level. For instance, 3rd-grade Lexile Measures range from 520L-750L and 4th-grade Lexile Measures range from 620L-910L. In this case of overlapping Lexile Measures, 0.5 was added to the lowest grade level approximation. Therefore, in this study, a Lexile Measure of 620L would be assigned a grade equivalence of 3.5.

Cognitive Demand

According to the Wisconsin Center for Educational Research (WCER, 2008), there are five cognitive demand categories for mathematics: (1) memorize; (2) perform procedures; (3) demonstrate understanding; (4) conjecture, prove, solve; and (5) apply/make connections. The researcher used his training provided by the WCER to train nine groups of K-12 mathematics teachers on how to rate assessment items based on these cognitive demand categories. Grade-level groups of three to five teachers rated TAKS mathematics assessment items. After teachers rated the items individually, they discussed their ratings with their group. Because some assessment items could address more than one cognitive demand category, the teachers could categorize an item in up to three cognitive demand categories. An overall average cognitive demand score based on the teachers' categorizations was computed for each assessment item.

In general, the cognitive demand of each TAKS assessment had an even distribution of the cognitive demand levels from 1 to 5. However, as illustrated in Figure 1, the cognitive demand ratings for the 5th- and 11th-grade assessments did not match this trend. The 5th-grade distribution of cognitive demand levels had a much smaller range of scores and included more outliers than any of the other grade levels. The 11th-grade interquartile range of cognitive demand levels was elevated as compared to the other grade levels. In both cases, the unusual distribution may be the result of a single outspoken teacher in each group. Future research may use other means to categorize the cognitive demand level of assessment items to minimize this issue.

Data Analysis

An analysis of covariance (ANCOVA) determined differences in student performance based on the between-subjects factor of RGL. On a mathematics assessment, the item's content strand and level of cognitive demand are known variables that influence student performance. Field (2009) noted that when variables not part of the main experiment have an influence on the dependent variable, then an ANCOVA should be used to control for the effect of these covariates. Field also explained that two assumptions in an ANCOVA should be tested: "(1) independence of

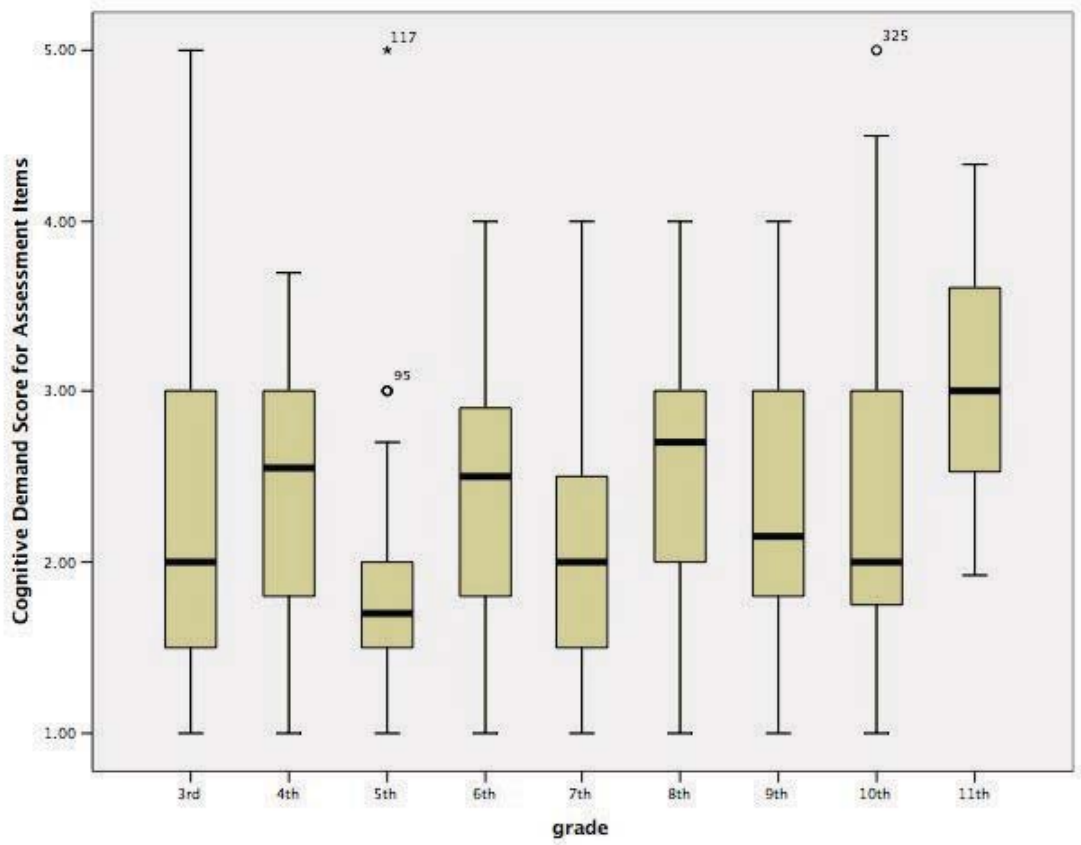

Figure 1. Box plots of cognitive demand scores for each TAKS grade level

the covariate and treatment effect, and (2) homogeneity of regression slopes" (p. 397). For these reasons, the data analysis in this study included tests of independence and homogeneity.

The mathematics TAKS assessment items were categorized based on whether the item's RGL is At-or-Below grade level (henceforth written as RGL At-or-Below) or above the student's grade level (henceforth written as RGL Above), establishing the independent variables in this study. For example, a 3rd-grade TAKS mathematics assessment item having a 2.5 RGL would be categorized as RGL At-or-Below, and a 3rd grade TAKS assessment item having a 5.0 RGL would be categorized as RGL Above. The dependent variable in this study was the percentage of students who correctly answered each assessment item.

An ANCOVA was conducted at three levels. The first level of analysis was a 2 x 9 factorial ANCOVA with RGL as the between-subjects factor and student grade level as the within-subjects factor. The covariates were cognitive demand and TAKS objective of each assessment item. The second level of analysis was a 2 x 3 factorial ANCOVA with RGL as the between-subjects factor and the assessment item's grade band (i.e., elementary school grades 3-5, middle school grades 6-8, or high school grades 9-11) as the within-subjects factor. The covariates of student grade level, cognitive demand, and TAKS objective data were used in this analysis. The final level of analysis grouped all assessment items from each TAKS grade level together. At this level, the covariates were grade level of each assessment item and the item's cognitive demand. SPSS 13 © was used for all statistical analysis in this study.

Tests of independence and homogeneity were also conducted for each ANCOVA, at each level of analysis. The test of independence consisted of a univariate analysis to determine if the RGL groups differed on each covariate. If no significant difference occurred between the RGL groups based on the

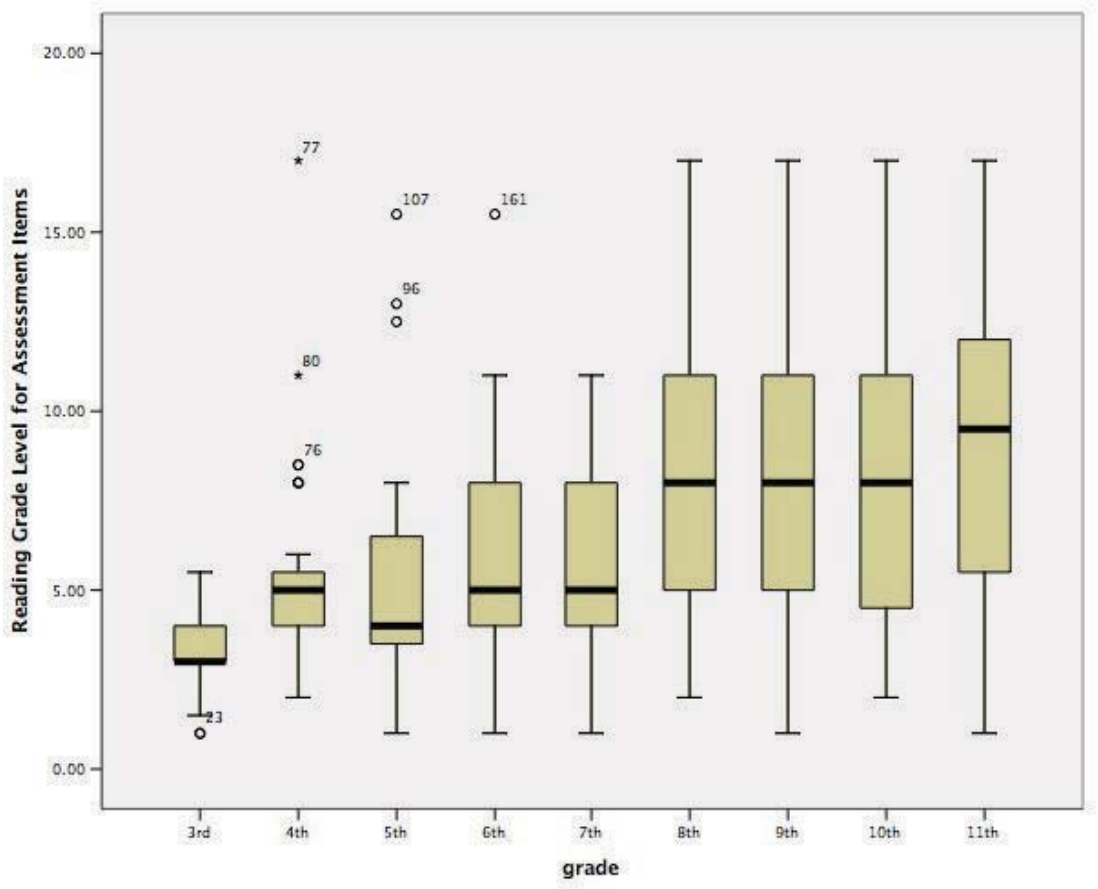

Figure 2. Box plot of the Reading Grade Level based on TAKS grade level

covariate, then the covariate was used in the ANCOVA. The test of homogeneity occurred after each ANCOVA with the researcher running the test a second time to check for any significant interactions between the independent variable of RGL and each covariate. If significant interaction was observed, the researcher further investigated this interaction using a univariate analysis between the RGL groups and covariate to further explain the significant interaction and determine effect (Tabachnick & Fidell, 1996).

Results

The average RGL for mathematics assessment items at each TAKS grade level ranged from 3.15 to 9.08 (see Table 3). All three elementary grades (3–5) had an average RGL greater than the student grade level being assessed. Excluding the 8th grade, all other grades had an average RGL below the student grade level being assessed.

As illustrated in Figure 2 and Table 3, the variance in RGL increased with the increase in the grade level. This increase in variance for higher grade levels was to be expected given that higher grade level assessments have a larger range of possible RGLs. It was also expected and confirmed that, at each grade level, there were assessment items written at a very low RGL. These assessment items were the standard mathematics questions requiring students to simply evaluate expressions or solve problems having no words other than action verbs and phrases like *solve*, *add*, *find the product*, or *evaluate*. Although there was a gradual increase in the range of RGL at each grade level, five out of the seven grade levels had assessment items with RGL measures at the graduate level (>16). Only the third grade assessment items were found to have a reasonable range of RGL measures, from 1 to 5.5.

The outliers identified in the cognitive demand and reading grade level variables presented some concern. Initial analysis at all three levels included assessment item data that were outliers in either RGL or cognitive

demand. Analysis was performed again, at all three levels, with these assessment items removed. At all three levels of analysis, there were no differences in results between the data with and without outliers. Therefore, in this study, the initial analysis results, using data that included outliers, is presented.

Table 3

Reading Grade Level Descriptive Statistics

TAKS Grade Level	Min. RGL	Max. RGL	Mean	Std. Dev.
3rd	1	5.5	3.15	0.99
4th	2	17	5.21	2.64
5th	1	155	5.23	2.89
6th	1	15.5	5.88	3.03
7th	1	11	5.81	2.58
8th	2	17	8.04	3.85
9th	1	17	8.15	4.03
10th	2	17	8.15	4.31
11th	1	17	9.08	4.18

A mean comparison of student performance based on the between-subjects factor of RGL was conducted at each grade level, each grade band, and over the entire set of mathematics assessment items. Items with an RGL At-or-Below had a higher percentage of successful students than items with an RGL Above in every grade level, grade band, and over the entire set of items except for 9th and 11th grade assessment items (see Table 5). These mean comparisons, however, do not account for the variance explained by other confounding variables. Therefore an analysis of covariance helped to determine differences in student performance at each grade level, grade band, and overall assessment items. The adjusted marginal means are provided in Table 5 to illustrate differences after controlling for various covariates. Inspection of differences between RGL group means, using either adjusted or unadjusted means, showed the 7th-grade RGL At-or-Below student performance average to be over 10 percentage points higher than the average for the 7th grade RGL Above. This adjusted mean difference between the two groups of items at the 7th grade level was the largest at any grade level, grade band, or over the entire set of assessment items. This observable difference at the 7th grade is discussed in great detail below.

Student Grade Level Analysis

The initial test of independence between the RGL groups and covariates of TAKS objective and cognitive demand resulted in no significant differences. This allowed for the inclusion of both covariates in the ANCOVA. Table 4 provides results from the ANCOVA performed at each grade level. It indicated a significant difference between RGL At-or-Below items and RGL Above items at the 7th-grade level ($F(1, 44) = 8.336$, $p = 0.006$, $\eta 2 = 0.159$, observed power = 0.81). Tests of homogeneity between the RGL groups and the covariates of cognitive demand and TAKS objective had no significant interactions except at the 10th-grade level. In this case, homogeneity was violated between the RGL groups and cognitive demand ($F(2, 51) = 6.453$, $p = 0.003$). Further investigation into this significant interaction revealed no significant difference in cognitive demand between the 10th-grade RGL groups ($F(1, 54) = 1.300$, $p = 0.259$). Therefore, the inclusion of the covariate in the analysis is justified (Tabachnick & Fidell, 1996).

Table 4

Tests of Between-Subjects Effects of RGL at Each Grade Level

Grade	F	Sig.	η^2	Observed Power
3rd	0.752	0.391	0.020	0.135
4th	1.778	0.190	0.045	0.255
5th	3.708	0.061	0.085	0.468
6th	2.571	0.116	0.058	0.347
7th	8.336	0.006	0.159	0.806
8th	0.060	0.808	0.001	0.057
9th	0.149	0.701	0.003	0.067
10th	2.788	0.101	0.051	0.374
11th	0.011	0.917	0.000	0.051

Note: $\alpha = 0.05$

Student Grade Band Analysis

The second phase of analysis of covariance yielded results related to differences in student performance at each of the three grade bands. The initial test for independence found that the covariates of cognitive demand, TAKS objective, and student grade level did not have any significant differences between the RGL

Table 5

Adjusted and Unadjusted Student Performance Means and Variability

Grade Level	RGL	N	Unadjusted		Adjusted	
			M	SD	M	SE
3rd	At-or-Below	28	81.04	11.12	81.10	2.42
	Above	12	77.42	17.44	77.26	3.70
	Total	40	79.95	13.20	79.18	2.20
4th	At-or-Below	20	82.45	7.79	82.51	1.60
	Above	22	79.59	7.78	79.54	1.53
	Total	42	80.95	7.82	81.02	1.10
5th	At-or-Below	30	82.27	6.57	82.31	1.31
	Above	14	77.86	7.93	77.77	1.93
	Total	44	80.86	7.24	80.04	1.15
6th	At-or-Below	33	78.03	11.03	77.98	2.06
	Above	13	71.62	13.06	71.75	3.29
	Total	46	76.22	11.85	74.86	1.94
7th	At-or-Below	34	**70.29**	9.95	**70.87**	1.90
	Above	14	**61.71**	13.74	**60.31**	3.03
	Total	48	67.79	11.72	65.59	1.75
8th	At-or-Below	31	67.97	12.96	67.74	2.25
	Above	19	66.47	11.10	66.84	2.89
	Total	50	67.40	12.19	67.29	1.82
9th	At-or-Below	32	61.03	12.97	61.34	2.41
	Above	20	63.35	14.49	62.86	3.06
	Total	52	61.92	13.48	62.10	1.92
10th	At-or-Below	40	66.25	13.25	65.49	1.87
	Above	16	57.56	11.06	59.47	3.01
	Total	56	63.77	13.18	62.48	1.75
11th	At-or-Below	43	40.86	14.46	40.86	2.20
	Above	17	41.29	15.06	41.29	3.51
	Total	60	40.98	14.50	41.08	2.07
Elementary	At-or-Below	78	81.87	8.66	81.96	1.04
	Above	48	78.54	10.80	78.40	1.32
	Total	126	80.60	9.63	80.18	0.84
Middle School	At-or-Below	98	72.16	12.01	71.95	1.20
	Above	46	66.48	12.81	66.94	1.76
	Total	144	70.35	12.51	69.44	1.06
High School	At-or-Below	115	55.30	17.68	55.65	1.37
	Above	53	54.53	16.49	53.78	2.03
	Total	168	55.06	17.27	54.71	1.22
Overall	At-or-Below	291	68.10	17.68	68.53	0.77
	Above	147	66.11	16.86	65.28	1.09
	Total	438	67.43	17.42	66.90	0.67

groups at the elementary and high school levels. However, a significant difference was found between the RGL groups based on the cognitive demand covariate at the middle school level (F (1, 142) = 5.540, p = 0.020). This led to the rejection of the assumption that the covariate and treatment effect were independent, and therefore the cognitive demand variable was omitted from the ANCOVA for the

Table 6

Tests of Between-Subjects Effects of RGL at each Grade Band

Grade Band	Source	Type III Sum of Squares	df	Mean Square	F	Sig.	η^2	Obs. Pwr. [a]
Elementary School	Cognitive Demand	22.126	1	22.13	0.264	0.608	0.00	0.08
	TAKS Objective	1097.001	1	1097.00	13.107	0.000	0.10	0.95
	Student Grade Level	23.005	1	23.01	0.275	0.601	0.00	0.08
	RGL	374.203	1	374.20	4.471	0.037	0.04	0.56
	Error	10127.313	121	83.70				
Middle School	TAKS Objective	24.382	1	24.382	0.173	0.678	0.00	0.07
	Student Grade Level	1612.186	1	1612.186	11.429	0.001	0.08	0.92
	RGL	773.341	1	773.341	5.482	0.021	0.04	0.64
	Error	19747.948	140	141.057				
High School	Cognitive Demand	1108.06	1	1108.06	5.139	0.025	0.03	0.62
	TAKS Objective	455.103	1	455.10	2.111	0.148	0.01	0.30
	Student Grade Level	14089.68	1	14089.68	65.34	0.000	0.29	1.00
	RGL	125.434	1	125.43	0.582	0.447	0.00	0.12
	Error	35148.711	163	215.64				

[a] Computed using $\alpha = 0.05$

Table 7

Tests of Between-Subjects Effects of RGL Over All Assessment Items

Source	Type III Sum of Squares	df	Mean Square	F	Sig.	η^2	Obs. Pwr. [a]
Cognitive Demand	0.025	1	0.025	0.000	0.990	0.00	0.05
Student Grade Level	54532.535	1	54532.535	315.162	0.000	0.42	1.00
RGL	1025.366	1	1025.366	5.926	0.015	0.01	0.68
Error	75095.139	434	173.030				

[a] Computed using $\alpha = 0.05$

middle school grade band. The ANCOVA yielded significant differences between the RGL At-or-Below assessment items and RGL Above assessment items in both the elementary (F (1, 121) = 4.471, p = 0.037, η^2 = 0.036, observed power = 0.56) and middle school (F (1, 140) = 5.482, p = 0.021, η^2 = 0.038, observed power = 0.64) grade bands (see Table 6). The adjusted mean difference at the elementary grade band resulted in the RGL At-or-Below student performance average being 3.52 percentage points higher than the RGL Above average. The middle school marginal mean difference had the RGL At-or-Below average 5.01 percentage points higher. At both of these grade bands, the students performed better on the assessment items written at-or-below their grade level. As illustrated by the adjusted means in Table 4, there was virtually no difference between RGL At-or-Below averages and RGL Above averages found at the high school grade band.

Tests of homogeneity were violated at all three grade bands. At the elementary grade band, a significant interaction existed between the RGL groups and TAKS objective covariate (F (2, 121) = 6.583, p = 0.002). This did not affect the ANCOVA results; no significant differences in TAKS objectives were found between the RGL groups at the elementary level (F (1, 124) = 0.078, p = 0.781). A significant interaction between the RGL groups and student grade level covariate at the middle school indicated homogeneity was violated (F (2, 139) = 5.832, p = 0.004). Similarly, because no significant differences in student grade level were found between the RGL groups at the middle school level (F (1, 142) = 1.063, p = 0.304), these interactions were not found to affect the ANCOVA results. Homogeneity was violated at the

high school level as well, with a significant interaction between the RGL groups and student grade level covariate (F (2, 163) = 29.315, p = 0.000). Like the elementary and middle grades, this interaction did not affect the ANCOVA, having no significant differences in student grade level between the RGL groups at the high school level (F (1, 166) = 1.260, p = 0.263).

Overall Assessment Item Analysis

Controlling for cognitive demand and student grade level, the final level of analysis of covariance determined a significant difference in student performance based on the RGL for all TAKS mathematics assessment items (F (1, 434) = 5.926, p = 0.015, η^2 = 0.013, observed power = 0.68) (see Table 7). The marginal mean difference between the RGL At-or-Below student performance average and RGL Above student performance average was 3.25 percentage points, indicating that students performed better on the RGL At-or-Below questions. Both the cognitive demand (F (1, 436) = 1.174, p = 0.279) and student grade level (F (1, 392) = 0.313, p = 0.576) covariates were independent from the RGL group variable. However, a significant interaction existed between the RGL group variable and student grade level covariate (F (2, 389) = 127.504, p = 0.000). This violation of homogeneity did not, however, have an effect on the ANCOVA, with no significant differences in student grade level found between the RGL groups.

Discussion and Conclusions

The purpose of this study was to determine if the reading difficulty of mathematics assessment items affected student performance on TAKS items. The claim that student performance would be affected by RGL of mathematics assessment items is supported by the results of this study. Analysis revealed that students in the state of Texas performed significantly lower on mathematics assessment items having RGL measures above their grade level than on items having RGL measures At-or-Below their grade level. This was especially true for elementary and middle school students. More specifically, the seventh grade students in the state of Texas were negatively affected by the RGL of mathematics assessment items.

Despite controlling for several extraneous variables at each level of analysis, the RGL explained very little of the variance as evidenced through the low effect size coefficients. Additionally, the observed power at nearly all levels of analysis yielded results illustrating the limited power of the RGL, except at the 7th grade level, where the observed power coefficient was above 0.8. This evidence of low effect size and power coefficients either suggests more extraneous variables could be identified or that the limitations of this study had an effect on the results.

In this study, a grade-equivalence measure was obtained for each mathematics assessment item's RGL to determine the item's categorization as being at-or-below versus above the student grade level. This measure constituted the greatest limitation of this study. Like Walker et al. (2008) noted, determining the reading grade level of a small passage of text, like that of a mathematics assessment item, lacks reliability. Lexile (2008) based their calculations on subsets of 125 words within the text sample, and no assessment item in this study had a word count close to 125 words. Therefore the calculations made using the Lexile Analyzer may not have provided the best reading grade-equivalence measure.

Another limitation to this study was the number of assessment items analyzed at each grade level. After items were categorized based on their RGL, nearly all subsets of RGL Above items had less than 20 items. A larger sampling of grade level items would have been ideal. Additionally, the utilization of publicly available data with respect to student performance measures limited the results of this study. Obtaining student level data that can be disaggregated would also be useful because determining how students from different ethnic, socioeconomic, regional, and gender groups could help in identifying how different subgroups of students are affected by the RGL of mathematics assessment items. However, the usage of state level data allowed this study to illustrate effects evident within the overall population of students in Texas.

Overall, this study provides support to the available research indicating the negative effect of a greater reading difficulty on student performance on mathematics assessment items (Bolt & Thurlow, 2007; Powell et al., 2009; Walker et al., 2008). Ideally, assessment items would minimize reading difficulty without jeopardizing mathematical complexity. Therefore, investigating ways of writing mathematics assessment items that require students to read and synthesize text without going beyond the students' reading grade level is imperative. Further empirical research is needed in this area.

Because this study relied on Texas state assessment items in mathematics that are used for determining AYP, implications regarding accountability practices should be considered. If schools and school districts are held accountable for student performance on standardized state mathematics assessments (NCLB, 2002), students (and hence

schools and districts) may be unduly penalized twice, once for low reading performance and once for low mathematics performance resulting from reading difficulties. To provide an accurate assessment of student mathematics performance, student results in mathematics may need to reflect individual student reading difficulties. Adjusted mathematics assessment scores could be created based on individual student reading levels, and these adjusted mathematics achievement results could be used in state accountability measures.

Students are called to acquire mathematical skills that are not only grounded in computation but also in complex problem solving (NCTM, 2000). These complex problem solving skills are assessed with items that unavoidably require a great deal of reading (Bolt & Thurlow, 2007; Walker et al., 2008); mathematics test items will continue to include reading passages in order to accurately present the mathematical situation. Teachers, administrators, test and textbook writers, students, and parents should understand that this dualistic nature of mathematics assessment items is inevitable. However, researchers should continue to investigate ways to minimize the reading difficulty of assessment items without limiting the mathematics content. There may soon be accountability measures that reflect the reading levels of students. But, until that time, teachers must continue to teach both the mathematics content and reading strategies in order for students to perform their best.

References

Boero, P., Douek, N., & Ferrari, P. L. (2008). Developing mastery of natural language: Approaches to some theoretical aspects of mathematics. In L. D. English (Ed.), *Handbook of International Research in Mathematics Education* (2nd ed., pp. 262–295). New York: Routledge.

Bolt, S. E., & Thurlow, M. L. (2007). Item-level effects of the read-aloud accommodation for students with reading disabilities. *Assessment for Effective Intervention*, *33*(1). 15–28.

Breen, M. J., Lehman, J., & Carlson, M. (1984). Achievement correlates of the Woodcock-Johnson reading and mathematics subtests, Keymath, and Woodcock Reading in and elementary aged learning disabled population. *Journal of Learning Disabilities*, *17*, 258–261.

Carter, T. A., & Dean, E. O. (2006). Mathematics intervention for grades 5-11: Teaching mathematics, reading, or both? *Reading Psychology*, *27*, 127–146.

Chall, J. S., & Dale, E. (1995). *Readability revisited: The new Dale-Chall Readability Formula*. Cambridge, MA: Brookline.

Field, A. (2009). *Discovering statistics using SPSS* (3rd Ed). London: Sage.

Flesch. R. (1948). A new readability yardstick. *Journal of Applied Psychology*, *32*, 221–233.

Fry, E. (1977). Fry's readability graph: Clarifications, validity, and extension to level 17. *Journal of Reading*, *21*, 242–252.

Fuchs, L. S., Fuchs, D., Eaton, S. B., Hamlett, C. L, & Karns, K. M. (2000). Supplementing teacher judgments of mathematics test accommodations with objective data sources. *School Psychology Review*, *29*, 65–85.

Gray, W. S., & Leary, B. E. (1935). *What makes a book readable?* Chicago: University of Chicago Press.

Helwig, R., Rozek-Tedesco, M. A., Tindal, G., Heath, B., & Almond, P. J. (1999). Reading as an access to mathematics problem solving on multiple-choice tests for sixth-grade students. *Journal of Educational Research*, *93*, 113–125.

Jerman, M. E., & Mirman, S. (1974). Linguistic and computational variables in problem solving in elementary mathematics. *Educational Studies in Mathematics*, *5*, 317–362.

Jiban, C. L., & Deno, S. L. (2007). Using math and reading curriculum-based measurements to predict state mathematics test performance: Are simple one-minute measures technically adequate? *Assessment for Effective Intervention*, *32*(2), 78–89.

Jordan, N. C., Hanich, L. B., & Kaplan, D. (2003). A longitudinal study of mathematical competencies in children with specific mathematics difficulties versus children with comorbid mathematics and reading difficulties. *Child Development*, *74* , 834–850.

Ketterlin-Geller, L. R., Yovanoff, P., & Tindal, G. (2007). Developing a new paradigm for conducting research on accommodations in mathematics testing, *Exceptional Children*, *73*, 331–347.

Kincaid, J. P., Fishburn, R. P., Rogers, R. L., & Chissom, B. S. (1975). *Derivation of new readability formulas (automated readability index, fog count and Flesch reading ease formula) for Navy enlisted personnel (Branch Report 8-75)*. Naval Technical Training Command, Millington, TN.

Lager, C. A. (2006). Types of mathematics-language reading interactions that unnecessarily hinder algebra learning and assessment. *Reading Psychology*, *27*, 165–204.

Lexile Framework for Reading. (2008). *Lexile Analyzer*. Retrieved May 1, 2008, from http://www.lexile.com

Lexile Frequently Asked Questions. (2008). FAQ. Retrieved August, 20, 2009, from http://www.lexile.com/

Mississippi Department of Education. (2008). *Mississippi public school accountability standards 2008*. Retrieved on September 8, 2009, from http://mde.k12.ms.us

National Council of Teachers of Mathematics. (2000). *Principles and Standards for School Mathematics*. Reston, VA: Author.

No Child Left Behind (NCLB) Act of 2001, Pub. L. No. 107-110, § 115, Stat. 1425 (2002).

Ohio Department of Education. (2007). *Addendum to the proposal to the United States Department of Education for employing a growth model for No Child Left Behind accountability purposes*. Retrieved March 30, 2010, from http://www.ode.state.oh.us/GD/DocumentManagement/

Pitts, R. J. (1952). Relationships between functional competence in mathematics and reading grade levels, mental ability, and age. *Journal of Educational Psychology*, *43*, 486–492.

Powell, S. R., Fuchs, L. S., Fuchs, D., Cirino, P. T., & Fletcher, J. M. (2009). Do word-problem features differentially affect problem difficulty as a function of students' mathematics difficulty with and without reading difficulty? *Journal of Learning Disabilities*, *42*, 99–110.

Reikerâs, E. K. L. (2007). Performance in solving arithmetic problems: a comparison of children with different levels of achievement in mathematics and reading. *European Journal of Special Needs Education*, *21*, 233–250.

Rubenstein, R. N. (2000). Word origins: Building communication connections. In A. R. Teppo (Ed.), *Reflecting on NCTM's Principles and Standards in elementary and middle school mathematics* (pp. 243–247). Reston, VA: National Council of Teachers of Mathematics.

Tabachnick, B. G., & Fidell, L. S. (1996). *Using multivariate statistics*. New York: Harper Collins.

Texas Education Agency (TEA). (2008a). *Assessment and testing*. Retrieved May 1, 2008, from http://www.tea.state.tx.us/assessment.html

TEA. (2008b). *Revised TAKS information booklets*. Retrieved March 16, 2010, from http://www.tea.state.tx.us/

TEA. (2009). *Texas growth proposal to the United States Department of Education*. Retrieved September 8, 2009, from http://www.tea.state.tx.us/index3.aspx?id=3688&menu_id3=793.

Thompson, E. N. (1967). *Readability and accessory remarks: Factors in problem solving in arithmetic*. (Doctoral dissertation, Stanford University, 1967). Dissertation Abstracts International, 28, 2464A.

Touchstone Applied Science Associates. (1999). *Degrees of reading power*. Brewster, NY: Author.

United States Department of Education (USDOE). (2008). *Growth models: Ensuring grade-level proficiency for all students by 2014*. Retrieved March 29, 2010, from http://www2.ed.gov/admins/lead/account/growthmodel/proficiency.html

USDOE. (2009). *Secretary Spellings approves additional growth model pilots for 2008-2009 school year*. Retrieved March 29, 2010, from http://www2.ed.gov/news/pressreleases/2009/01/01082009a.html

Walker, C. M., Zhang, B., & Surber, J. (2008). Using a multidimensional differential item functioning framework to determine if reading ability affects student performance in mathematics. *Applied Measurement in Education*, *21*, 162–181.

Wilson, A., Archer, D., & Rayson, P. (Eds.) (2006). *Corpus linguistics around the world*. New York: Editions.

Wisconsin Center for Educational Research. (2008). *Coding procedures for curriculum content analyses*. Retrieved May 1, 2008, from http://seconline.wceruw.org/Reference/CodingProcedures2008.pdf