All Reading Tests Are Not Created Equal: A Comparison of the State of Texas Assessment of Academic Readiness (STAAR) and the Gray Oral Reading Test-4 (GORT-4)

Kary A. Johnson
Celia M. Wilson
Texas Wesleyan University

Dara Williams-Rossi
Southern Methodist University

Abstract

This exploratory study investigated how reading comprehension was conceptualized on the new high-stakes test, the 2011-2012 State of Texas Assessment of Academic Readiness (STAAR). Specifically, comprehension, rate, and accuracy scores on the Gray Oral Reading Test 4(GORT-4) from a group of struggling, low-SES, Hispanic middle school students (n = 59) were set as predictor variables to examine possible relationships with the STAAR. Initial bivariate correlations showed a weak relationship between GORT-4 predictor variables (comprehension, rate, accuracy) and STAAR ELA scores. Moreover, the overall regression model was not a good fit, with the linear combination of the GORT-4 components of comprehension, rate, and accuracy accounting for only 3.5 % of the variance in STAAR scores. The weak relationship between STAAR test results and the GORT-4 is examined in light of the current research on high-stakes testing, particularly for the at-risk population studied.

Since the inception of the No Child Left Behind Act of 2001 (NCLB, 2002), formally known as the reauthorization of the Elementary and Secondary Education Act (ESEA, 2002), public school K-12 education has changed. Implementation of high-stakes testing has altered the national teaching landscape in terms of how concepts are taught and how instructional time is allocated. Teacher practices have become increasingly standardized by district mandates; a seemingly rational response for a system desperately striving to meet the demands of federally mandated legislation requiring testing implementation (Amrien & Berliner, 2002; Au, 2011). Yet, research shows achievement has not truly increased. For example, according to a large-scale longitudinal study conducted on participants from 27 states involved in high-stakes testing, while student scores on high-stakes measures have steadily increased in reading and math, corresponding student scores on the National Association of Educational Progress (NAEP) have not increased (Amrien & Berliner, 2002). In fact, many of the states studied report flat rates of achievement on the NAEP examinations since the advent of high stakes tests (Amrien & Berliner, 2002; Shepard, 2003). Thus, while concerns ushered in by state testing requirements are wide and varied, pressing initial concerns about what high-

stakes instruments actually measure, specifically in terms of reading and language arts, must be addressed (Bracey, 2005). Furthermore, ongoing concerns about singular use of high-stakes testing for diagnostic and intervention purposes must also be addressed (Hale & Fiorello, 2001).

Purpose of the Study

The two research questions were examined for the current exploratory comparative study. First, the research team sought to determine how the 2011-2012 State of Texas Assessment of Academic Readiness (STAAR) measured reading achievement as compared to how a widely utilized nationally normed test, the Gray Oral Reading Test (GORT-4), measures reading. And second, the research team sought to determine if the specific constructs tested by the GORT-4 (reading comprehension, reading rate, and reading accuracy) were in some way predictive of achievement on the STAAR, thereby indicating potential diagnostic utility of the STAAR in terms of intervention planning for struggling students in 6th through 8th grades.

What Do Tests of Reading Measure?

When investigating various measures of the same domain of achievement, many practitioners and researchers logically assume tests which measure similarly named constructs actually measure the same thing (Amrien & Berliner, 2002). Thus students who struggle with performance on various facets of high-stakes exams, such as the STAAR, simply should receive intervention in the areas of weakness as indicated by the high stakes assessment. Therefore, if all tests of reading comprehension can be held equal, then a student, who performs poorly on a high stakes measure of reading comprehension such as the STAAR, should simply be provided ensuing interventions in reading comprehension. Unfortunately, though, all tests of reading comprehension are not created equal.

High-stakes assessments are created to measure student mastery of curricular expectations included in the state curriculum (Hintz & Silberglitt, 2005). Paradoxically, student performance on high-stakes tests of reading such as the ELA portion of the STAAR often do not correlate with or transfer to performance in the classroom and on other measures of reading performance (Shepard, 2003). Lack of transfer likely occurs because different tests of reading measure different constructs, especially in terms of the complex domain of reading comprehension. For instance, in a recent study conducted on low income, urban, middle school students (n = 91), researchers found reading skill as measured by traditional reading assessments did not predict performance on high stakes measures of reading. Instead, executive function skills such as self-monitoring and metacognitive awareness accounted for 40% of the variance in high stakes reading test scores (Waber, Gerber, Turcios, Wagner, & Forbes, 2006). Other recent research suggests various measures of reading comprehension are differentially reliant on the factors of listening comprehension and verbal ability, as compared to decoding ability (Keenan, Betjemann, & Olson, 2008). Still other researchers assert reading speed also accounts for unique variance on high stakes measures of reading (Cutting & Scarborough, 2006).

Use of High Stake Measures as Diagnostic Instruments

While not all researchers agree upon which underlying deficits impede reading comprehension and achievement, many agree provision of multiple measures is superior to use of singular measures when determining skills to target for reading intervention (Hale & Fiorello, 2001). Moreover, as specific tests of reading comprehension have been empirically tied to various related and underlying factors (reading rate, IQ, language ability, listening comprehension, decoding accuracy, and sustained attention), use of multiple assessment tools provides those who plan and implement intervention a more thorough view of areas to target for instruction (Cutting & Scarborough, 2006). Meta-analytic research reaffirms the complex nature of the process of learning in general and learning to read and comprehend in particular, especially for those who struggle in reading acquisition (Adams, 1990; Hale & Fiorello, 2010; NIH 2000; Pennington, 2009; Shaywitz and Shaywitz, 2007). As reading comprehension is not a singular construct, concerns about use of one measure of reading for diagnostic and intervention purposes exist. More specifically, the lack of specific diagnostic information provided by high-stakes achievement tests is particularly concerning. To this end, in 2001, The National Research Council called for a system-wide improvement of the diagnostic data provided by state-mandated high stakes measures, encouraging test developers to provide more thorough feedback to teachers about the strategies children employ when problem solving on such examinations (Madaus & Russell, 2010). Although concerns about the diagnostic use of high-stakes testing exist, it can be assumed the STAAR, much like its high-stakes predecessor, the Texas Assessment of Knowledge and Skills (TAKS) will continue to be used by teachers and districts to make instructional decisions especially for those students who struggle in reading (Edwards & Pula, 2011, Guskey, 2003).

Impacts of High-Stakes Testing

High-stakes testing is a hotly debated and controversial topic in what many call the era of accountability (Assaf, 2006). Specifically tied to the federal government's No Child Left Behind act of 2001 (NCLB, 2002), testing of all students in reading and in math has become a phenomenon in the American K-12 public education system (Assaf, 2006; Au, 2011). Ensuing system-wide implementation of highly controlled, narrow, test-driven, curriculum has also become the norm of many school leaders (Zhao, 2012). While high stakes tests are supposed to produce a more rigorous system of education, many systematic studies indicate unintended negative results are produced, including increased drop-out rates, decreased graduation rates, decreasing student and teacher motivation, and a narrowed curriculum (Amrien & Berliner, 2003; Madaus & Russell, 2010).

Advocates of high-stakes testing insist a narrowed curriculum allows educators to get "back to the basics" of reading, writing, and arithmetic. And while it is true that more time is allocated to these critical areas, standardization has also led to cuts in non-tested subject areas (Au, 2011; Lobascher, 2011). For instance, Au reported that as of 2010, 71% of US districts had cut one subject to increase time in

math or reading due to the increased high stakes testing focus contained within NCLB (Au, 2011). Moreover, beyond narrowing the content of the curriculum to the basics of reading and math, curriculum has become controlled, not by those in the classroom, but instead by upper level bureaucrats, who often advocate for concepts to be taught in small, discrete, linear units (Assaf, 2006; Au, 2011). For reading curriculum, practicing educators contend the opposite should occur; units of reading instruction should spiral, with reintroduction of important concepts (i.e. main idea, summarization, authors purpose etc.) occurring regularly within various contexts and genres of literature (Atwell, 2007). According to Berliner (2011), curriculum narrowing is the most serious of sins associated with high-stakes testing as it naturally restricts learners from engaging in enjoyable and creative activities, thereby reducing higher level thinking.

Several studies have also linked the advent of high-stakes testing to further marginalization of children living in low socioeconomic status (SES). For example, Marder, Bansal, and Kadanoff (2009), analyzed data from 4.6 million students who took the TAKS in 2003 and in 2007, and found the single most significant predictor of student performance on high stake exams was income level (after reviewing all possible predictors such as past TAKS scores, random guessing, retention rate, and transience). Sadly, this influence of SES on high stakes achievement worsens throughout the middle school years, leading to retention and eventual drop out (Marder, Bansal, & Kadanoff, 2009). Further, in a 2010 study of 14,059 5th grade children who were given Florida's high-stakes assessment, the Florida Comprehensive Achievement Test (FCAT), researchers found only 39% of the low SES students passed, as compared to 65% of the high SES students (Baker & Johnston, 2010).

Method

Setting and Participants

The 59 participants for the present study were 6th, 7th, and 8th grade students enrolled in reading improvement classes at one urban middle school in Texas. There were 52 Hispanic students, 3 Caucasian (non-Hispanic) students, 3 Asian students, and one African American student. The sample included 36 females and 23 males. The majority of the students were coded as economically disadvantaged (specific economic codes for individual students within the tested sample were not available to the researchers). Finally, the mobility rate for the campus was approximately 17%, a rate proportionate to overall mobility levels for the state of Texas.

Instrumentation

For purposes of the present study, comparisons were made between the Gray Oral Reading Test 4th Edition (GORT-4) and the STAAR. The GORT-4 is a classically created, norm-referenced, assessment measuring reading rate, reading accuracy, and reading comprehension for students in 2nd through 12th grade (Wiederholt & Bryant, 2001). GORT-4 internal consistency (reliability) coefficients, reported by the test authors for all areas of the GORT rate, accuracy, comprehension all met or exceeded $\alpha = .90$ (Wiederholt & Bryant, 2001). For purposes of the present study, internal consistency metrics were also computed with a resultant Cronbach's alpha of $\alpha = .87$ or better for all tested areas (rate, accuracy, comprehension). As alphas

of .70 or above are generally considered sufficient; an alpha of .87 or above is well within the acceptable range, indicating the GORT-4 was internally consistent for the normative sample as well as the current sample (Henson, 2001).

The STAAR is a newly developed, criterion-referenced, high-stakes test aligned to the State of Texas Curriculum Standards, the Texas Essential Knowledge and Skills (TEKS). In 6th through 8th grade, the English Language Arts portion of the STAAR provides an individual student total raw score, as well as, raw scores in three subscales: reading comprehension of literary text (including fiction, literary non-fiction, poetry and drama subtypes), reading comprehension of information text (including expository and persuasive subtypes), and understanding and analysis across genres (comparing across all genres in literary text and information text above) (Texas Education Agency, 2011). The STAAR, like many other high-stakes evaluations, was created based on latent trait theory (also called item response theory). Scores given in tests created with this newer latent trait theoretical foundation are based on a different perspective than classical test creation methodology (e.g. norm referenced tests like the GORT-4). Most notably, instead of basing scores on norms within the population, the test is scored based on a continuum of the trait being examined (Mason, 2007).

Data Collection and Analysis

Data collection consisted of first administering the Gray Oral Reading Test-4 form A, individually to students in the sample (n = 59). GORT-4 examiners included trained members of the research team, as well as research assistants who had advanced degrees in reading (minimum Master's level), and specific training in test administration. The testing environment was controlled and quiet and all testing procedures outlined in the GORT-4 examiner's manual were implemented accordingly.

After assessing individual students, GORT-4 assessment protocols were scored by the research team. Age-based scores for the following three individual constructs were calculated: reading rate, reading accuracy, and reading comprehension. In examining the GORT-4 scores for the study sample, all mean scores were more than one standard deviation below the population outcomes, as presented by the authors of the GORT-4 (M = 100, σ = 15). This is an expected finding as the present sample only included identified struggling readers. Moreover, present study standard deviations were smaller (between 12.88 and 8.39) indicating a smaller range of scores for participants than for the typical distribution of individuals (Weiderholt & Bryant, 2001). Descriptive statistics for GORT-4 scores are shown in Table 1 located at the end of the article.

After scoring GORT-4 protocols, the research team converted raw scores for individual student STAAR performance to percentage correct scores. Scores for the total STAAR test and three subscales of understanding across genres, literary text, and information text were included for purposes of descriptive understanding prior to implementation of multiple regression analysis (See Table 2 located at the end of the article). When the present study was conducted, no passing standard was set by the state, yet note all mean scores indicate

values of less than 50% correct. In contrast to relatively narrow standard deviations for the GORT-4 scores, STAAR standard deviations were large (between 13.07 and 19.75) indicating more variability in the data.

Results

After computing descriptive statistics, a multiple regression model was created using SPSS to determine if performance on the GORT-4 was predictive of, or related to, performance on the STAAR (Field, 2009). Overall model summary findings, as well as, specific contributions of three predictor variables of comprehension, reading accuracy and reading rate from the GORT-4 were analyzed. As previous studies indicate reading rate and reading accuracy scores may be predictive of comprehension scores, the predictors of rate and accuracy were retained in addition to comprehension (Cutting & Scarborough, 2006; Keenan, Betjemann, & Olson, 2008; Wiederholt & Bryant, 2001).

When examining the correlations between the STAAR and GORT-4, no predictor was strongly or significantly associated with the outcome variable of the STAAR total score (r = .131 for rate; r = .211 for accuracy; r = .243 for comprehension). As such, student performance on the GORT-4 (reading rate, reading accuracy; reading comprehension) cannot be used to indicate areas for reading intervention simply by analyzing performance on the STAAR. Further, findings indicated the multiple regression model (see Table 3 located at the end of the article) was not significant (p =.1 79, $\alpha < .05$). Thus, GORT-4 predictor variables (rate, accuracy, comprehension) did not explain STARR total scores (adjusted as the R^2 value was .035), as there was only 3.5% of the variance in STAAR outcomes. As such, there is little relationship between the ELA STAAR total score and reading as measured by the GORT-4 for the studied participants.

Discussion

Based on the results of the current study, the GORT-4 and the STAAR do not measure reading comprehension in a similar manner and questions remain as to what the STAAR is measuring. As such, future research is warranted to determine how the STAAR measures the complex construct of reading comprehension, especially for those students most at risk for failure, including but not limited to, those students in high poverty, minority, and learning disabled groups (Baker & Johnston, 2010; Shepard, 2003). Further investigation into other potential confounding factors (SES, gender, ethnicity, etc.) influencing STAAR outcomes, beyond the GORT-4 predictors of reading comprehension, reading rate, and accuracy is also warranted.

Given the lack of relationship between the STAAR and GORT-4, the complex nature of the reading process, and the complex process of learning, STAAR ELA scores are likely influenced by various abilities and proficiencies (Waber et al., 2006). The present study shows while the STARR may somehow measure reading for middle school students, it does not measure comprehension, rate, or accuracy in the same manner as other commonly used diagnostic reading measures such as the GORT-4. This is a problem and educators should not rely solely on the STAAR test for decisions regarding instructional planning. In addition, practicing ELA middle school

educators are encouraged to use multiple formal and informal reading assessment tools (in addition to STAAR scores) to pinpoint areas of reading difficulty and plan reading intervention for individual students who struggle (Cutting & Scarborough, 2006; Hale & Fiorello, 2010).

At the policy level, the institutionalized practice of using STAAR reading scores for unilateral decision making in terms of student promotion and retention, student graduation, and school and district performance is also called into question. The lack of relationship between STAAR outcomes and GORT-4 outcomes in the current study suggests we may not know what facets of reading the ELA STAAR measures, especially for at-risk populations. As such overreliance on STAAR as a diagnostic indicator of reading performance for students seems not only premature but potentially harmful, as unintended, negative, consequences including, but not limited to, student distress, teacher burnout, and curriculum narrowing may occur (Berliner, 2011).

Table 1

GORT-4 Scores

Construct	*M*	*SD*
GORT-4 Reading Comprehension	84.49	8.39
GORT-4 Reading Rate	83.73	9.45
GORT-4 Reading Accuracy	76.78	10.49

Table 2

STAAR Descriptive Statistics

Construct	*M*	*SD*
STAAR Total Score	47.46%	13.07
STAAR Subscale: Between Genres	48.31%	19.75
STAAR Subscale: Literary Text	46.14%	14.54
STAAR Subscale: Informational Text	46.41%	14.47

Table 3

Multiple Regression Summary of GORT-4 Predictors on STAAR outcome

	SS	*Df*	MS	*F*	*P*	R^2	*Adjusted* R^2
Regression	837.282	3	279.094	1.694	.179	.085	.035
Residual	9063.362	55	164.778				
Total	9900.644	58					

References

Adams, M. A. (1990). *Beginning to read: Thinking and learning about print.* Cambridge, MA: MIT Press.

Amrien, A. L., & Berliner, D.C. (2002). *The impact of high-stakes tests on student achievement performance: An analysis of NAEP results in the states with high-stakes tests and ACT, SAT, and AP test results in states with high school graduation exams*. Retrieved from Arizona State University, (EPSL-0211-126-EPRU) Education Policy Studies Laboratory website: http://edpolicylab.org.

Amrien, A. L., & Berliner, D. C. (2003, February). The effects of high-stakes testing on student motivation and learning. *Educational Leadership, 60(5)*. 32-38.

Au, W. (2011). Teaching under the Taylorism: High-stakes testing and the standardization of the 21st century curriculum. *Journal of Curriculum Studies, 43(*1), 25-45.

Assaf, L. (2006). One reading specialist's response to high-stakes testing pressures. *The Reading Teacher, 60*(2), 158-167.

Atwell, N. (2007). *The reading zone*. New York, NY: Scholastic

Baker, M., & Johnston, P. (2010). The impact of socioeconomic status on high-stakes testing reexamined. *Journal of Instructional Psychology, 37*(3), 193-199.

Berliner, D. (2011). Rational responses to high stakes testing: The case of curriculum narrowing and the harm that follows. *Cambridge Journal of Education, 41*(3), 287-302.

Cutting, L. E. & Scarborough, H. S. (2006). Prediction of reading comprehension: relative contributions of word recognition, language proficiency, and other cognitive skills can depend on how comprehension is measured. *Scientific Studies of Reading, 10*(3), 277-299.

Edwards, A. T. & Pula, J. J. (2011, Summer). Back to high school: A teacher educator's hands-on encounter with the pressures of high-stakes testing. *Delta Kappa Gamma Bulletin*, *77*(3) 11-14.

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

Guskey, T. R. (2003, February). How classroom assessments improve learning. *Educational Leadership*, *60*(5) 7-11.

Hale, B., & Fiorello, C. A., (2010). *School neuropsychology: A practitioner's handbook*. (2nd Ed.). New York, NY: Guilford Press.

Henson, R. K., (2001). Understanding internal consistency reliability estimates: A conceptual primer on coefficient alpha. *Measurement and Evaluation in Counseling and Development*, *34*, 177-189.

Hintz, J. M., & Silberglitt, B. (2005). A longitudinal examination of the diagnostic accuracy and predictive validity of the R-CBM and high-stakes testing. *School Psychology Review, 43*(3), 372-386.

Keenan, J. M., Betjemann, R. S. & Olson, R. K. (2008). Reading comprehension tests vary in the skills they assess: Differential dependence on decoding and oral comprehension. *Scientific Studies of Reading, 12*(3), 281-300.

Lobascher, S. (2011). What are the potential impacts of high-stakes testing in literacy education in Australia? *Literacy Learning: The Middle Years*, *19*(2), 9-19.

Madaus, G., & Russell. M. (2010). Paradoxes of high-stakes testing. *Journal of Education*, *190*(1), 21-30.

Marder, M., Bansal, D., & Kadanoff, L. P. (2009). Flow and diffusion of high-stakes test scores. *Proceedings of the National Academy of Science of the United States of America*, *106*(41), 17267-17270.

Mason, E. J. (2007). Measurement issues in high-stakes testing: validity and reliability. *Journal of Applied School Psychology, 23*(2), 27-46.

National Institute of Child Health and Human Development. (2000). *Report of the National Reading Panel. Teaching children to read: An evidence-based assessment of the scientific research literature on reading and its implications for reading instruction* (NIH Publication No.00-4769). Washington, DC: U.S. Government Printing Office.

No Child Left Behind Act, 20 U.S.C. § 6319 (2001).

Pennington, B. E. (2009). *Diagnosing learning disorders*, (2nd Ed.). New York: NY: Guilford Press.

Shaywitz, S. E. & Shaywitz, B. A. (2007, March). What neuroscience really says about reading instruction. *Educational Leadership*, *64*(6), 73-76.

Shepard, L. A. (2002, Winter). The hazards of high stakes testing. *Issues in Science and Technology Online*, Retrieved from http://www.issues.org/19.2/shepard.htm.

Waber, D. P., Gerber, E. B., Turcios, V. Y., Wagner, E. R., & Forbes, P. W. (2006). Executive functions and performance on high-stakes testing in children from urban schools. *Developmental Neuropsychology, 29*(3), 459-477.

Wiederholt, J. L. & Bryant, B. R. (2001). *GORT-4: Gray oral reading test*. Austin, TX: Pro-Ed.

Zhao, Y. (2012). *World class learners: Educating creative and entrepreneurial students*. Thousand Oaks, CA: Corwin Press, A Sage Company.