2015

Evaluating Technical Adequacy Features of Sentence Verification Technique as a General Outcome Measure of Content Knowledge

Ren&eacutee E. Lastrapes
Louisiana State University and Agricultural and Mechanical College

EVALUATING TECHNICAL ADEQUACY FEATURES OF SENTENCE VERIFICATION TECHNIQUE AS A GENERAL OUTCOME MEASURE OF CONTENT KNOWLEDGE

A Dissertation

Submitted to the Graduate Faculty of the
Louisiana State University and
Agricultural and Mechanical College
in partial fulfillment of the
requirements for the degree of
Doctor of Philosophy

in

The College of Human Sciences and Education

by
Renée Elissa Lastrapes
B.A. Louisiana State University, 1992
M.Ed. Louisiana State University, 1993
August 2015

ACKNOWLEDGEMENTS

I would like to thank everyone who helped me on this journey to achieve a lifelong goal. First, thanks to my committee, my co-chairs Dr. Eugene Kennedy and Dr. Paul Mooney, and Dr. Keena Arbuthnot, Dr. Brian Marx, and Dr. Anna Long. Dr. Mooney, thank you for including me in the grant and giving me the opportunity to investigate this topic, and thank you for your humor and patience as you endured my many phone calls, texts, and emails over the past months as I progressed through this study. Thank you also for your editing and proofreading, your writing is an inspiration! Dr. Kennedy and Dr. Marx, thank you so much for your statistical advice, guidance, and patience as I grappled with the analysis for this project. I have learned so much from the both of you! Dr. Arbuthnot, thank you for your help in conceptualizing this project, I remember very well that conversation in your office when you advised me to pick one thing about the grant and make it my dissertation…and here it is! Dr. Long, it has been a pleasure to have you as the Dean's representative, and your expertise has been a wonderful complement to the committee.

Thanks go to Dr. James Royer who graciously responded to my email requests for documents I was not able to access, and for generously giving me an annotated bibliography he had maintained throughout his career regarding studies pertaining to sentence verification technique. I look forward to sharing this dissertation with you!

Many thanks go to the wonderful people in the parish who were involved in this study, and especially to a very special and hardworking teacher, Holly Benedetto, who created the instruments used in this study. You are a wonderful teacher, a tireless worker, and I am proud to consider you a friend! Thanks go also to Dr. Dawn Washington and the other administrators and teachers who helped facilitate a smooth process for data collection. Thank you to Geralyn

Callegan who made sure that teachers were administering the monthly SVTs. Thank you to Mr. Ellis and Mr. Robert who made sure the links were always active on the parish website!

Finally, I would like to thank my family for their support, patience, and encouragement throughout this process. Thank you to my partner, Kay, for her patience and proofreading and for supporting our family while I pursued this goal. Thanks go to my mother Jenny for all those nights while I was at school, being there when the children got off the bus and making dinner for us all. Thank you to my children, Alexander and Nora, for being patient with me during those many weekends I spent in front of the computer or studying. Also, thanks to my sister Robin and my brother Lee, with whom I talked on the phone many times during that hour-long drive to and from LSU, and finally thank you to my wonderful neighbor and friend Brent, who watched my children so many days when I had to go to LSU during the summer. I share this with all of you.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

ABSTRACT

Once students have mastered the mechanics of reading, they are expected to learn new material by reading. This new material, however, becomes increasingly more complex as students enter upper elementary and especially middle and high school. If students fail to comprehend what they read, they risk failure in content courses such as science and social studies. Early assessment of risk and appropriate response to that risk is a goal of effective education. One problem with the risk reduction sequence is that there are limited formative assessments that have been validated as technically adequate for assessing content knowledge.

The present study examined an established reading comprehension assessment called sentence verification technique (SVT) as a formative measure of science content knowledge. SVT probes were administered to 130 fourth, fifth, and sixth grade students at 2 PK-6 schools for 5 months, as well as the abbreviated Stanford Achievement Test, 10th Edition, as a criterion measure. Monthly SVT probes were analyzed for internal scores consistency reliability, as well as for predictive and criterion validity. Multilevel modeling was used to determine if SVT was a significant predictor of student growth. Item types were examined to determine if there were significant differences in scores based on race or gender.

Results indicated that SVT probes had internal consistency reliability estimates that ranged from .45 to .84, and criterion validity estimates ranged from .33 to .53. Sentence verification technique was found to have predictive validity for fifth and sixth grade, accounting for 24% to 40% of the variability in the criterion measure. Estimates for fourth and fifth grade showed that SVT was a significant indicator of growth. Finally, item analysis showed that there were marginally significant differences for gender and highly significant differences for race on items by type.

SVT shows potential for use as a general outcome measure of content. While it has been shown to demonstrate internal reliability, predictive and criterion validity, and growth measurement capacity, more research is needed. Findings to date suggest that given the more complex nature of instruction in content, SVT may work best in combination with other validated general outcome measures, including those with academic language indicators such as vocabulary matching or critical content monitoring.

CHAPTER 1: INTRODUCTION

It could be argued that a stated desire for change – read by political and/or professional change agents to mean improvement – is a constant in today's American educational landscape. In 2009, as a recent example, the National Governors Association (NGA) and the Council of Chief State School Officers (CCSSO) convened a group of stakeholders to develop the Common Core State Standards (CCSS; NGA & CCSSO, 2010). These new educational standards were purportedly developed across the states to ensure that students graduating from high school were ready for entrance into institutes of higher education or the workforce (NGA & CCSSO). In this context of change, as of the fall 2014, 46 states and a number of U.S. territories had adopted the CCSS with at least two states subsequently withdrawing their support, contributing to public and political expressions of concern with the standards (Bidwell, 2014).

One of the areas of focus in terms of better preparing students for college and career has been reading improvement. Reading for comprehension is a necessity for success in school as well as in life. The RAND Reading Study Group (2002) suggested that reading involves a reader, process, and purpose and defined reading comprehension as "the process of simultaneously extracting and constructing meaning through interaction and involvement with written language" (p. 11). Reading improvement for public school students has long been a concern for a number of educational constituencies.

According to the most recent Nation's Report Card (NRC; 2013), only 34% of public school students in reading performed at or above proficient in grades four and eight. Generally, to be considered proficient, students must demonstrate solid academic performance and competency over challenging subject matter (NRC). Guidelines for reading in the CCSS address complex texts and academic language. Students, CCSS documents assert, are expected

to develop reading comprehension skills as they progress through the grades, with the focus on academic vocabulary from the content areas. Students are supposed to read carefully and understand information based on evidence in the text as well as answer questions requiring inferences based on careful attention to informational text (NGA & CCSSO, 2010). Reading for evidence and making inferences are reading behaviors that are expected to build student knowledge.

Difficulties resulting from the CCSS' reportedly higher standards are exacerbated in content courses such as science and social studies. As students progress through school, the texts that students encounter become more difficult. Texts in the content areas of science and social studies have been described as inconsiderate of student and teacher needs (Espin, Busch, Lembke, Hampton, Seo, & Zukowski, 2013), with an overabundance of difficult vocabulary that introduce a wide range of topics within a short amount of time. In addition, with the CCSS, the expectation is now that students will not only be able to comprehend content-area texts, but use their comprehension to interpret and integrate this knowledge of the content areas.

In order to determine whether students comprehend, interpret, and integrate content knowledge, it is critical that there be effective assessment frameworks. In the present context it is essential that educators effectively assess what content knowledge students gain as a result of reading. Particularly in upper elementary and secondary school environments where students are taught less how to read and are expected to learn new content as a result of reading, knowing what students know becomes essential. Traditional assessments such as teacher-made tests have been shown to lack psychometric standards of technical adequacy (Fuchs & Fuchs, 1999; Tindal, Fuchs, Fuchs, Shinn, Deno, & Germann, 1985). States have implemented large-scale accountability tests, but these summative assessments are not useful for guiding instruction

because they happen once throughout the year, with the results usually not available until after the school year has ended (Fuchs & Fuchs; Linn, 2002). It is critical, then, for teachers to use assessment formatively to guide their group and individual student instruction and/or intervention practices (Fuchs, Fuchs, Hosp, & Jenkins, 2001).

General outcome measurement (GOM) is an instructional assessment framework that was designed specifically to document academic learning. It is a type of formative assessment that targets an entire school year curriculum and what a student should know or demonstrate by the end of a grade or subject. As part of the framework, purportedly equivalent measures known as probes are administered periodically to determine end-of-year competence in a subject area (Fuchs & Deno, 1991; Mooney, McCarter, Schraven, & Callicoatte, 2013). General outcome measurement uses generic stimulus materials for the probes rather than specific texts used by the teacher. To date, several general outcome measures have been investigated in content courses. These include maze (MZ; a cloze-type procedure where every seventh word is omitted and replaced with three choices; students show reading comprehension competence by selecting the correct word); vocabulary matching (where students match content-area vocabulary words to their respective definitions); and oral reading fluency (ORF; where students read aloud for one minute, and words read correctly are counted as the student's score) (Busch & Espin, 2003; Wayman, Wallace, Wiley, Tichá, & Espin, 2007).

Another comprehension tool, Sentence Verification Technique (SVT; Royer, Hastings & Hook, 1979), is now receiving attention as a potential general outcome measure after previously being extensively studied as a measure of reading and listening comprehension. In a study conducted by Marcotte and Hintze (2009), SVT was added to a suite of formative assessments of reading comprehension, namely ORF, MZ, retell fluency (RTF), and written retell (WRT) with

fourth grade students. They examined these measures for incremental and concurrent validity. Results showed that MZ, SVT, and WRT added to the variance associated with reading ability when combined with ORF. Findings further indicated that these measures were reliable indicators of a student's performance on a state accountability test and a standardized measure of reading. More recently, Mooney, Lastrapes, Marcotte, and Matthews (2015) evaluated the effectiveness of SVT, WRT, and an adaptation of vocabulary matching known as critical content monitoring as predictors of achievement in content courses. Results indicated that both critical content monitoring (CCM) and SVT were statistically significant predictors of achievement across fifth grade science and social studies content.

The potential promise of SVT as an objective indicator of student academic performance and progress prompted the present inquiry. With the CCSS call for student comprehension, interpretation, and integration of content en route to college and career readiness, it is critical that educational stakeholders have useful assessment frameworks to document learning. While GOM provides a vehicle for documenting student performance and progress, its emphasis historically has been directed primarily at reading comprehension and secondarily at math understanding. There has been a dearth of scholarship targeting content learning. The present study expanded the breadth of SVT research by examining its technical characteristics as a measure of performance and progress. Establishment of SVT as a general outcome measure of content learning in social studies and science may provide teachers with another tool to inform instructional decision making, including helping teachers determine what students know, how students are progressing, and what group and/or individual instructional decisions teachers need to make in the present CCSS culture wherein focus is directed at understanding complex informational text.

Before providing a summary of the SVT literature and elaborating on the rationale for the present study, the following key terms and accompanying definitions are listed.

Definition of Terms

- Sentence Verification Technique (SVT): It is a measure of reading and listening comprehension based on the theoretical assumption that comprehension is a constructive process that entails an interaction between the context, linguistic message, and knowledge base of the listener or reader.
- Formative assessment: These are assessment procedures teachers use throughout the year in order to modify teaching and learning activities to improve student achievement both individually and collectively.
- General Outcome Measurement (GOM): This is a type of formative assessment that assesses proficiency on general outcomes from which the entire curriculum is focused rather than on specific skills; it is assessment of what the teacher wants the students to have mastered at the end of the year.
- Criterion validity: This is a measure of how well scores from a new or predictor measure compare with those from other "gold standard" or already established tests in the same subject area.
- Predictive validity: This is a measure of the extent to which a score on a test predicts scores on a criterion measure.
- Reliability: This refers to the overall consistency of a measure; a measure is considered to have high reliability if it produces similar results with repeated administrations of the measure.

CHAPTER 2: REVIEW OF LITERATURE

In introducing SVT, Royer and Cunningham (1978; 1981) presented a theory of comprehension with the stated goal of determining the best way to measure or assess reading or listening comprehension. (Reading comprehension is emphasized herein.) The model suggested that the act of comprehension must involve an interaction between the incoming linguistic message and a reader's world knowledge. Comprehension was described as a constructive process resulting from the interaction between context, the linguistic message, and prior knowledge of the reader. The model posited that the construction of meaning from an incoming linguistic message maintained the meaning of the message but not necessarily its exact structure (Kintsch & van Dijk, 1978; Royer, Hastings, & Hook, 1979).

Relative to assessment, this meaning representation could then be measured by determining if readers had established a memory from something read. Sentence verification technique was designed to determine this (Royer, 2004). Royer et al. (1979) claimed that if readers had comprehended a text, and established a meaningful memory of what they had read, then they should be able to determine if a paraphrased sentence from the original passage preserved the meaning of the original sentence from which it came. Similarly, they proposed that a reader should be able to determine that an exact copy of a sentence was the same as in the original passage. On an SVT test, the participant would read a passage, then without the passage, select "yes" if the meaning is preserved, and "no" if the meaning is not preserved, to a series of test items. If presented with a sentence that was slightly different or opposite from the original passage then the reader should reject this sentence as meaning the same thing as the original passage. Similarly, the reader should be able to determine that a distractor sentence (one

similar in structure but not from the original passage) did not have the same meaning as the original.

Royer and his associates reasoned that if a test contained all four types of sentences (an original, a paraphrase, a meaning change, and a distractor), a strategy of relying on similar or different wording in the test items would not work (Royer et al., 1978; 1981; Royer, 2004). Test items that have different wording than an original passage sentence sometimes have the same meaning (paraphrases) and sometimes have a different meaning (distractors). Test items that appear identical to the original sometimes mean the same (original) and sometimes do not (meaning change). Royer et al. (1979) reasoned that including all four types of items eliminated the reliability of cues in the syntax or structure of the sentence as hints to selecting the correct answer. In order to achieve on the SVT test, a reader must successfully make a correct memory representation of what he or she has read.

In the literature an SVT test consists of two to six passages and a set of test sentences that includes an equal number of each of the sentence types (Royer, Greene & Sinatra, 1987; Royer et al., 1979). In school-based assessments, the reading level of the passages typically overlaps the target population. For example, if administering a three-passage assessment to a group of sixth graders, the first passage would be a fifth grade passage, the second, a sixth grade passage, and the third, a seventh grade passage. If the test is a 16-sentence test, then four originals, paraphrases, meaning-changes, and distractors would be constructed. Once the sentence types are developed, they are randomly arranged in the test with the caveat that the first eight items come from the first half of the sentences in the passage. This restriction is to avoid having the first sentence that an examinee encounters come from the last sentence in the passage that had just been read. This seemingly eliminates the possibility that a correct answer could potentially

be a product of short-term memory rather than a meaningful memory representation having been established in the reader's mind (Royer et al., 1987; Royer, Carlo, & Ciserco, 1992).

SVT tests have been scored in three ways in the literature. The first method has been to compute proportion correct scores (Royer et al., 1992) which can be calculated for overall performance, performance on each passage, or performance on sentence item type. A second scoring technique is to use signal detection theory (Swets, Tanner, & Birdsall, 1961). Signal detection theory is dependent upon two criteria: The ability to detect a signal when it is present (i.e., decide if a test sentence has the same meaning as a passage sentence), and the criteria which the subject establishes in order to judge if a signal is present (i.e., say a test sentence is a yes sentence) (Royer et al., 1992). The technique involves the use of signal detection analysis, the precursor to what is now commonly referred to as receiver operating characteristic (ROC) curve analysis (Swets, 2014). In this type of analysis, the purpose is to establish whether a signal detection parameter (what Royer refers to as *d'*) is distinguishable from what Royer refers to as *c*, the cutoff parameter. A third more recent scoring method, utilized by Marcotte and Hintze (2009) and Mooney et al. (2015), counted the number of test items marked correctly.

The research on SVT in its capacity as a determiner of reading comprehension is extensive. The reliability of SVT tests has been summarized in various studies (Royer, 2004; Royer, Carlo, Cisero, 1992; Greene, Royer & Anzalone, 1990; Marchant, et al, 1988; Royer & Carlo, 1991a; Royer, Sinatra, Greene, & Tirre, 1989). Cronbach's alpha has been calculated for SVT made of three passages with 16 items each (48 total) have coefficients from .5 to .6; SVT with four passages (64 items total) have coefficients from .70 to .80; and SVT with six passages have coefficients ranging from .80 to .90. Linear relationships for reading comprehension tasks have been found to be consistently higher than for listening comprehension (Royer, 2004). For

example, Ulusoy and Cetinkaya (2012) reported Kuder Richardson-20 statistics from .67 to .74 for a 40-item measure. To date, reliability estimates have not been reported for two-passage SVTs (Royer, 2004).

Validity research has emphasized comparisons with standardized criterion measures. Correlations with the Iowa Test of Basic Skills (fifth and sixth grade students) was 0.73; with the California Achievement Test .52 (fourth and sixth grade students) and with Stanford Achievement Test (fourth and sixth grade students) correlation was 0.50 (Royer, 2004). Other research studies have demonstrated that SVT has been shown to be sensitive to text difficulty (Royer, et al., 1979; Royer, Kulhavy, Lee, & Peterson, 1986) and different skill levels of reading (Rasool & Royer, 1986; Royer et al., 1979; Royer et al., 1986). Royer, Lynch, Hambleton, and Bulgareli (1984) also provided evidence that SVT is a measure of passage rather than sentence comprehension.

Other research has targeted SVT's diagnostic utility. For example, Carlisle (1989a; 1989b) argued that reading and listening comprehension measures should be included in any diagnostic assessment of students with comprehension difficulties. She used SVT as a measure of both listening and reading comprehension, and analyzed the errors on the four different types of SVT items to determine where the students' comprehension problems lie. She found that good comprehenders showed similar comprehension problems when examined in both the listening and reading tests (Royer et al., 1992). Poor comprehenders tended to answer incorrectly on originals and meaning change items when part of both the listening and reading assessments. They tended to answer incorrectly on paraphrases when part of the reading test and distractors when part of the listening test. From her findings Carlisle (1989b) suggested that SVT could be used to identify readers' difficulties in comprehension, essentially laying out an

intervention formula. That is, SVT could be used formatively in that the way students responded to certain test items could guide teachers on how to adjust their instruction in order to ensure that students continue to make progress.

Recently, the SVT instrument has been utilized across a number of literacy-related subjects. For example, Harper (2014) used SVT in creating a health literacy assessment tool for young adults with a sample of 144 undergraduate students. Using item response theory and goodness of fit statistics, it was determined that of 20 comprehension questions, eight SVT items and 12 cloze (a reading technique in which an examinee reads a passage with blanks and supplies the missing word) items were retained in the final assessment. Item discrimination and difficulty were also investigated resulting in the elimination of six of the eight SVT items and two cloze items due to poor item response theory (IRT) discrimination values (below 0.3). They recommend further research with a much larger sample.

In a 2 x 2 experimental design study, Marchand, Nardi, Reynolds, and Pamoukov (2014) manipulated the temperature, ventilation, lighting, and acoustics of a testing room to determine if these parameters set to comfortable levels or just outside the comfort zone affected the learning, mood, and perceptions of environmental influence on the performance of undergraduate students on listening and reading tasks. One of the comprehension measures was a 40-sentence SVT passage with 10 of each type of sentence item. Results indicated that participants in the environment outside of the comfort zone had lower scores on a listening comprehension test than those in the normal listening condition, but that no difference was detected between conditions for reading comprehension.

In a study completed for a dissertation, a researcher used SVT to examine the effects of video-based peer modeling on the question asking, reading motivation, and text comprehension

of struggling adolescent readers (Tsikalas, 2012). Jones and Smith (2014) compared the use of SVT and cloze tests to a variant of SVT called Meaning Identification Technique (MIT), (which only presents paraphrases and meaning change items), and C-tests which are a variant of cloze. Like the cloze procedure, C-Tests measure the reader's capacity to predict information from context. Unlike in Cloze tests, in the C-Test the words in the text are only partly deleted. These assessments were used to investigate the understandability of accounting documents. They found that the SVT and MIT tests were measures of the understandability of accounting tests, but what the Cloze tests and C-tests was measuring was uncertain. The sample in Jones and Smith was relatively small (44 participants) and the authors suggested that further research in SVT/MIT as a measure of accounting document understanding was warranted.

General Outcome Measurement

The central focus of the present inquiry is SVT's potential to serve as a general outcome measure. General outcome measurement (GOM) is one of two models of "instructionally relevant measurement" (Fuchs & Deno, 1991, p. 488) that are described in the literature, the other being mastery measurement. General outcome measurement targets an entire curriculum domain (e.g., sixth grade science) and develops equivalent tests that sample from the whole domain and indicate end-of-year skill or subject competence. Two approaches have been used in developing general outcome measures (Fuchs, 2004; Fuchs, Fuchs, & Zumeta, 2008). One method has been to identify robust indicators of curricular proficiency, which are capstone tasks such as oral reading fluency that provide strong correlations with the component skills that comprise the relevant academic domain. The other practice has been to systematically sample from the skills that comprise the annual curriculum (e.g., mathematics concepts and applications) in such a way that each probe represents an equivalent snapshot of the grade-level curriculum.

No matter the probe development approach, ongoing assessment using equivalent measures indicates both student performance at a moment in time and growth over time, thereby providing a system that is believed to be sensitive to instruction over the long term.

Curriculum-Based Measurement (CBM; Deno, 1985) is an example of a GOM system. Curriculum-Based Measurement has an extensive literature supporting its function as a method for assessing the growth of students in basic academic skills (Fuchs, 2004). It has progressed to the point where it "has been proposed as a means for predicting performance on and monitoring progress toward rigorous, state-defined academic standards for individual students" (Wallace, Espin, McMaster, Deno, & Foegen, 2007, p. 66). Yet although hundreds of studies over the past four decades have been aimed at demonstrating the efficacy of reading, spelling, writing, and math measures, predominantly in the elementary grades, there has been a relative dearth of research targeting the utility of content-focused CBM/GOM assessments (Burns, Scholin, & Zaslofsky, 2011).

General Outcome Measurement for Content Courses

In the content areas, there have been five measures other than SVT that have received scholarly attention as potential general outcome measures. From that list, which includes concept maze (Ketterlin-Geller, McCoy, Twyman, & Tindal, 2006), mainstream consultation agreements (Tindal & Germann, 1991; Tindal, Parker & Germann, 1990), key vocabulary progress monitoring (Vannest, Parker, & Dyer, 2011), and critical content monitoring (Mooney et al., 2013), the measure with the largest research base is vocabulary matching (Espin, Busch, Shin, & Kruschwitz, 2001; Espin & Deno, 1993, 1994-1995; Espin & Foegen, 1996; Espin, Shin, & Busch, 2005). Vocabulary matching is an example of an assessment of academic language. Early vocabulary matching probes have traditionally included randomly selected vocabulary

terms and their accompanying definitions that were drawn from a classroom textbook and teacher notes and presentation materials (Espin et al., 2001). Paper-pencil versions of the probes have included vocabulary terms that were placed in alphabetical order on the left side of a page with the accompanying definitions organized in a random order on the right side. Students have been instructed to match the term with the appropriate definition and given five minutes to complete the task. Probes have been scored according to the number of correct matches in the time frame (Espin et al., 2001).

Stages of General Outcome Measurement Research

In her article regarding the past, present, and future of CBM research, L. S. Fuchs (2004) delineated three stages of CBM/GOM research as it relates to evaluating the tenability of any measure. For technical adequacy purposes specifically, that means the reliability and validity of the measure. These stages can serve as a guide to conducting research regarding different GOM techniques. In stage 1, technical features of the static score are investigated. This is the stage where the reliability and criterion validity of a GOM measure are established with scores obtained at one point in time. Fuchs described stage 2 as studies that examine the "technical features of the slope" (p. 189). In stage 2 research, investigators attempt to establish that repeated measures of a GOM show academic growth over time. It is in this stage that an individual's growth rate, as measured by the slope, is evaluated to show academic improvement. In stage 3 of GOM research, Fuchs described studies that investigate the instructional utility of GOM measures. This stage is concerned with how information obtained by GOM measures are utilized by practitioners to monitor student achievement and initiate change in their teaching practices when students are failing to progress. Fuchs stated that there is very little recent research in this stage of GOM, with most of the research focusing on stage 1. She made the

claim that this may be because research in stages 2 and 3 is "more laborious, requiring ongoing data collection for stage 2 and additionally necessitating practitioners' data utilization for stage 3" (p. 191).

Vocabulary Matching

The literature to date on vocabulary matching is unique in the content areas in that it is the only content focused general outcome measure that has technical adequacy issues findings related to its static score and slope. A summary of Stage 1 and 2 statistical findings offered promise for the use of GOM content-focused tools for measuring academic learning. Researcher-developed vocabulary matching probes: (a) had mean scores that were moderately correlated with multiple criterion measures including standardized subject matter and general knowledge and statewide accountability tests (Espin et al., 2001; Mooney, McCarter, Schraven, & Haydel, 2010); (b) had criterion-related correlations with a statewide accountability test that were significantly stronger than were linear relations between the criterion and general outcome measures of reading and writing (Mooney, McCarter, Schraven, et al., 2013); (c) shared the greatest proportion of variance with the statewide test score in predictive regression models, with unique variance attributed to it even when a standardized measure of general vocabulary was included in the model (Mooney, McCarter, Schraven et al.); (d) evidenced statistically significant growth patterns for weekly probe administrations for time periods ranging from 11 to 24 weeks (Borsuk, 2010; Espin et al., 2005; Mooney, McCarter, Schraven, et al.); and (e) demonstrated strong interrater reliability correlations across studies (Borsuk, 2010; Espin et al., 2001; Mooney et al., 2010; Mooney, Schraven, & Cox, 2010).

A recent online adaptation of critical content monitoring (Mooney, McCarter, Russo, & Blackwood; 2013, 2014) has been a target of recent content area inquiry. Critical content

monitoring is a multiple-choice oriented content test that is administered online to students with content selection and timing features that are identical to vocabulary matching design. Mooney, McCarter, Russo et al. (2013) evaluated the criterion validity and passage equivalence capacity of the online adaptation in science content and reported moderately strong correlations with a statewide accountability content test and that probe correlations were and mean scores were not equivalent for a population of generally high achieving fifth-grade students. Mooney et al. (2014) replicated the moderately strong correlations in social studies content and also reported similar magnitude correlations for probes that differed in terms of length, time, and content makeup.

SVT as a General Outcome Measure

The rationale for SVT's inclusion in the content area GOM literature originated with a reading comprehension focused study conducted by Marcotte and Hintze (2009). Marcotte and Hintze examined the incremental and concurrent validity of four different formative measures of reading comprehension when combined with ORF: Maze, RTF, WRT, and SVT. They hypothesized that these formative assessment measures would account for variability in reading proficiency, as indicated by two criterion measures of reading comprehension, beyond what was accounted for by ORF alone. They found that of the four measures, only RTF did not significantly contribute to the estimation of the variance in the criterion measure. The other three measures (i.e., MZ, SVT, and WRT), in combination with ORF, were reliable indicators of student performance on a grade-level literacy test and a state criterion-referenced test.

Findings contrasting with those of Marcotte and Hintze (2009) were reported by Christ, White, Ardoin, and Eckert (2013). While not the primary target of research focus, Christ et al. examined the incremental validity of an alternate form of SVT as a reading comprehension

measure in conjunction with ORF. Findings indicated that SVT did not add significance to the model predicting academic performance for students in grades two through five.

Although general outcome measures are typically timed measures, there have been instruments utilized that were not timed. Vannest, Parker, and Dyer (2011) developed key vocabulary probes, which is a combination of vocabulary matching presented in "cloze" sentences with the missing term the key vocabulary word. Students read the cloze sentences and selected the appropriate word. Probes were presented via computer and were untimed. Marcotte and Hintze (2009) used SVT probes that were also untimed while indicating that the 4-passage SVT assessments utilized in their study required 30 minutes to administer.

The lone study to date examining the utility of SVT as a measure of content comprehension was an extension of Marcotte and Hintze (2009). Mooney et al. (2015) made comparisons of SVT, WRT, and critical content monitoring to a standardized measure of content achievement in science and social studies, the abbreviated online Stanford Achievement Test-10th Edition. Findings demonstrated that both SVT and critical content monitoring correlated moderately with both content tests and contributed uniquely to regression models documenting variability accounted for. In both cases, critical content monitoring accounted for the greatest share of variability. However, commonality analyses documented that while the two measures shared variance, that each contributed uniquely to science and social studies models.

Rationale for the Study

Given today's CCSS emphasis on student comprehension, interpretation, and integration of content as a precursor to college and career readiness, there remains a need for formative assessment in the content areas. General outcome measurement has a demonstrated track record in terms of regularly identifying students performing at expected levels as well as those who are

at risk for new or continued academic failure. Responsiveness to intervention (RTI) preventative frameworks have become common in schools, particularly in the elementary grades and in the areas of reading development and comprehension. Stereotypical systems involve brief, standardized assessment of all students to differentiate students on track from those at risk for falling or remaining behind.

When implemented as intended, school systems provide students identified as at risk for academic failure with research-validated interventions in areas of targeted need. These students are assessed more frequently to determine if the core plus supplemental instruction and intervention program is reducing risk of academic failure. The assessment systems utilized in RTI frameworks – be it periodic benchmark testing of all or progress monitoring of those at risk – generally rely on general outcome measures to document performance and progress. Such systems, which evolved from individual student assessment of students with individualized education programs in the 1970s to today's use with all students – were designed to be practical, efficient, inexpensive, and technically sound (Deno, 1985). Assessment frameworks in the area of content learning and in the upper elementary and secondary grades are significantly less developed than those targeting reading development and those for the lower elementary grades (L.S. Fuchs & Vaughn, 2012).

The SVT assessment holds promise as a potential determinant of student performance and progress for use in RTI systems. First, it has an evidence base as a reliable and valid indicator of reading comprehension, which allows it to serve multiple purposes and subject areas, thus making it an efficient determinant of learning for upper elementary and secondary school teachers. Second, these teachers who generally serve more students on a daily basis and have less instructional time per individual student than do lower elementary teachers and therefore

face greater demands on their time in administering and scoring many frequent formative assessments. The SVT is easily presented and scored in an online version, thus eliminating this challenge and possibly increasing its instructional utility, which is the goal of Stage 3 research. Third, SVT has a theoretical framework to support its implementation, giving researchers and practitioners the opportunity to evaluate systematically the utility of developing mental constructions of material read, as well as the ability of SVT to distinguish between good and poor comprehenders and serve the formative assessment role of informing instruction.

Fourth, since SVT requires that students read text-based content and determine whether subsequent statements directly relate to the read content or not, there is a chance that the assessment more naturally mirrors student expectation and learning action than do some of the other content-focused general outcome measures. Specifically, part of the expectation of students in content courses is that they read, understand, and use information provided in textbooks. Passages in SVT probes come directly from textbooks and the assessment process requires students to read first and make determinations as to reading material accuracy. Such actions may be closer to instructional reality than the demonstration of academic language knowledge that is required in vocabulary matching and critical content monitoring probes. The SVT actions may also be closer to instructional reality than student actions in completing MZ probes, in which content or non-content passages are adapted to include blanks that are replaced with multiple options for student choosing. Finally, and related to the formative assessment function indicated earlier, SVT scores have the potential to directly inform instruction, something that general outcome measures often lack the ability to do. Because students read content passages and then construct different types of comprehension representations, there is the possibility for performance patterns in SVT testing to provide teachers with ideas as to how to

proceed instructionally. Such a possibility enables teachers a chance to positively impact academic performance in reading and content comprehension areas. The listed reasons provide a rationale for continuing to explore the utility of SVT to serve as a general outcome measure in content courses beyond the criterion validity related research that has been conducted to date.

Research Questions

In the discussed context, the following research questions are provided:

Research Question 1: What are the distribution of scores for the SVT across the five months and for the SAT for each of the demographic groups?

Research Question 2: What is the internal consistency reliability of SVT measures?

Research Question 3: What is the predictive validity of SVT score(s) for performance on the SAT-10 online? What is the criterion validity of the SVT?

Research Question 4: Is there evidence for growth in probe scores over the duration of the study and what was the expected growth rate? Were there growth rate differences among student subgroups?

Research Question 5**:** Are there differences in item response patterns for the different item types in SVT based on race or gender?

CHAPTER 3: METHOD

Participants

District

The school district consisted of all public schools is an entire parish, located in a predominantly rural part of southern Louisiana less than 50 miles from a large urban area. According to data from the 2011-12 school year provided by the National Center for Education Statistics (NCES), the district was comprised of 4,569 students and 363 teachers with a student-teacher ratio of 12.59. In the district, 0.46% (n = 21) of students were classified as English language learners (ELL) and 9.8% (n = 450) as students with disabilities (NCES). The parish had a median household income of $44,000 and 17% of the parish lived below the poverty level (U.S. Census, 2008-2012).

Students

The target population of this study was all fourth, fifth, and sixth graders. The accessible population were those students in those three grades in science classes in two schools. Parish administrators provided permission to solicit students from two schools, one rural and one in a town of approximately 7,100. School A was a small prekindergarten (PK) through grade 6 school located in what is described by NCES as rural/fringe. The student population was approximately 300. School A had 88% of its students eligible for free or reduced price lunches. The student population was 78% African-American and 22% White. School B was a much larger PK-6 school, with a total population of 1,100 students. NCES described the locale as large suburban. Eighty-nine percent of the students at School B were eligible for free or reduced lunch. The racial makeup of the School B was 1 (0.09%) Native American, 6 (0.5%)

Asian/Pacific Islander, 11 (1%) Hispanic, 200 (17.4%) White, 927 (81%) African American, and 3 (0.26%) multi-ethnic.

Parent permission was solicited for all grades 4-6 students at School B first, and then School A. A sufficient number of grade 4 permission slips at School B led to the recruitment of only grades 5 and 6 students at School A. Once permission slips for the two schools were received, student assent procedures were undertaken. For participants, testing took place at School A on March 26-27, 2015, and School B on March 30-31. The demographic characteristics of the sample are provided in Table 1. School A had 17 fifth graders in one teacher's class and 10 sixth graders from another teacher's class. School B consisted of 50 fourth graders from two teachers' classes, 23 fifth graders from a single teacher's class, and 30 sixth graders from a single teacher's class. The students' ages ranged from 9.5 to 13.6 with the average student age 11.5 years.

Table 1. Demographic Characteristics of the Overall Student Sample by School

	Total		School A		School B	
	N=130	%	*N*=27	%	*N*=103	%
Gender						
Male	59	45.4	17	63	42	40.8
Female	71	54.6	10	37	61	59.2
Race/Ethnicity						
African American	123	94.7	24	89	99	96.1
White	5	3.8	3	11	2	94.6
Hispanic/Latino	2	1.5	0	0	2	1.5
Education						
General education	122	94	26	96.3	96	93.2
Special education	8	6	1	3.7	7	6.8
Socioeconomic status (SES)						
High SES	3	2.3	0	0	3	2.9
Low SES	127	97.8	27	100	100	97.1

The entire district's fourth through sixth graders (approximately 1,100 students) completed monthly SVT probes, but only 130 took the criterion measure. Because the second and the last research questions did not pertain to the criterion measure, the larger data set was used, however, for research question 5 only those students for whom there was complete data for the five months (N = 567) were included. The demographic summary for the sample can be found in Table 2. Analysis was limited to comparisons between gender and race/ethnicity, namely, African American students and White students.

Table 2. Demographics of Total Group for Item Analysis

	Total	%	4th Grade	%	5th Grade	%	6th Grade	%
Full Data Set	N = 567		N = 204		N = 200		N = 163	
Race / Ethnicity								
Asian	3	0.5	1	0.5	0	0	2	1.2
African American	359	63	119	58	142	71	98	60
American Indian	2	0.3	1	0.5	0	0	1	0.6
Hispanic/Latino	7	1.2	2	1	3	1.5	2	1.2
≥ Two races	3	0.5	0	0	0	0	3	1.8
White	193	34	81	40	55	28	57	35
Gender								
Male	289	51	111	54	103	52	76	47
Female	278	49	93	46	97	48	88	53
Socioeconomic status								
Low	468	83	165	81	175	87.5	128	77
High	99	17	39	19	25	12.5	35	33
Educational Services								
Regular Education	531	93.6	188	92	187	94	156	96
Special Education	36	6.7	16	8	13	6	7	4

Note. Item analysis by race was limited to African American and White due to the unequal sample sizes for the other race/ethnic groups.

Research Design

The present study employed a non-experimental correlational research design in order to establish the reliability and validity of SVT. Analysis of covariance (ANCOVA) was used to determine whether the SVT predicted achievement in the criterion measure, and partial

correlations were used to examine criterion validity. Multilevel modeling was used to determine if SVT was sensitive to student growth over time, and to examine differences in growth patterns in students by gender and teacher/school. Multivariate analysis of variance (MANOVA) was used to examine differences in responses to the four different item types by race and gender.

Instrumentation

Two measures were compared in the present study, one criterion measure and one predictor variable. The criterion measure was the science content test of the online abbreviated Stanford Achievement Tests-Tenth Edition (SAT-10; Pearson Education, n.d.). The predictor variable was SVT.

Criterion Measure – Stanford Achievement Test, 10th Edition

The abbreviated form of the online Stanford Achievement Test, 10th Edition (SAT-10) is a standardized, norm-referenced achievement test battery that measures reading, mathematics, spelling, language, listening, science, and social studies performance for students in kindergarten through 12th grade. The science test was described by publishers as aligned with national and state content standards. The test-derived scaled score was used in the present study. The scaled score is vertically equated across each subject test, reportedly allowing for the tracking of performance across grades (Pearson Education, n.d.). The science test assesses science as inquiry, knowledge of life, physical, and earth sciences. The abbreviated battery content test consisted of 30 multiple-choice questions and was untimed. Two online Buros Institute *Mental Measurements Yearbook* reviewers (Carney, n.d.; Morse, n.d.) provided support for the use of SAT-10 in measuring achievement in K-12 settings. Both described evidence of alternate-form reliability and content validity for the test as a whole. Mooney et al. (2015) reported a .64 correlation with the Louisiana state accountability science test in fifth grade.

Predictor Measure – Sentence Verification Technique

The SVT September probe was developed by the researcher. All subsequent probes were developed by the researcher and a fifth grade teacher who has highly qualified certification status in science and was trained in probe development. Internal reliability has not been previously reported for two-passage SVT tests (with 32 test items) to date. Reliability using Cronbach's alpha has been reported for three-passage SVT probes which ranged between .5 and .6 and for four-passage SVT probes which have ranged from .70 to .80 (Royer, 2004). The internal consistency of the probes for the current study were examined using Cronbach's alpha. Previous criterion validity correlations for a paper-pencil version of SVT were .46 (95% confidence interval [CI]; 0.21, 0.66) with the Louisiana Integrated Education Assessment Program (iLEAP) and .49 (95% CI; 0.25 0.67) with SAT-10 online science subtest (Mooney et. al, 2015).

Procedure

Students took five monthly assessments, September through February, excluding December. The assessments were delivered via Qualtrics, an online survey software system (Qualtrics Labs, n.d.). The reading levels of the passages were investigated using the Flesch-Kincaid scale (Kincaid, Fishburne, Rogers, & Chissom, 1975). Readability grade levels are reported in Table 3, with passages and corresponding items found in Appendix F.

Table 3. SVT Reading Level Passages as Measured by the Flesch-Kincaid Readability Scale

	Flesch-Kincaid Readability Level		
Month	Passage 1	Passage 2	Average
September	9.4	11.8	10.6
October	6.8	7.0	6.9
November	5.7	7.1	6.4
January	6.8	7.8	7.3
February	7.6	7.1	7.4

Passages were drawn from science texts used by the school district and topics covered were in accordance with Louisiana Grade Level Expectations (LDE, n.d.). The researcher inputed the assessment into the software, which then generated a web link. The web links were placed on the district's website for the classroom teacher to guide students through the test-taking process. Students typed in their name and grade and selected their teacher. They were then led through standardized directions and completed the test.

For the first day of testing in September, two researchers administered the tests together at the smallest school in the district to ensure that there was fidelity of implementation, and subsequently the researchers went to all schools and instructed the teachers in the testing process. After October, the district teachers administered the tests independently. Researchers periodically visited the schools throughout the school year to ensure that all teachers were consistently administering the assessments. No fidelity of test implementation checks were conducted.

Parental consent and youth assent forms were distributed at two schools to approximately 150 students in February. One hundred thirty-two students returned parent permission and assent to take the exam. During the testing window, the researcher delivered the test to 27 students at School A with the assistance of a classroom teacher. The tests were administered to 103 students at school B with the assistance of school staff for a total of $N = 130$. Two students whose parents had returned consent forms were absent on the day of testing and therefore not tested. Results were immediately reported to the researcher upon student completion of the assessment. Five SVT probes were analyzed (September, October, November, January, and February) for internal consistency reliability, as well as examined for their predictive validity and

growth in relation to this criterion measure. Achievement on SVT items by type (originals, paraphrases, meaning changes and distractors) were also examined by race and gender.

Data Analysis

Each month, the data were downloaded from the Qualtrics website using SPSS, Version 22. The SVT was then scored by item and a total score was computed. A master file of total scores was maintained in Excel and for each month there was a master in an SPSS file with item level information. Probe and SAT-10 online scores were analyzed using correlational methods as well as linear mixed modeling. Below listed are the research questions and accompanying data analyses.

Research Question 1: What were the distributions of scores for the SVT across the five months and for the SAT-10 online for each of the demographic groups?

The sample for this study was homogenous in terms of race/ethnicity, socioeconomic (SES) status, and educational classification. Due to this, the statistical analysis comparing groups was limited. In order to see the differences among the different demographic groups, mean scores and standard deviations, as well as 95% CIs of the monthly SVT probes were calculated for each of the groups by grade level.

Research Question 2: What was the internal consistency reliability of SVT measures?

In order to establish SVT for use as a general outcome measure, a number of questions related to issues of technical adequacy, its reliability and criterion validity, needed to be established, including whether it adequately models growth. In stage 1 research, the purpose of the study was to determine a single probe score's reliability and validity. For the present study, the internal consistency reliability of the measures themselves were determined using Cronbach's alpha for each month of a school year that equivalent SVT probes were

administered. This is a different approach to answering the question of reliability than what has usually been indicated in the GOM literature. Because this is what was reported in the literature regarding SVT, this is what was determined here. The goal of GOM is to create short probes. For the present study the researcher used two passages with 16 test items per passage. Two passage internal consistency reliability has not been established for SVT to date (Royer, 2004). The overall Cronbach's alpha was calculated for each monthly probe, as well as the Cronbach's alpha for each passage and corresponding 16 item test to determine if it is possible to further shorten the probes in accordance with GOM probe development.

Research Question 3: What was the predictive validity of SVT score(s) for performance on the SAT-10 online? What was the strength of the relationship between SVT and SAT-10 online?

The data were examined by grade level. The criterion measure SAT-10 online was administered at two participating schools; however for the fourth grade, all the students participating were from two different teachers' classes at one school. There were 50 students total. Because the students came from one school, and the group was largely homogenous, main effects for teacher, gender, and SVT score, as well as the respective interactions were examined.

For the fifth (N = 40) and sixth grades (N = 40), there was only one teacher per grade per school. Because the students came from one teacher per school, and the fact that the group was largely homogenous, main effects for school/teacher, gender, and SVT score, as well as the respective interactions were examined.

The data were analyzed with random effects for school using linear mixed modeling, however the model failed to converge. The model was then analyzed as an ANCOVA, with the SVT score as the covariate and school and gender as fixed effects. A separate ANCOVA was conducted for each month, in order to address the nature of the research question, that at each

month, do SVT scores predict achievement on the criterion measure? To average these scores over the year or to evaluate them in one calculation does not allow for the fact that these were monthly scores intended to guide instruction throughout the course of the school year. In September, the October score was not yet available, so it was not logical to evaluate them together. Similarly, in October, both September and October scores were available, but practically, teachers evaluate scores at each month. In addition, the equality of the probes had not yet been established. Therefore, the analysis was conducted separately by the month.

To establish criterion validity of the SVT in relation to SAT-10, the strength of the relationship between each monthly SVT score and the SAT-10 online was examined. Partial correlations were calculated, controlling for the effects of teacher and gender for the fourth grade and controlling for the effects of school/teacher and gender in fifth and sixth grades. Assumptions of all tests were addressed in the results section.

Research Question 4: Is there evidence for growth in probe scores during the study and what was the expected growth rate? Were there growth rate differences among student subgroups?

Multilevel modeling (MLM) was used to determine the significance of the mean growth rate of participants, the variability in growth rate among students, and the difference in growth rate between students by subgroup. While the lowest level of data in MLM can be the individual, MLM can be used to analyze repeated measurements of individuals (Tabachnick & Fidell, 2013). Multilevel modeling for repeated measures can be used in the presence of missing data as well as when the time periods between measurements are not equal, features which make it a more desirable choice than repeated-measures analysis of variance (Bell, Ene, Smiley, & Schoeneberger, 2013).

Research Question 5: Are there differences in item response patterns for the different item types in SVT based on race or gender?

In order to address this question, the monthly scores for each item type originals, paraphrases, meaning changes, and distractors were totaled. For each monthly assessment, there were eight of each item type. Mean scores were calculated for each month by item type overall. To better see overall trends, composite scores were calculated by item type, giving a possibility of 40 correct responses (8 per month x 5 months) per item type. These composite scores were calculated for the overall group. They were then compared by race and gender and the interactions of race and gender. Bar graphs (Figures 1 – 8) display these results.

Correlations were also calculated for the overall sample with composite items to examine the nature of the relationship between the types of items. Originals and paraphrases have "yes" as the correct answer while meaning changes and distractors have "no" as correct answers. Correlations for all subgroups were calculated but they mirrored the overall composite very closely, therefore only that table (i.e., Table 16) is reported.

Multivariate analysis of variance (MANOVA) was used to calculate the mean differences and statistical significance of differences among groups (Tabachnick & Fidell, 2013). A MANOVA was conducted with all four item types as dependent variables and with race and gender as independent variables to see whether the SVT functions in the same way across demographic groups. Assumptions for this statistical procedure were addressed in the results section.

Institutional Review Board Approval

Permission for this study was requested from the Institutional Review Board (IRB) at Louisiana State University. The description for the project was located in Appendix A. Approval

from Louisiana State University was located in Appendices B and C. Parent consent to allow students to take the assessments was located in Appendix D, with youth consent form to participate included in Appendix E.

CHAPTER 4: RESULTS

The purpose of this study was to conduct both stage 1 performance and stage 2 progress analyses of SVT as part of a larger determination of the efficacy of SVT as a general outcome measure of science content comprehension. The stage 1 portion of the study examined the technical adequacy of five monthly SVT measures to determine the internal consistency reliability and how well the SVT measures predicted success on a standardized measure of science content, the SAT-10 abbreviated online test. The stage 2 portion of the study focused on determining the degree to which SVT measures predicted student growth over time in the area of science comprehension. Due to the unique design of SVT, patterns of achievement were also investigated by item type (i.e., originals, paraphrases, meaning changes and distractors), by race and gender.

This study included students in a large, rural school district in the southeastern U.S. Mean scores, standard deviations, and 95% CIs were examined for the monthly SVT probes by grade (fourth, fifth, and sixth) as well as by gender, race/ethnicity, education classification, and socioeconomic status (SES). Findings related to each of four major research questions follow.

Research Question 1: What were the distributions of scores for the SVT across the five months and for the SAT-10 online for each of the demographic groups?

Examination of Tables 4-6 shows the monthly distribution of scores for all groups over the five months of testing. SVT scores were strongest for February for all grades. For both the fourth and fifth grades, students identified as receiving special education services performed below that of their peers in all monthly assessments except for January (fifth grade). The SAT was administered at the end of March. By examining the group means, it can be seen that males outperformed females in all grades on the SAT-10, but that for the SVT, females outperformed males on every measure in all grades but February (fourth grade) and January and February (fifth

Table 4. Descriptive Statistics for 4th Grade Science Content, $N = 50$

	Male	Female	AA	White	Latino	High SES	Low SES	Gen Ed	Sp Ed	Total
September										
$\bar{X}$	16.5	17.6	17.1	16.5	19	17.7	17.1	17.4	15	17.1
SD	2.9	3.6	3.4	2.1	-	.58	3.5	3.4	2.4	3.3
N	17	25	39	2	1	3	39	38	4	42
95% CI	16.6, 18.0	16.1, 19.1	16.0, 18.2	-2.5,35.6	-	16.2, 19.1	16, 18.2	16.3, 18.5	11.1, 18.9	16.1, 18.2
October										
$\bar{X}$	19.7	21.2	20.6	15	28	16	20.8	21	15	20.5
SD	3.8	2.8	3.0	-	-	1.4	3.3	3.1	-	3.4
N	18	19	34	2	1	2	35	34	3	37
95% CI	17.8, 21.6	19.9, 22.6	19.6, 21.7	-	-	3.3, 28.7	19.6, 21.9	19.9, 22.1	-	19.4, 21.6
November										
$\bar{X}$	17.3	17.8	17.7	14	23	16.7	17.7	18.6	16.5	17.6
SD	3.3	3.1	3.0	4.2	-	4.9	3.1	3.8	4.1	3.2
N	18	23	38	2	1	3	38	37	4	41
95% CI	15.7, 19.0	16.5, 19.2	16.7, 18.6	-24, 52	-	4.4, 28.9	16.7, 18.7	17.4, 19.7	9.9, 23.1	16.6, 18.6
January										
$\bar{X}$	18.4	22.1	18.5	16	19	17.3	18.5	18.6	17	18.4
SD	2.6	4.3	3.7	2.8	-	4.1	3.7	3.8	2.2	3.6
N	20	29	46	2	1	3	46	44	5	49
95% CI	17.2, 19.6	20.5, 23.8	17.4, 19.6	-9.4, 41	-	7.3, 27.3	17.4, 19.6	17.4, 19.7	14.2, 19.8	17.4, 19.5
February										
$\bar{X}$	22.3	22.1	22.2	21	26	21	22.3	22.5	18	22.2
SD	4.0	4.3	4.3	1.4	-	1	4.3	4.0	3.7	4.2
N	18	30	45	2	1	3	45	44	4	48
95% CI	20.3, 24.3	20.5, 22.5	20.9, 23.4	8.3, 33.7	-	18.5, 23.5	21, 23.6	21.3, 23.8	12.2, 23.8	21, 23.4
SAT-10										
$\bar{X}$	616.1	614.7	612.8	650	660	641.7	613.5	616.5	603.8	615.2
SD	37.2	21.3	26.7	43.8	-	35.2	27.5	25.7	49.4	28.4
N	20	30	47	2	1	3	47	45	5	50
95% CI	599, 633	607, 623	605, 621	256,1044	-	554, 729	605, 622	609, 624	542, 665	607, 623

Note. AA = African American; $\bar{X}$ = Mean; CI = Confidence Interval; SD = Standard Deviation; Gen Ed = general education; Sp Ed = special education; SES = socioeconomic status.

Table 5. Descriptive Statistics for 5th Grade Science Content, N = 40

	Male	Female	AA	White	Latino	High SES	Low SES	Gen Ed	Sp Ed	Total
September										
$\bar{X}$	18.3	20.0	18.7	20.5	-	-	18.9	19	18.7	19
SD	3.7	2.3	3.4	.71	-	-	3.3	3.4	2.9	3.3
N	23	15	36	2	-	-	37	35	3	38
95% CI	16.7, 19.9	18.7, 21.3	17.7, 20.0	14.1, 26.8	-	-	17.8, 20	17.8, 20.1	11.5, 25.8	17.8, 20
October										
$\bar{X}$	17.8	19.5	18.6	15	-	-	18.4	18.6	17	18.5
SD	2.8	4.8	3.8	-	-	-	3.8	3.9	2.8	3.8
N	21	15	35	1	-	-	35	34	2	36
95% CI	16.5, 19.1	16.9	17.7, 20	-	-	-	17.1, 19.7	17.2, 20	-8.4, 42	17.2, 19.8
November										
$\bar{X}$	18.2	20.4	19.2	17.5	-	-	19.1	19.1	15.5	19.1
SD	2.4	4.2	3.4	3.5	-	-	3.4	3.5	1.5	3.4
N	24	16	38	2	-	-	40	37	3	40
95% CI	17.2, 19.2	18.2, 22.5	18.1, 20.3	-14.3, 49	-	-	18, 20.2	17.9, 20.2	15.5, 23.1	18, 20.1
January										
$\bar{X}$	17.1	16.8	16.7	18.5	-	-	17	16.9	17.7	17
SD	3.0	2.6	2.8	2.1	-	-	2.8	2.9	5.3	2.8
N	24	16	38	2	-	-	39	37	3	40
95% CI	15.9, 18.4	15.4, 18.2	16, 17.8	-.6, 37.5	-	-	16, 17.9	15.9, 17.9	11.9, 23.4	16.1, 17.9
February										
$\bar{X}$	22.3	21.8	22.1	22.0	-	-	22	22.2	20.7	22.1
SD	5.9	4.3	5.4	2.0	-	-	5.3	5.2	6.7	5.3
N	23	15	36	2	-	-	37	35	3	38
95% CI	19.7, 24.9	19.4, 24.2	20.2, 24	9.3, 34.7	-	-	20, 23.7	20.4, 24.1	4.2, 37.2	20.3, 23.9
SAT-10										
$\bar{X}$	627.4	621.6	623.3	662.5	-	-	625.2	626.7	607	625.2
SD	23.7	31.6	25.6	26.2	-	-	26.7	26.8	20.0	26.7
N	25	15	38	2	-	-	40	37	3	40
95% CI	618, 637	604, 639	615, 632	427, 898	-	-	617, 634	618, 636	557, 657	617, 634

Note. AA = African American; $\bar{X}$ = Mean; CI = Confidence Interval; SD = Standard Deviation; Gen Ed = general education; Sp Ed = special education; SES = socioeconomic status.

Table 6. Descriptive Statistics for 6th Grade Science Content, N = 40

	Male	Female	AA	White	Latino	High SES	Low SES	Gen Ed	Sp Ed	Total
September										
$\bar{X}$	18.8	20.4	19.7	23	-	-	20	19.8	-	20
SD	3.2	3.4	3.4	-	-	-	3.4	3.4	-	3.4
N	13	23	35	1	-	-	35	36	-	36
95% CI	16.9, 20.7	18.9, 21.8	18.5, 20.8	-	-	-	18.7, 21.1	18.6, 20.9	-	18.6, 20.9
October										
$\bar{X}$	21.2	21.6	21.2	29	-	-	21.5	21.4	-	21.4
SD	3.8	5.2	4.5	-	-	-	4.7	4.7	-	4.7
N	12	20	31	1	-	-	31	32	-	32
95% CI	18.7, 23.6	19.1, 24.0	19.5, 22.8	-	-	-	20, 23.2	19.7, 23.1	-	19.7, 23.1
November										
$\bar{X}$	19.8	20.4	20.5	19	10	-	20.4	20.1	-	20.1
SD	3.8	5.6	4.8	-	-	-	4.9	4.9	-	4.9
N	13	22	33	1	1	-	34	35	-	35
95% CI	17.5, 22.1	17.9, 22.8	18.8, 22.2	-	-	-	18.7, 22.1	18.5, 21.8	-	18.5, 21.8
January										
$\bar{X}$	14.5	19.8	17.4	23	25	-	18.1	17.8	-	17.8
SD	4.1	4.9	5.2	-	-	-	5.1	5.3	-	5.3
N	13	22	33	1	1	-	34	35	-	35
95% CI	12.0, 17.0	17.6, 22.0	15.6, 19.3	-	-	-	16.2, 19.8	16, 19.6	-	16, 19.6
February										
$\bar{X}$	18.9	24.6	22.2	26	-	-	22.4	22.3	-	22.3
SD	4.6	4.4	5.3	-	-	-	5.3	5.3	-	5.3
N	14	21	34	1	-	-	34	35	-	35
95% CI	16.2, 21.5	22.6, 26.6	20.4, 24.1	-	-	-	20.6, 24.3	20.5, 24.1	-	20.5, 24.1
SAT-10										
$\bar{X}$	644.4	640.2	638.7	703	693	-	641.7	641.7	-	641.7
SD	28.7	21.1	20.2	-	-	-	23.7	23.7	-	23.7
N	14	26	38	1	1	-	40	40	-	40
95% CI	628, 661	631, 649	632, 645	-	-	-	634, 649	634, 649	-	634, 649

Note. AA = African American; $\bar{X}$ = Mean; CI = Confidence Interval; SD = Standard Deviation; Gen Ed = general education; Sp Ed = special education; SES = socioeconomic status.

grade). Due to the small and unequal group sizes, examination of group score distributions were only evaluated qualitatively, not statistically.

Research Question 2: What was the internal consistency reliability of SVT measures?

To address stage 1 concerns regarding reliability, the SVT was examined using Cronbach's alpha. Cronbach's alpha measures a test's internal consistency, namely, whether a set of test items as a whole are related. Cronbach's alpha detects error in measurement due to the content of the measurement and variability in the measurement items (Reynolds, Livingston, & Willson, 2009). It can be applied to items that are either scored dichotomously, like the SVT, or that have multiple values (multiple-choice). To determine the reliability of the measures by examining the consistency of the responses of all the individual items of the test, Cronbach's alpha coefficients were calculated for each month's assessment, see Table 7. Because a characteristic of GOM is to have short, easily administered probes, Cronbach's alpha was also calculated for each month by passage, see Table 8.

Table 7. Cronbach's Alpha Coefficient by Month

Month	Cronbach's Alpha	*N*
September	.58	931
October	.71	921
November	.77	898
January	.79	943
February	.89	902

Table 8. Cronbach's Alpha Coefficient by Passage by Month

Month	Passage One	Passage Two	*N*
	Cronbach's Alpha		
September	.35	.53	931
October	.53	.60	921
November	.60	.65	898
January	.67	.69	943
February	.84	.79	902

Research Question 3: What were the predictive validity of SVT score(s) for performance on the SAT-10 online? What was the strength of the relationship between SVT and SAT-10 online?

The model for the fourth grade was as follows.

$$Y_{ijk} = \alpha + (\text{Gender})_i + (\text{Teacher})_j + \beta(\text{SVTmonth})_k + (\text{Gender})*(\text{Teacher})_{ij} + \varepsilon_{k(ij)}$$

Data were analyzed with random effects for teacher using multilevel modeling. However, the model failed to converge. Tabachnick and Fidell (2013) stated that convergence problems in linear mixed modeling are common. This is due to the fact that this procedure uses maximum likelihood (when the data are nested), or restricted maximum likelihood, both of which require iterations, and often these fail to converge. This may be due to inaccurately specifying the model, or because the sample is small. The solution Tabachnick and Fidell suggested was to change the random predictors to fixed predictors. For this analysis, the model was then analyzed as an analysis of covariance (ANCOVA), with the SVT score as the covariate and teacher and gender as fixed effects.

For the fifth (N = 40) and sixth grades (N = 40), there was only one teacher per grade per school. Due to this, teacher and school effects at the fifth and sixth grades were confounded with each other. The effects of school and teacher therefore could not be separated. Because the students came from one teacher per school, and the group was largely homogenous, only main effects for school/teacher, gender, and SVT score, as well as the respective interactions were examined. The model was specified as:

$$Y_{ijk} = \alpha + (\text{Gender})_i + (\text{School/Teacher})_j + \beta(\text{SVTmonth})_k + (\text{Gender})*(\text{School/Teacher})_{ij} + \varepsilon_{k(ij)}$$

Again, data were analyzed with random effects for school using multilevel modeling, with the model failing to converge. The model was then analyzed as an ANCOVA, with the SVT score as the covariate and school and gender as fixed effects.

Due to the fact that there were five measures that were measured over time, and the fact that the information of interest was the predictive validity of each of the probes, each month was analyzed separately by grade. To adjust for variance represented by the two different teachers (fourth grade) and by the schools/teachers (fifth and sixth grade), these were entered into the model as fixed factors, with SVT as the covariate. Usually, the covariate is a variance measure that is not of interest; rather the interest is the differences in groups. However, for the present study, the interest is whether the covariate is a significant predictor, after having controlled for group differences.

Before conducting the ANCOVA, the homogeneity-of-regression (slope) assumption was evaluated for all months. The test evaluated the interaction between the covariate and the factor (independent variable) in the prediction of the dependent variable. A significant interaction between the covariate and the factor(s) would suggest that the differences on the dependent variable among groups vary as a function of the covariate, and therefore ANCOVA was not a meaningful procedure (Tabachnick & Fidel, 2013). For all grades in which the monthly SVT measure was a significant predictor, this assumption was met, as well as homogeneity of variance as measured by Levene's test. For fourth grade, the covariate (SVT score) was not significant for any of the months. For fifth grade, September was the only month that SVT was a significant predictor of SAT-10 online, $F(1, 31) = 7.19$, $p < .05$, $R^2 = .464$, $Adj.\ R^2 = .395$. Standardized residuals and Cook's D values were examined for each month for which SVT was significant and all indicated no outliers or influential points. Additionally, the interactions of

SVT and gender*school/teacher were not significant for all months. For sixth grade, tests of the covariate indicated that the following months, SVT was a significant predictor of achievement on SAT-10:

- September, $F(1, 28) = 5.17$, $p < .05$, $R^2 = .264$, *Adj.* $R^2 = .081$
- October, $F(1, 27) = 11.14$, $p < .05$, $R^2 = .333$, *Adj.* $R^2 = .235$
- January, $F(1, 30) = 4.54$, $p < .05$, $R^2 = .142$, *Adj.* $R^2 = .028$
- February, $F(1, 31) = 7.67$, $p < .05$, $R^2 = .209$, *Adj.* $R^2 = .107$.

Parameter estimates for each grade and month are presented in Table 9. Evaluating the magnitude of effect sizes (partial η^2) showed that the strongest predictors for sixth grade achievement on the SAT-10 were October and February.

Table 9. Parameter Estimates by Grade and Month for SVT Predicting Achievement on SAT-10

	B	SE	*t*	P value	95% CI	Partial η^2
Fourth Grade						
	--	--	--	--	--	--
Fifth Grade						
Intercept	554.4	22.0	25.2	.000	[510, 599]	.95
September SVT	3.4	1.2	2.7	.012	[.82, 6.1]	.19
Sixth Grade						
Intercept	592.5	51.4	11.5	.000	[487, 698]	.83
September SVT	2.3	1.1	2.1	.04	[.08, 4.6]	.13
Intercept	601.6	16.5	36.4	.000	[568, 635]	.98
October SVT	2.4	.72	3.3	.002	[.92, 3.9]	.29
Intercept	614.6	15.3	39.6	.000	[583, 646]	.98
January SVT	1.8	.87	2.1	.04	[.07, 3.6]	.13
Intercept	603.2	16.7	35.8	.000	[569, 637]	.98
February SVT	2.2	.78	2.8	.009	[.57, 3.7]	.20

To establish criterion validity of the SVT in relation to SAT-10, the strength of the relationship between each monthly SVT score and the SAT-10 online was examined. Partial

correlations were calculated, controlling for the effects of teacher and gender for the fourth grade and controlling for the effects of school and gender in fifth and sixth grades. Results are reported in Tables 10-12 as well as results with a Bonferroni adjustment to control for family-wise error rate of .05.

To evaluate the criterion validity of the SVT monthly measures in relation to the SAT-10, fourth grade showed only one marginally significant partial correlation, that of October at r = .33, (95% CI; .01, .59); however, with the Bonferroni adjustment, there were no significant correlations. For fifth grade, the September SVT probe showed a significant linear relationship with the SAT-10 with the Bonferroni correction (r = .47). Finally, in sixth grade, four probes (all but November) were significantly correlated with the criterion measure, and with the Bonferroni correction, only the months of October (r = .53) and February (r = .42) were significantly correlated with SAT-10.

Research Question 4: Was there evidence for growth in probe scores during the study and what was the expected growth rate? Were there growth rate differences among student subgroups?

Graphically inspecting the data (see Figure 1) shows that the variance is fairly consistent over time for the repeated measures, with the exception of February (Time 5) for which the mean was slightly higher than the other months. The profile plot (see Figure 2) of individual scores over the five months of assessment shows an inconsistent pattern of high and low scores for individual participants.

Multilevel modeling was used to determine the significance of the mean growth rate of participants, the variability in growth rate among students, and the difference in growth rate between students by subgroup

Table 10. Predictor Means and Partial Correlations Controlling for Teacher and Gender with the Standardized Science Test in Fourth Grade

								SAT-10 Correlations	
			Means			Distributions		Bonferroni	
	N	Range	Mean	SD	95% CI	Skew	Kurt	r	r	95% CI
September	42	7-25	17.1	3.3	[16.1, 18.2]	-1.0	1.8	.03	.03	[-.28, .33]
October	37	15-28	20.5	3.4	[19.4, 21.6]	.27	-.67	.33*	.33	[.01, .59]
November	41	10-23	17.6	3.2	[16.6, 18.6]	-.40	-.15	-.06	-.06	[-.36, .25]
January	49	8-29	18.4	3.7	[17.3, 19.4]	.27	1.4	-.15	-.15	[-.41, .14]
February	48	10-29	22.2	4.2	[21.0, 23.4]	-.57	.33	.20	.20	[-.09, .46]
SAT-10	50	551-681	615.2	28.4	[607, 623]	.37	.29	--	--	--

Note. All correlations non-significant but October, which showed marginal significance at $p = .049$. SAT-10 = Stanford Achievement Test-Tenth Edition, abbreviated form; Shapiro-Wilk statistic indicates SAT-10 and all months but September normally distributed.

Table 11. Predictor Means and Partial Correlations Controlling for School and Gender with the Standardized Science Test in Fifth Grade

			Means			Distributions		SAT-10 Correlations Bonferroni		
	N	Range	Mean	SD	95% CI	Skew	Kurt	r	r	95% CI
September	36	11-28	19.0	3.4	[17.9, 20.1]	-.28	1.1	.47*	.47^	[.17, .69]
October	34	14-31	18.5	3.9	[17.2, 19.9]	1.1	1.8	.09	.09	[-.26, .41]
November	39	14-32	19.2	3.4	[18.1, 20.3]	1.4	3.9	-.01	-.01	[-.32, .31]
January	38	11-26	17.0	2.9	[16.0, 17.9]	.27	1.8	-.10	-.10	[-.41, .23]
February	36	12-31	22.3	5.3	[20.5, 24.1]	-.63	-.36	.07	.07	[-.27, .39]
SAT-10	40	551-681	625.2	26.7	[617, 633]	-.25	.39	--	--	--

Note. All correlations non-significant but September, which was significance at p = .005; ^ significant with Bonferroni adjustment; SAT-10 = Stanford Achievement Test-Tenth Edition, abbreviated form; Shapiro-Wilk statistic indicates SAT-10 and all months but October, November, and February were normally distributed.

Table 12. Predictor Means and Partial Correlations Controlling for School and Gender with the Standardized Science Test in Sixth Grade

	Means					Distribuitons		SAT-10 Correlations	Bonferroni	
	N	Range	Mean	SD	95% CI	Skew	Kurt	r	r	95% CI
September	36	11-27	19.8	3.4	[18.6, 20.9]	-.40	.18	.35*	.35	[.24, .61]
October	32	10-29	21.4	4.7	[19.7, 23.1]	-.57	-.02	.53**	.53^	[.22, .74]
November	35	8-31	20.1	4.9	[18.5, 21.8]	-.06	.92	.06	.06	[-.28, .39]
January	35	9-31	17.8	5.3	[16.1, 19.6]	.51	-.10	.39**	.39	[.07, .64]
February	36	12-32	22.1	5.4	[20.2, 23.9]	-.08	-.69	.42**	.42^	[.11, .66]
SAT-10	40	594-703	641.7	23.7	[634, 649]	.32	.29	--	--	--

Note. All correlations marked * significant at $p < .05$, ** at $p < .005$; ^ significant with Bonferroni adjustment. SAT-10 = Stanford Achievement Test-Tenth Edition, abbreviated form; Shapiro-Wilk statistic indicates all months and SAT-10 were normally distributed.

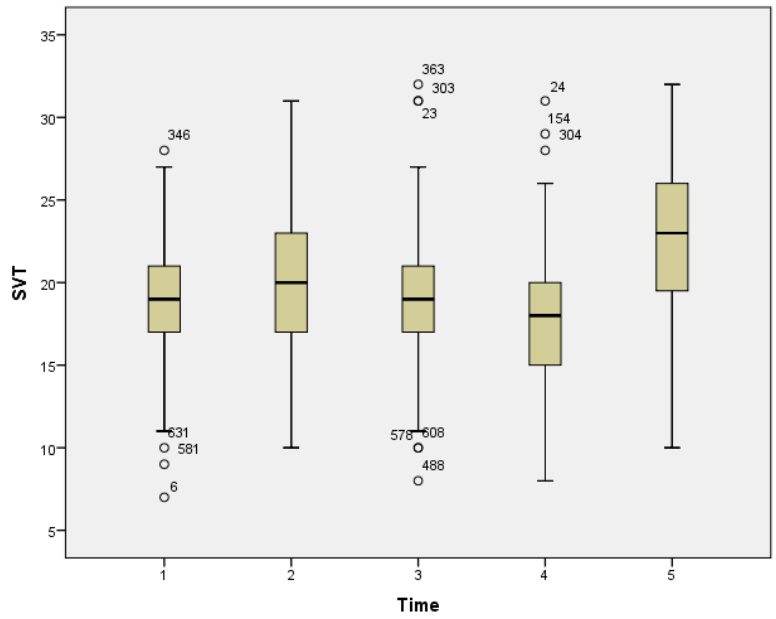

Figure 1. Distribution of Variability of Scores Over the Five Months of the Study

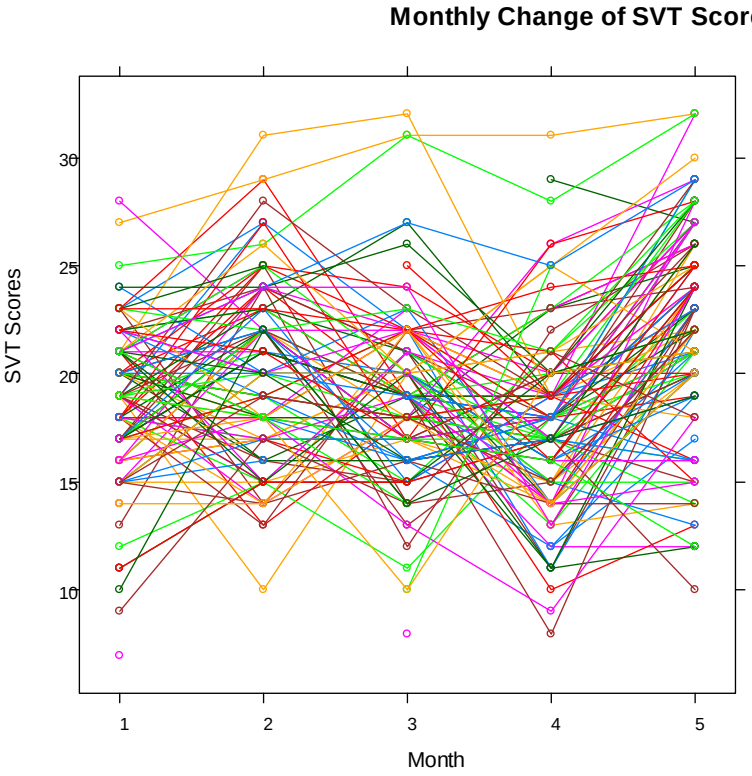

Figure 2. Profile Plot of Individual SVT Scores Across Months, 1 = September, 2 = October, 3 = November, 4 = January, 5 = February

Analysis was evaluated for each grade separately and was completed using maximum likelihood. The model building process started with an unconditional or "null" model, in which the intercept was allowed to vary randomly and there were no predictor variables. The intraclass correlation was then calculated to determine if there was enough variability in individuals to warrant this type of analysis. Once this was established, time was entered into the model as a fixed factor to establish the mean growth rate. Akaike Information Criteria (AIC) was used to evaluate the model at each step. As the magnitude of this information criteria decreased, it indicates a better fit for the model, and therefore can be used to guide the model building process (Bell, Ene, Smiley, & Schoeneberger, 2013). The intercept was centered at the September assessment, the initial point of data collection, in order to establish if there were individual differences in students' SVT scores at the beginning of the study. Results are presented in Tables 13-15.

Table 13. Fourth Grade Growth as Measured by SVT

Fixed effects	Model 1	Model 2	Model 3	Final Model
Intercept	19.2* (.32)	17.3* (.49)	17.4* (.43)	15.9* (.59)
Time	--	.88* (.17)	.84* (.19)	.88* (1.7)
Gender	--	--	--	-.06 (.55)
Teacher	--	--	--	2.7* (.54)
Random effects				
Error Variance				
Level 1	14.8 (1.6)	12.6* (1.4)	12.7* (1.4)	12.6*(1.4)
Intercept	1.5* (1.1)	2.3* (1.1)	--	.59 (.77)
Time			.28 (.14)	--
Model Fit				
AIC (Akaike Information Criteria)	1224.5	1202.4	1203.5^	1186.2

Note. *Statistically significant, $p < .05$; Intraclass correlation = .1; Entries show parameter estimates and standard errors in parenthesis. Estimation method = maximum likelihood. ^AIC value increased, indicating worse model.

For fourth grade, the null model (Model 1) revealed that fixed effects for the intercept were statistically significant with a magnitude of 19.2. This was mathematically equivalent to the grand mean of SVT scores at the initial point of data collection (September). The significant intercept for the error variance indicated that the students differed in SVT scores in September. The intraclass correlation indicated that 10% of the variability in SVT scores existed among students. To determine if SVT was an indicator of average growth for students, the predictor time was added as a fixed effect to the model (Model 2). The result .88 was statistically significant and indicated that for every month, students showed an improvement in SVT scores by .88 points. To determine if SVT growth varied across students, time was added as a random effect. Note that the AIC increased, indicating that the previous model, with fixed effects for time, was the better model. Interactions were found to be non-significant, but only the final model was shown. Students in Teacher A's group scored 2.4 points higher than Teacher B.

Table 14. Fifth Grade Growth as Measured by SVT

Fixed effects	Model 1	Model 2	Model 3	Final Model
Intercept	19.2* (.34)	18.1* (.55)	18.6* (.51)	16.8* (.63)
Time	--	.51* (.20)	.49* (.22)	.51* (.21)
Gender	--	--	--	1.3* (.63)
School	--	--	--	2.1* (.61)
Random effects				
Error Variance				
Level 1	15.7* (1.9)	15.1* (1.8)	14.7* (1.8)	15.1* (1.8)
Intercept	1.7 (1.2)	1.9 (1.2)	--	.45 (.92)
Time			.30 (.19)	--
Model Fit				
AIC (Akaike Information Criteria)	1045.4	1041.5	1040.6^	1033

Note. *Statistically significant, $p < .05$; Intraclass correlation = .097; Entries show parameter estimates and standard errors in parenthesis. Estimation method = maximum likelihood.
^AIC value increased, indicating worse model.

For fifth grade, the null model (Model 1) revealed that fixed effects for the intercept was statistically significant with a magnitude of 19.2, very similar to that of fourth grade, with only the standard error slightly different. This was mathematically equivalent to the grand mean of SVT scores at the initial point of data collection (September). The non-significant intercept for the error variance indicated that the students did not differ in SVT scores in September. The intraclass correlation indicated that 9.7% of the variability in SVT scores existed among students. To determine if SVT was an indicator of average growth for students, the predictor time was added as a fixed effect to the model (Model 2). The result .51 was statistically significant and indicated that for every month, students showed an improvement in SVT scores by .51 points. To determine if SVT growth varied across students, time was added as a random effect. Note that previously the AIC indicated improvement in model selection, but here it increased, indicating that the previous model, with fixed effects for time, was the better model. Finally, to determine if there were significant differences in gender and teacher, interactions were examined and found to be non-significant, but only the final model was shown. Results indicated that there were significant differences for gender, that females performed 1.3 points higher than males, and significant differences in school, with students in School A scoring 2.1 points higher than those in School B.

For sixth grade, results were a bit more complicated. The null model (Model 1) revealed that fixed effects for the intercept was statistically significant with a magnitude of 20. This is mathematically equivalent to the grand mean of SVT scores at the initial point of data collection (September). The significant intercept for the error variance indicated that the students differed in SVT scores in September. The intraclass correlation indicated that 43% of the variability in SVT scores existed among students. To determine if SVT was an indicator of average growth

for students, the predictor time was added as a fixed effect to the model (Model 2). The time variable was non-significant, indicating that the slope was not different from zero, namely, SVT did not change over time. To determine if SVT growth varied across students, time was added as a random effect; however, there was an increase in AIC. To determine if there were significant differences in gender and school, interactions were examined and found to be significant. Results shown in the final model indicated that there were significant differences for the interaction of time and school, indicating that as time progressed, results were not the same depending on what school the student attended.

Table 15. Sixth Grade Growth as Measured by SVT

Fixed effects	Model 1	Model 2	Model 3	Final Model
Intercept	20.0* (.59)	19.6* (.72)	19.9* (.53)	19.2* (.81)
Time	--	.16 (.20)	.09 (.27)	.49* (.22)
School	--	--	--	1.8 (1.6)
Time*School	--	--	--	-1.4* (.45)
Random effects				
Error Variance				
Level 1	13.9* (1.7)	13.8* (1.7)	14.5* (1.8)	13.0* (1.6)
Intercept	10.7* (3.2)	10.8* (3.2)	--	10.5* (3.1)
Time			1.4* (.43)	--
Model Fit				
AIC (Akaike Information Criteria)	1015.8	1017.1^	1024.6^	1011.8

Note. *Statistically significant, $p < .05$; Intraclass correlation = .43; Entries show parameter estimates and standard errors in parenthesis. Estimation method = maximum likelihood. ^AIC value increased, indicating worse model.

Research Question 5: Were there differences in item response patterns for the different item types in SVT based on race or gender?

In order to examine the differences in performance on item type by different demographic groups, the full data set was employed for analysis. Due to the fact that September

was researcher-created and it had 9 yes items (5 paraphrases, 4 originals) and 7 no items (4 meaning changes and 3 distractors) and it had lower reliability, these analyses only included the months of October, November, January, and February. A larger body of students took monthly assessments than took the criterion measure, and due to this and the fact that in the larger sample there was more diversity, the larger sample was examined but was limited to those students for whom there was complete data for the four months (N = 567). Demographic summary for the sample can be found in the methods section in Table 2. Analysis was limited to comparisons between gender and race/ethnicity, namely, African American students and White students.

The items scores were totaled in SPSS by month. In each SVT there were eight originals and eight paraphrases (both correct answer "yes"), eight meaning changes and eight distractors (both correct answer "no"). Scores were coded "1" if the student got the item correct and "0" if the student got the item incorrect. The total number of correct items by type were calculated for each student for each month, the maximum a student could obtain on any item type for each month was eight. Then, means for the overall group were calculated. Bar graphs were generated by item type for the overall sample by month, see Figure 3.

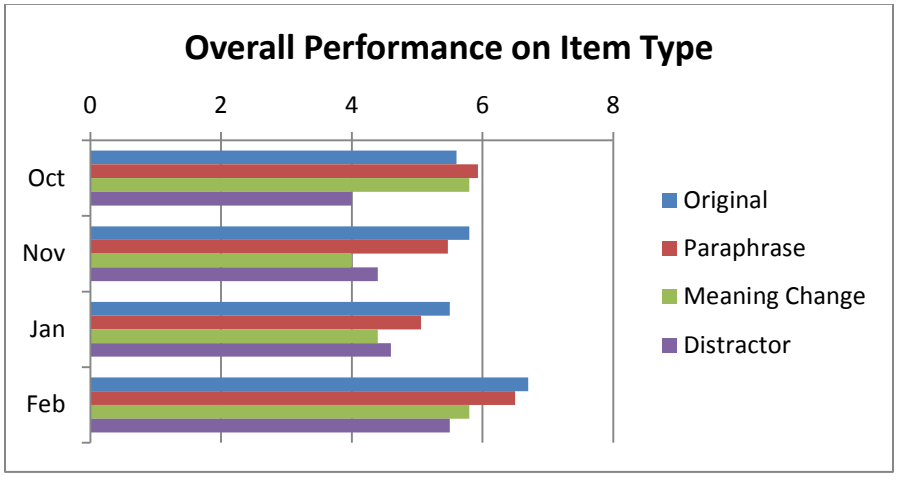

Figure 3. Scores by Item Type Over the Course of the Study

For ease of comparison, the items were condensed over the course of the study into overall original, paraphrase, meaning change, and distractor scores with the total correct of 32 (8 per item x 4 months). In Figure 4, the students as a whole were most successful on originals, followed by paraphrases, meaning changes and then distractors. Referring back to Figure 3, February appeared to be the month where students were the most successful and the only month that tracked the same as the composite scores.

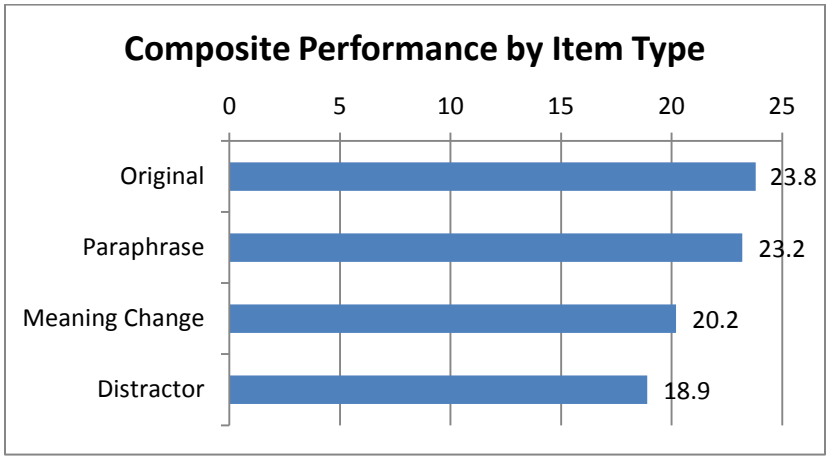

Figure 4. Overall Scores by Item Type (Total Number Possible Correct = 32)

To examine relationships between females and males, the data were analyzed by gender (see Figure 5). Results indicated that females performed consistently stronger than males on all item types, with a trend following the overall trend of best performance on originals, paraphrases, meaning changes, and distractors in that order.

When examined across race (see Figure 6), there appeared to be parity between African American and White students on originals and paraphrases, the "yes" items. On the "no" items, White students achieved at a higher rate than African American students with the greatest discrepancy for distractors. When examined across race for female students (see Figure 7), the pattern of success on each item mirrored that of the overall African American group (see Figure

6), with the exception that African American females achieved slightly higher on originals than White females.

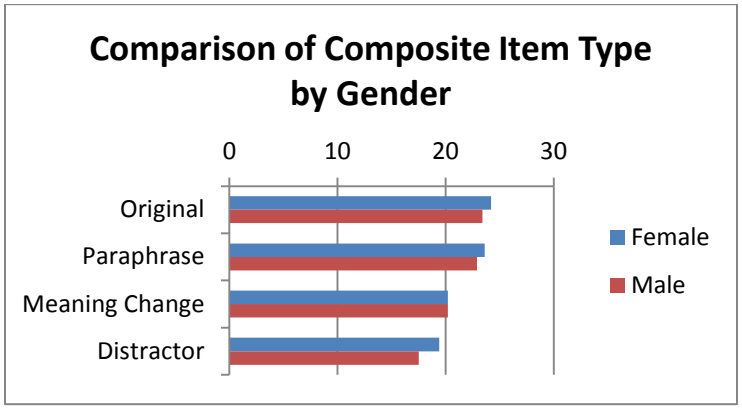

Figure 5. Item Achievement Analyzed by Gender

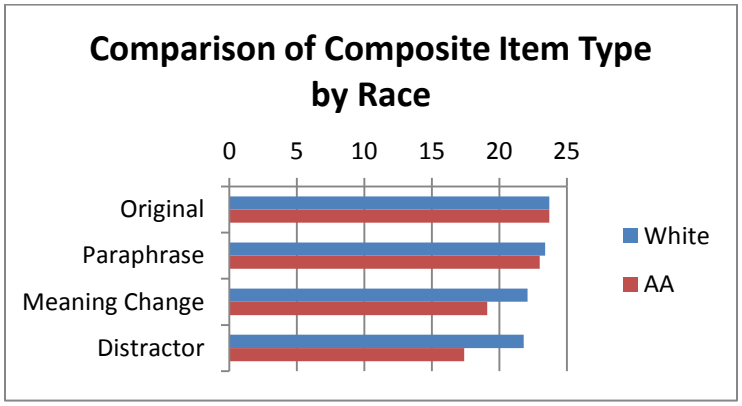

Figure 6. Item Achievement Analyzed by Race.

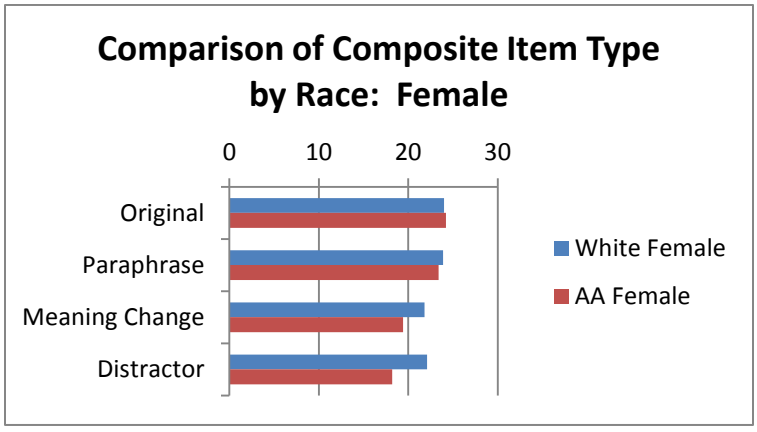

Figure 7. Item Achievement Analyzed by Race for Females

When examined across race for male students (see Figure 8), the pattern of success on each item mirrored the overall achievement by race (see Figure 6).

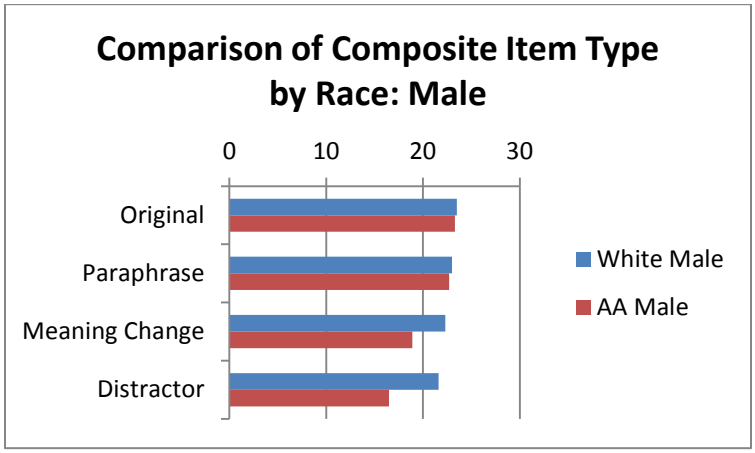

Figure 8. Item Achievement Analyzed by Race for Males

The trends in item achievement for African American males and females in Figure 9 mirrored that of overall males and females (see Figure 5).

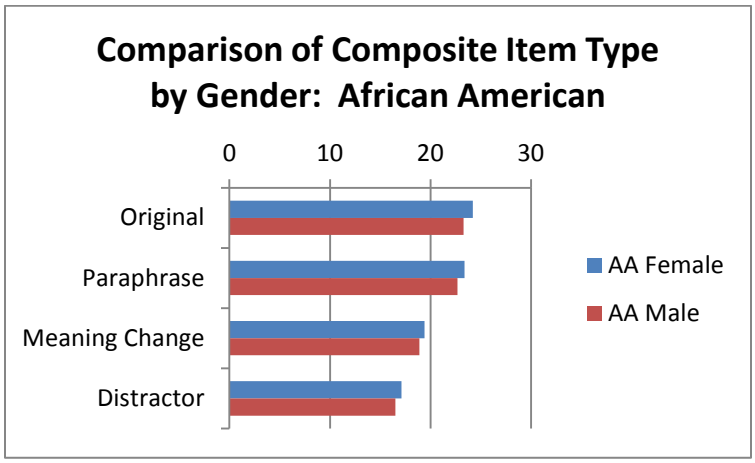

Figure 9. Item Achievement Analyzed by Gender for African American Students.

Finally, in examining Figure 10, the item achievement tracked that of the overall achievement as seen in Figure 4, with the exception that for White students, females greatly outperformed males on distractors, a trend not seen in the overall comparison by gender (see Figure 5) or by gender for African American students (see Figure 9).

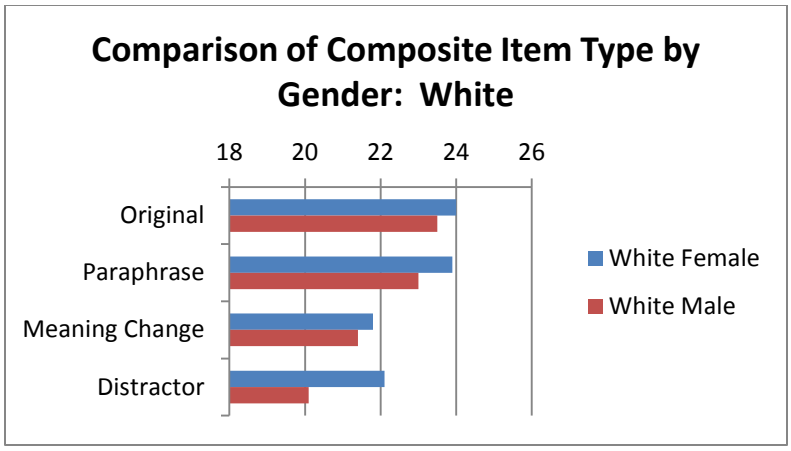

Figure 10. Item Achievement Analyzed by Gender for White Students.

In order to examine the relationships between item types, Pearson correlation coefficients were calculated among the items (see Table 16). Overall, the relationship was strong between the "yes" items, originals and paraphrases (r = .71), with the relationship, while still statistically significant, was much smaller for "no" items, meaning changes and distractors (r = .40).

Table 16. Correlations Among Composite Items for All Students

	Original	Paraphrase	Meaning Change
Paraphrase	.71**		
Meaning Change	.33**	.38**	
Distractor	.13**	0.05	.40**

Note. ** = Significant at $p < .001$

Correlations were analyzed for all subgroups, and will be summarized. Interestingly, in the large group, the relationship between originals (yes item) and distractors (no item) was highly significant at $p < .001$ level, but when disaggregated by race, that relationship was highly significant at the $p < .001$ level for African Americans overall (r = .18) and not significant for the White group. When looking at the White group only, that relationship was significant at the $p < .05$ level for the females (r = .15). When looking at the African American group only, that relationship was highly significant for the females (r = .23).

To better understand the patterns of achievement for the group by race and gender, a two-way MANOVA was conducted with the originals, paraphrases, meaning changes, and distractors as the dependent variables and the two categorical predictors, gender and race, as independent variables. Only subjects with complete data for all item types for all months were analyzed, (N = 567), and due to small sample sizes for races other than White and African American, only those two categories were compared as well as gender. Data were screened for normality and all of the dependent variables were found to be non-normal as measured by the Shapiro-Wilks statistic. Examining histograms revealed that all dependent variables appeared left-skewed, however, MANOVA is robust to violations of normality (Tabachnick & Fidell, 2013). Box's test revealed equality of covariance matrices and the Levene's test revealed equal variances for all dependent variables except paraphrases. The main effects for gender were marginally significant (Wilks' Λ= .98, $F(4, 545) = 2.5$, $p < .042$, partial $\eta^2 = .02$). The main effects for race/ethnicity were highly significant (Wilks' Λ= .88, $F(4, 545) = 19.5$, $p < .001$, partial $\eta^2 = .13$). The interaction between race and gender, however, was not found to be statistically significant. Univariate ANOVAs were used to isolate the source of the difference. Results indicated that the greatest differences occurred for gender on originals, $F(1, 548) = 4.33$, p = .04 and paraphrases, $F(1, 548) = 4.8$, p = .03 with girls outperforming boys, and for race/ethnicity on meaning changes, $F(1, 548) = 48.1$, p <.001 and distractors, $F(1, 548) = 46.7$, p < .001 with White students outperforming African American students; however the size of the effect, as measured by partial η^2 was small (< .01 for all variables).

CHAPTER 5: DISCUSSION

The purpose of this study was to determine Stage 1 and 2 (Fuchs, 2004) technical adequacy characteristics for SVT. The research was designed to ascertain the efficacy of using SVT as a general outcome measure of science content knowledge. The research questions addressed in this study included demographic comparisons in the measures by month; internal consistency reliability of SVT measures; criterion and predictive validity of the SVT on SAT-10 online; evidence for student growth in probe scores; and differences in item response patterns for SVT items based on race or gender.

Summary of Findings

Demographic Comparisons by Month

The first research question sought to describe the personal characteristics of the sample and the distribution of mean scores on the monthly SVT probes. Means, standard deviations, and 95% CIs were calculated and compared across groups. Due to the homogeneity of the sample, only very small subgroups were present, and therefore were not able to be examined statistically. Qualitative inspection of differences indicated that students in all three grades performed the highest on the February measures. Males outperformed females for all grades on the criterion measure, but females outperformed males in almost all months of the SVT measures. This may be due, in part, because the SAT-10 online was a multiple choice test of science knowledge, where the SVT is a language-based assessment where students are intended to read and comprehend science information in order to select whether test items retained the meaning of the original passage. Females have been shown to outperform males in tests of reading and language (Watson, Kehler, & Martino, 2010; Corbett, Hill, & St Rose, 2008).

Internal Consistency Reliability of SVT measures

The second research question sought to determine the internal consistency reliability of the SVT. Most GOM studies report alternate-forms reliability or test-retest reliability. All reliability estimates for SVT thus far have been measured by utilizing Cronbach's alpha or Spearman Brown formula (Royer, 2004). Results indicated an increasing trend in the reliability as measured by Cronbach's alpha. The lowest score, September, also happened to be the one where the readability level based on the Flesch-Kincaid scale was the highest. It was also the only probe developed solely by the researcher, without any assistance or input from the highly qualified science teacher, and taken directly from a 5^{th} grade text. Subsequent passages were also taken from 4^{th} – 6^{th} grade texts, but the verbiage in the passages was changed by the science teacher to lower the reading level. To date, this is the only study that has reported internal consistency reliability for a two-passage SVT. Internal consistency reliability is preferred when a test is designed to be implemented once, for tests that are to be given more than once to the same individuals, test-retest or alternate-forms reliability is advised (Reynolds et al., 2009).

In the present study, the goal was to determine if students were able to read and comprehend science content from month to month, with each month presenting new content utilizing the curriculum sampling method of GOM probe construction. The reliability estimates ranged from $\alpha = .58$ to .89, with an average of .75, prompting the question what level is acceptable? Reynolds et al. (2009) stated that when high-stakes decisions are being made as a result of the test score, that reliability estimates should be greater than .9 and that for teacher made tests and tests for screening purposes, that the estimates should be at least .7. In relation to this criterion, the SVT met the screening standard for all months but September. Because the goal of GOM is to provide teachers with a quick measure of student progress, the monthly

assessments were investigated by passage (16 item assessments). These alpha coefficients ranged from .35 (September) to .84 (February) with 7 of the 10 passages .6 or higher. These coefficients indicate that there is promise in reliability for shorter SVTs. It is possible that the test items could be increased to 20 and thereby possibly increase the reliability. The fact that the estimates were highest for February did, however, show promise in that the creators of the assessments were improving their skills in test development.

Predictive and Criterion Validity

The third research question set out to determine if SVT significantly predicted achievement on the SAT-10. Results indicated that SVT failed to predict achievement for any monthly probes for the fourth grade. Fifth grade produced one significant predictor, that of September, (R^2 = .464, *Adj.* R^2 = .395), indicating that the September probe shared 40% of the variability in the criterion measure. The sixth grade produced the most probes that were significant predictors of achievement on the SAT-10. The months of September, October, January and February were found to be significant with effect sizes ranging from R^2 = .142, *Adj.* R^2 = .028 (January) to R^2 = .333, *Adj.* R^2 = .235 (October). These values were comparable to predictive validity estimates of SVT found in Marcotte and Hintze (2009) and with the combination of critical content monitoring and SVT (13% shared variability with SAT) in Mooney et al. (2015).

To evaluate the criterion validity of the SVT monthly measures in relation to the SAT-10, partial correlations were examined for all grades. Results for the fourth grade showed one significant linear relationships with the criterion measure for the month of October at r = .33, (95% CI; .01, .59); however, with the Bonferroni adjustment, this probe was not significant. For the fifth grade, only one significant partial correlation was found, that of September, r = .47,

(95% CI; .17, .69). Sixth grade partial correlations were the strongest, with four of the five months statistically significant, listed in order of increasing magnitude: September, r = .35, (95% CI; .24, .61), January, r = .39, (95% CI; .07, .64), February, r = .42, (95% CI; .11, .66), and October, r = .53, (95% CI; .22, .74). These criterion coefficients were similar in magnitude to those reported for vocabulary matching which had mean scores that were moderately correlated with multiple criterion measures including standardized subject matter and general knowledge and statewide accountability tests (Espin et al., 2001; Mooney, McCarter, Schraven, et al., 2013). Mooney, McCarter, Russo, et al. (2013) had a pooled estimate of .45 for CCM with a state accountability test, and vocabulary matching tests were higher for social studies (.7) than for science (.46 - .47) (Mooney et al., 2015).

Cohen (1988) established criteria for evaluating the magnitude of correlation coefficients. According to Cohen, coefficients of .10 are considered "small," those of .30 are "medium," and those of .50 are "large" (see pp. 77–81). Following this very well-established guideline would put the October correlation for fourth grade in the medium range, however, more recent investigation in psychological measurements have been examined. Hemphill (2003) evaluated two large summaries of the literature regarding psychological assessment (Meyer, Finn, Eyde, Kay, Moreland, Dies, 2001; Lipsey & Wilson, 1993) totaling 380 metaanalytic reviews. Hemphill asserted that researchers often judge the magnitude of correlations based on other guidelines, such as perfect correlation (something hardly ever achieved in applied research), reliability coefficients which he states often are larger than validity coefficients, and "…monomethod correlation coefficients, which yield results that are artificially large compared with associations found between real-world, independently measured variables" (p. 78). From the studies, he compiled correlations based on converted measures of effect size reported as

Cohen's *d* into Pearson product-moment correlations and established empirical guidelines based on the results of this. He stated that a correlation of .5 would correspond to the 89th percentile for the Meyer et al. (2001) studies and 97th percentile for the Lipsey et al. (1993) studies, implying that Cohen's "high" benchmark may be unrealistic. He compiled all results into a table with empirical guidelines of <.20 in the lower third, .20 to .30 in the middle third, and >.30 in the upper third.

Following these guidelines, the sixth grade correlation of SVT with SAT-10 would be in the upper third in magnitude. Fifth grade showed only September statistically significant r = .47, (95% CI; .17, .69). This is very close to Cohen's (1988) cutoff for "high" measure of effect size, and in the upper third of Hemphill's criteria. These range from medium to high effect sizes according to Cohen's criteria and are all in the upper third according to Hemphill's criteria.

Measures of Growth

The fourth research question set out to determine the significance of the mean growth rate of participants, the variability in growth rate among students, and the difference in growth rate between students by teacher or school and by gender. The analyses were performed by grade level. Results indicated that fourth grade showed significant mean growth rate over time (.88 items per month) and fifth grade showed (.51 items per month). Sixth grade had a significant interaction of time and school, and the mean growth rate was not significant. The fourth and fifth grade growth estimates were similar to those found for vocabulary matching and reported in Borsuk (2010), and Espin et al. (2013). The variability in growth scores across grade was also similar to stage 2 findings in the content GOM literature.

Item Response Patterns by Race and Gender

Finally, to answer the fifth research question, SVT items (originals, paraphrases, meaning changes, and distractors) were examined across race and gender for the entire group of students for whom there were complete data ($N = 567$). Items were investigated across the five months of the study by comparing mean scores on each item type by race and gender. Due to small samples of students identifying as other than African American or White, only African American and White students were compared.

Results from a MANOVA indicated that when compared by gender, female students performed at a higher rate on all item types. When disaggregated by race, African American and White students performed roughly equally on originals and paraphrases (the "yes" items) and White students performed statistically significantly better than African American students on meaning changes and distractors (the "no" items). When comparing White females to African American females, both groups performed roughly equally on paraphrases, with African American females slightly outperforming White females on originals and White females performing better on meaning changes and distractors. White males outperformed African American males on all item types. African American females outperformed African American males on all items, and White females outperformed White males on all item types.

What does different performance on item type mean? Recall, that the theoretical assumption upon which SVT was developed is that comprehension is a constructive process (Royer et al., 1979). When a reader comprehends what has been read, he or she makes a memory representation developed from the incoming linguistic message and his or her own prior knowledge. If this memory representation is consistent with the text, it can then be assessed by

presenting the reader with items that seek to determine if the student can recognize if the idea that was present in the original passage is being presented in a test item.

Originals and paraphrases are measures of the comprehension of language, demonstrating whether a student can recognize ideas that are present in a text. Because originals are sentences that are exact copies of those in the passage, being able to recognize these and answer correctly "yes" may be simply a measure of basic comprehension of language, where performing successfully on paraphrases may be measuring a more complex level of language comprehension (Carlisle & Felbinger, 1991). To correctly identify paraphrases, students need to not only recognize ideas that were present in the text but also decipher that meaning from a different set of words or grammatical structures from that of the original passage. Carlisle and Felbinger posited that patterns of errors on meaning changes and distractors potentially indicated a strategy developed by students to compensate for meaning construction difficulties. If students are making errors on meaning changes it is possible that students are not paying careful attention to the wording and meaning of the sentence because these test items only differ on one or two words from that of the passage. Carlisle and Felbinger further posited that distractors potentially measure a student's difficulty in identifying ideas within the passage, because these items "test whether a student has developed a sense of the ideational boundaries of the text (p. 347)." They suggested that readers with a significant number of errors on meaning changes or distractors may indicate an inattention to the exact wording in the passage for meaning changes or a dependence on background knowledge in the case of distractors. Their study, however, focused on students identified as poor readers or good readers. In the present study, significant differences were found not based on reading ability, but on racial group and marginally significant differences in

gender. This raises the question, are the types of items presented in the SVT valid for all racial/ethnic groups and gender?

The *Standards for Educational and Psychological Testing* (American Educational Research Association, American Psychological Association, & National Council on Measurement in Education, 1999) delineate four ways to address test fairness, or the "moral, philosophical, or legal issue on which reasonable people can disagree" (Brown, Reynolds, & Whitaker, 1999 as cited in Reynolds et al., 2009; p. 429). These ways of defining fairness are absence of bias, equitable treatment, opportunity to learn, and equal outcomes. In the present study, the male and female students as well as African American and White students came from the entire district, comprised of seven schools. It is entirely possible that these students did not have equitable treatment and equal opportunities to learn in each of these seven schools. Messick (1989) addressed the concept of validity in respect to the social consequences of testing. If the SVT were being used to make high-stakes decisions such as whether to promote students to the next grade then the consequences of the different patterns of responses to SVT items based on gender and race would be more dire than consequences emanating from screening and progress monitoring processes. In light of the present study, that SVT is being used to guide instruction, the consequences are not negligible, however. If males and African Americans continue to underperform in relation to females and White students, it is possible that they may referred to Tier 2 (small group) instruction in Responsiveness-to-Intervention frameworks more often than needed, not as a result of student performance but as a result of bias in the instrument.

Mestre and Royer (1991) addressed the concept of test fairness in the realm of language-minority students. They stated that a test which may be designed to be sensitive both culturally and linguistically for English language learners (ELL) may fail to be valid for that group. They

cited research that was current at that time which reported that tests that seemed to significantly predict future achievement for native English speakers may not be as efficient for predicting performance for ELL students (Hedges & Majer, 1976; Houston, 1980; Mestre, 1981) and advised caution in how these test scores were interpreted among language minority populations.

Willingham and Cole (1997) stated that there exists research that investigates patterns of gender similarities and differences across race/ethnicity, but they claimed that there is little research that looked at the patterns of gender similarities and differences within racial/ethnic groups. They looked at gender differences within ethnic groups for three different types of tests: undergraduate admissions tests, advanced course placement tests and graduate admissions tests. They investigated ACT and SAT results for 1992 and found very little variation in gender differences across race/ethnicities but for African American examinees, the patterns of difference were markedly different than the White group. They reported that females showed a stronger mean difference than males in the African American group than they found in the White group. For advanced placement tests, more African American women took the exams than African American men, but their performance was similar to that of the ACT/SAT performance despite outnumbering the men by almost 2 to 1. Finally in the 1992 GRE results, African American women outperformed African American men at a rate higher than White women outperformed White men.

For the present study, these findings were supported. African American females outperformed African American males overall, but were outperformed by White females on the "no" items. Based on the results found here, caution about the fairness and validity of the SVT for language minority students should be extended to address racial and ethnic minority populations as well as gender.

Implications

Students are increasingly being expected to develop reading comprehension skills as they progress through the grades, with the focus on academic vocabulary from the content areas as specified in the CCSS. Students are supposed to read carefully and understand information based on evidence in the text, then to use information gained from reading to answer questions requiring inferences based on careful attention to informational text, therefore building students' content knowledge (NGA & CCSSO, 2010). In light of this, students who face reading comprehension challenges can be expected to also face content comprehension challenges.

General outcome measurement has a long history of effectively identifying and monitoring students who are not performing at expected levels. As responsiveness to intervention (RTI) becomes prevalent in schools as the metric to identify and monitor students failing to show progress, effective forms of GOMs are needed, specifically for students who are in upper grades and content area courses. These GOMs are used to benchmark test the entire student population and are predominately used in the lower grades to identify students who are struggling with reading, writing, and/or math. As students progress through the upper grades and enter middle school, they are presented with a host of challenges such as a new school, changing classes with different teachers for content classes, and more stringent curricular demands (Johnson & Smith, 2008). It follows that some students might face learning challenges with these increased demands. Furthermore, as students progress through the grades, academic deficits compound and become more severe the farther they fall behind (L.S. Fuchs, Fuchs & Compton, 2010). Students in middle and high school may exhibit reading challenges from word recognition to metacognitive skills. "Shortfalls in any of these areas have been implicated as a significant contributor to comprehension failure appreciably decreasing students' ability to use

text to acquire new vocabulary, information, and knowledge" (Fuchs et al., p. 25-26). Given the increasing demands placed on students regarding comprehension, integration, and interpretation of content as a prerequisite to college and career readiness, there needs to be a measure, or suite of measures, that can be used to assess older students' understanding of science concepts, as presented in written form. To date, although vocabulary matching and more recently its online version critical content monitoring have been the most investigated, SVT shows promise for this purpose as well.

Stage 1 and 2 Validity Evidence

In evaluating the reliability and criterion validity of the monthly SVT probes, results from the present study indicated that the measures were, especially when broken out by passage, had a reliability estimates that were in the moderate range (Reynolds et al., 2009). This is promising in that the probes have the potential to be presented as one passage with 16 items, further increasing its utility as a short, easily-administered assessment. The fact that the probes can be administered and scored via computer further increases its utility, especially for teachers in the upper grades who serve far more students per day than do elementary school teachers. Royer, Carlo, and Cisero (1992) posited that SVT could be used to track students' progress because the SVT has demonstrated that students who are better readers perform better on SVT as well as older students perform better than younger students and that their performance has been shown to improve as the year progresses.

Sentence Verification Technique in the present study has also been shown to have predictive and criterion validity for the fifth and sixth grades. This is a promising first step to validating measures for use as a general outcome measure. Finally, SVT was shown to be a significant measure of student growth for fourth and fifth grades, with sixth grade growth

contingent on school. Research has shown that growth from the beginning of the year to middle of the year has been documented, with less growth from the middle to the end of the year (Ardoin & Christ, 2008; Graney, Missall, Martinez & Bergstrom, 2009; Tindal, 2013). It is possible that not having administered SVT in December may have affected this outcome.

SVT Viability as a General Outcome Measure

SVT shows promise as an addition to the suite of GOM for content courses, vocabulary matching and critical content monitoring. It has long been established as a measure of reading comprehension and has a solid theoretical framework to support its implementation. While the results reported here show that SVT is reliable and has some predictive and criterion validity, as well as a measure of student growth, caution should be maintained in evaluating its effectiveness. The results of this sample may not be generalizable to a larger population. It is conceivable that there will not be a robust indicator discovered to measure content knowledge. Vocabulary matching shows promise, but SVT may work best in a multiple measures approach to GOM to formatively assess and guide instruction.

SVT may lend face validity to formative assessment of content comprehension in that it requires students to read text-based content and determine whether test items directly relate to content or not. This type of assessment possibly reflects student expectation and learning action better than do some of the other content-focused general outcome measures. The fact that students are not allowed to return to the passage to answer questions, while part of the theory behind SVT that a memory representation is, in fact, a product of comprehension does not reflect the nature of testing today. Students are encouraged to return to the text to look for evidence, and therefore future studies could examine allowing students to do this in SVT.

Finally, SVT scores have the potential for instructional utility, in that the probes are easily created by teachers and come from content found either in textbooks or ancillary content materials, and are easily administered and tracked online. In this way, they have the possibility to directly inform instruction. Because students read content passages and then construct different types of comprehension representations, there is the possibility for performance patterns in SVT testing to provide teachers with ideas as to how to proceed instructionally. Such a possibility enables teachers a chance to positively impact academic performance in reading and content comprehension areas.

Limitations of the Study

There are several limitations to the study. The study was conducted in a predominantly rural school district in south Louisiana. The target population of the study was fourth, fifth, and sixth grade science students and their respective teachers as part of a large grant intended to improve science education. The accessible population were students at two schools in the district. Because the focus this school year will be on science, it may affect the generalizability of the study when these methods are used with students whose focus of instruction is not on improving science education district-wide.

Another limitation is that the demographic of the population is predominately African American and students whose families are considered low SES. The percentage of students in the two schools is 95% African American and 98% low SES. The demographics of this district may not be generalizable to many other districts across the nation.

Finally, the school year 2014-2015 was initially intended to be the inaugural year that the state would implement the Partnership for Assessment of Readiness for College and Careers (PARRC) test, a test that has been developed in relation to the CCSS. Because of the potential

problems that accompany a new assessment, and the current legislative battle happening in Louisiana regarding Common Core, it was decided to use a commercially available standardized test with which to establish the criterion validity of the SVT. Because of the costs involved in having students complete standardized assessments, the sample size was intended to be 50 per grade level. The actual sample sizes were 50 for fourth grade and 40 for fifth and sixth grades respectively. While this number is sufficient for parametric statistical analysis, it is limiting in terms of generalizability of the study.

Recommendations for Future Research

Future research should be conducted with a larger and more diverse sample, possibly with more frequent assessments to get a better picture of the ability of SVT to model student growth. Furthermore, future research should address other content courses such as history and social studies. SVT could also be evaluated with secondary measures of reading comprehension as well, increasing its potential as an English/Language Arts GOM. Additionally, future research should include Stage 3 research, perhaps qualitative in nature, to determine what teachers see as benefits and limitations to this type of assessment. This may serve to guide future experimental research to determine if the use of SVT as a general outcome measure successfully affects teachers' use of GOM to formatively guide their content instruction.

REFERENCES

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education (1999). *Standards for educational and psychological testing*. Washington, DC: American Educational Research Association.

Ardoin, S. P., & Christ, T. J. (2008). Evaluating curriculum-based measurement slope estimates using data from triannual universal screenings, *School Psychology Review*. 37, 109–125.

Bell, B. A., Ene, M., Smiley, W., & Schoeneberger, J. A. (2013). A multilevel model primer using SAS® PROC MIXED. *SAS Global Forum,* (0-19).

Bidwell, A. (2014, August 20). Common Core in Flux as States Debate Standards. *US News and World Report.* Retrieved from http://www.usnews.com/news/articles/2014/07/15/common-core-status-in-flux-as-states-debate-standards-tests.

Borsuk, E. R. (2010). Examination of an administrator-read vocabulary-matching measure as an indicator of science achievement. *Assessment for Effective Intervention, 35*(3), 168-177.

Brown, R. T., Reynolds, C. R., & Whitaker, J. S. (1999). Bias in mental testing since Bias in Mental Testing. *School Psychology Quarterly*, *14*(3), 208.

Burns, M. K., Scholin, S. E., & Zaslofsky, A. F. (2011). Advances in Assessment Through Research: What Have We Learned in the Past 3 Years? *Assessment for Effective Intervention, 36*(2), 107-112.

Busch, T. W., & Espin, C. A. (2003). Using curriculum-based measurement to prevent failure and assess learning in the content areas. *Assessment for Effective Intervention, 28*(3-4), 49-58.

Carlisle, J. F. (1989a). Diagnosing Comprehension Deficits Through Listening and Reading. *The Annals of Dyslexia*, 39, 159-176.

Carlisle, J. F. (1989b) The Use of the Sentence Verification Technique in Diagnostic Assessment of Listening and Reading Comprehension. *Learning Disabilities Research, 5*(1), 33-44.

Carlisle, J. F., & Felbinger, L. (1991). Profiles of listening and reading comprehension. *The Journal of Educational Research*, *84*(6), 345-354.

Carlo, M. S., Sinatra, G. M., & Royer, J. M. (1989). *Using the Sentence Verification Technique to measure transfer of comprehension skills from native to second language*. Paper presented at the annual meeting of the American Education Research Association, San Francisco.

Carney, R. N. (n.d.). Review of the Stanford Achievement Test, Tenth Edition. *Mental Measurements Yearbook.* http://buros.org/mental-measurements-yearbook.

Christ, T. J., White, M. J., Ardoin, S. P., Eckert, T. L., & VanDerHeyden, A. (2013). Curriculum Based Measurement of Reading: Consistency and Validity Across Best, Fastest, and Question Reading Conditions. *School Psychology Review, 42*(4).

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Hillsdale, NJ: Erlbaum.

Common Core State Standards: Key shifts in Language Arts (2014), retrieved from: http://www.corestandards.org/other-resources/key-shifts-in-english-language-arts/.

Corbett, C., Hill, C., & St Rose, A. (2008). *Where the girls are: The facts about gender equity in education*. American Association of University Women Educational Foundation. 1111 Sixteenth Street NW, Washington, DC 20036.

Corson, D. (1997). The learning and use of academic English words. *Language Learning, 47*(4), 671-718.

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

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

Deno, S. L. (2003). Developments in curriculum-based measurement. *The Journal of Special Education, 37*(3), 184-192.

Espin, C. A., Busch, T. W., Lembke, E. S., Hampton, D. D., Seo, K., & Zukowski, B. A. (2013). Curriculum-based measurement in science learning vocabulary-matching as an indicator of performance and progress. *Assessment for Effective Intervention, 38*(4), 203-213.

Espin, C. A., Busch, T., Shin, J., & Kruschwitz, R. (2001). Curriculum-based measures in the content areas: Validity of vocabulary-matching measures as indicators of performance in social studies. *Learning Disabilities Research & Practice, 16,* 142–151.

Espin, C. A., & Deno, S. L. (1993a). Content-Specific and General Reading Disabilities of Secondary-Level Students Identification and Educational Relevance. *The Journal of Special Education, 27*(3), 321-337.

Espin, C. A., & Deno, S. L. (1993b). Performance in reading from content area text as an indicator of achievement. *Remedial and Special Education, 14*(6), 47-59.

Espin, C. A., & Deno, S. L. (1994-1995). Curriculum-based measures for secondary students: Utility and task specificity of text-based reading and vocabulary measures for predicting performance on content-area tasks. *Diagnostique, 20,* 121-142.

Espin, C. A., & Foegen, A. (1996). Validity of three general outcome measures for predicting secondary students' performance on content-area tasks. *Exceptional Children, 62*(6), 497-514.

Espin, C. A., Shin, J., & Busch, T. W. (2005). Curriculum-based measurement in the content areas: Vocabulary matching as an indicator of progress in social studies learning. *Journal of Learning Disabilities, 38,* 353–363.

Espin, C., Wallace, T., Lembke, E., Campbell, H., & Long, J. D. (2010). Creating a Progress-Monitoring System in Reading for Middle-School Students: Tracking Progress Toward Meeting High-Stakes Standards. *Learning Disabilities Research & Practice*, *25*(2), 60-75.

Fewster, S., & Macmillan, P. D. (2002). School-based evidence for the validity of curriculum-based measurement of reading and writing. *Remedial and Special Education, 23*(3), 149-156.

Fuchs, D., & Fuchs, L. S. (2006). Introduction to response to intervention: What, why, and how valid is it?. *Reading Research Quarterly*, *41*(1), 93-99.

Fuchs, D., Mock, D., Morgan, P. L., & Young, C. L. (2003). Responsiveness-to-intervention: Definitions, evidence, and implications for the learning disabilities construct. *Learning Disabilities Research & Practice*, *18*(3), 157-171.

Fuchs, L. S. (2004). The past, present, and future of curriculum-based measurement research. *School Psychology Review 33*(2), 188-192.

Fuchs, L. S., & Deno, S. L. (1991). Paradigmatic distinctions between instructionally relevant measurement models. *Exceptional Children, 57*(6), 488-99.

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

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

Fuchs, L. S., Fuchs, D., & Compton, D. L. (2010). Rethinking response to intervention at middle and high school. *School Psychology Review, 39*(1), 22-28.

Fuchs, L. S., Fuchs, D., Hosp, M. K., & Jenkins, 1. R. (2001). Oral Reading Fluency as an Indicator of Reading Competence: A Theoretical, Empirical, and Historical Analysis. *Scientific Studies of Reading*, 5, 239-256.

Fuchs, L. S., Fuchs, D., & Zumeta, R. O. (2008). A curricular-sampling approach to progress monitoring: Mathematics concepts and applications. *Assessment for Effective Intervention*.

Fuchs, L. S., & Vaughn, S. (2012). Responsiveness-to-Intervention: A decade later. *Journal of Learning Disabilities*, *45*, 195-203. doi: 10.1177/0022219412442150.

Good, R. H., Simmons, D. C., & Kame'enui E. 1. (2001). The importance and decision making utility of a continuum of fluency-based indicators of foundational reading skills for third-grade high-stakes outcomes. *Scientific Studies of Reading*, 5, 257-288.

Greene, B. A., Royer, J. M., & Anzalone, S. (1990). A new technique for measuring listening and reading literacy in developing countries. *International Review of Education, 36*(1), 57-68.

Harper, R. (2014). Development of a Health Literacy Assessment for Young Adult College Students: A Pilot Study. *Journal of American College Health*, *62*(2), 125-134. DOI:10.1080/07448481.2013.865625.

Hedges, L.V., & Majer, K. (1976).An attempt to improve prediction of college success of minority students by adjusting for high school characteristics. *Educational and Psychological Measurement*, 36, 953-957.

Hemphill, J. F. (2003). Interpreting the magnitudes of correlation coefficients. *American Psychologist*, 58(1), 78-79.

Hintze, J. M., Christ, T. J., & Methe, S. A. (2006). Curriculum-based assessment. *Psychology in the Schools*, *43*(1), 45-56.

Hoover, H. D., Hieronymus, A. N., Frisbie, D. A., & Dunbar, S. B. (1992). *Iowa test of basic skills*. Chicago, IL: Riverside Publishing Company.

Hosp, J. (2011). Using assessment data to make decisions about teaching and learning. In K. Harris, S. Graham, &T. Urdan (Eds.), *APA Educational psychology handbook.* (Vol. 3, pp. 87-110). Washington, DC: American Psychological Association.

Houston, L.N. (1980).Predicting academic achievement among specially-admitted Black female college students. *Educational and Psychological Measurement, 40,* 1189-1195.

Howe, K. B., Scierka, B. J., Gibbons, K. A., & Silberglitt, B. (2003). A school-wide organization system for raising reading achievement using general outcome measures and evidence-

based instruction: One education district's experience. *Assessment for Effective Intervention, 28,* 59-72.

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

Jones, M., & Smith, M. (2014). Traditional and alternative methods of measuring the understandability of accounting narratives. *Accounting, Auditing & Accountability Journal*, *27*(1), 183-208.

Ketterlin-Geller, L. R., McCoy, J. D., Twyman, T., & Tindal, G. (2006). Using a concept maze to assess student understanding of secondary-level content. *Assessment for Effective Intervention*, *31*(2), 39-50.

Kincaid, J.P., Fishburne, 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. Research Branch Report 8-75*. Chief of Naval Technical Training: Naval Air Station Memphis.

Kintsch, W. (1994). Text comprehension, memory, and learning. *American Psychologist*, *49*(4), 294.

Kintsch, W., & van Dijk, T. (1978). Toward a model of text comprehension and production. *Psychological Review*, 85, 363-394.

Kovalenko, A. (2013). *Disclosure of the Persuasive Intent behind the Placement of Risky Products in Movies: Consequences for Cognitive and Affective Processing* (Doctoral dissertation, University of Otago).

Linn, R L. (2002). Validation of the uses and interpretations of results of state assessment and accountability systems. In Tindal, G. & Haladyna, T. M. (Eds.), *Large-scale assessment programs/or all students* (p.p. 49-66). Mahwah, NJ: Erlbaum.

Lipsey, M. W., & Wilson, D. B. (1993). The efficacy of psychological, educational, and behavioral treatment: Confirmation from meta-analysis. *American Psychologist, 48,* 1181–1209.

Louisiana Department of Education. (LDE; n.d.). Academic and Grade Level Expectations. Available: http://www.louisianabelieves.com/resources/library/academic-standards.

Marchand, G. C., Nardi, N. M., Reynolds, D., & Pamoukov, S. (2014). The impact of the classroom built environment on student perceptions and learning. *Journal of Environmental Psychology, 40,* 187-197. DOI: 10.1016/j.jenvp.2014.06.009.

Marchant, H. G., Royer, J. M., & Greene, B. A. (1988). Superior reliability and validity for a new form of the Sentence Verification Technique for measuring comprehension. *Educational and Psychological Measurement, 48*(3), 827-834.

Marcotte, A. M., & Hintze, J. M. (2009). Incremental and predictive utility of formative assessment methods of reading comprehension. *Journal of school psychology, 47*(5), 315-335.

Messick, S. (1989). Validity. In R. L. Linn (Ed.), *Educational measurement* (3rd ed., pp. 13-103). Upper Saddle River, NJ: Merrill Prentice Hall.

Mestre, J.P. (1981).Predicting academic achievement among bilingual Hispanic college technical students. *Educational and Psychological Measurement, 41*, 1255-1264.

Mestre, J. P., & Royer, J. M. (1991). Cultural and linguistic influences on Latino testing. In G. Keller, J. Deneen, & R.Magallan (Eds.), *Assessment and access: Hispanics in higher education*. Albany, N. Y.: State University of New York Press. (pp. 39-66).

Meyer, G. J., Finn, S. E., Eyde, L. D., Kay, G. G., Moreland, K. L., Dies, R. R., Eisman, E. J., & Kubiszyn, T. W. (2001). Psychological testing and psychological assessment: A review of evidence and issues. *American Psychologist, 56*, 128–165.

Mooney, P., Lastrapes, R. E., Marcotte, A. M. & Matthews, A. (2015). *Validity Evidence for Critical Content Monitoring and Sentence Verification Technique as Indicators of Student Science and Social Studies Achievement.* Manuscript submitted for publication.

Mooney, P., McCarter, K. S., Russo, R. J., & Blackwood, D. L. (2013). Examining an Online Content General Outcome Measure Technical Features of the Static Score. *Assessment for Effective Intervention, 38*(4), 249-260.

Mooney, P., McCarter, K. S., Russo, R. J., & Blackwood, D. L. (2014). The structure of an online assessment of science and social studies content: Testing optional formats of a general outcome measure. *Social Welfare Interdisciplinary Approach, 4*(1).

Mooney, P., McCarter, K. S., Schraven, J., & Callicoatte, S. (2013). Additional Performance and Progress Validity Findings Targeting the Content-Focused Vocabulary Matching. *Exceptional Children, 80*(1), 85-100.

Mooney, P., McCarter, K. S., Schraven, J., & Haydel, B. (2010). The Relationship Between Content Area General Outcome Measurement and Statewide Testing in Sixth-Grade World History. *Assessment for Effective Intervention, 35*(3), 148-158.

Mooney, P., Schraven, J., & Cox, B. (2010). Test-retest reliability of vocabulary matching in sixth-grade world history. *International Journal of Psychology: A Biopsychosocial Approach, 6*, 29-40.

Morse, D. T. (n.d.). Review of the Stanford Achievement Test, Tenth Edition. *Mental Measurements Yearbook*. http://buros.org/mental-measurements-yearbook.

Morrison, G. M. (2013). *Fundamentals of Early Childhood Education* (7th ed.). Boston: Pearson.

Muyskens, P., & Marston, D. B. (2006). The relationship between curriculum-based measurement and outcomes on high-stakes tests with secondary students. *Minneapolis Public Schools. Unpublished manuscript*.

National Center for Education Statistics (2013). *The Nation's Report Card: A First Look: Mathematics and Reading* (NCES 2014-451). Institute of Education Sciences, U.S. Department of Education, Washington, D.C. Retrieved from: http://nationsreportcard.gov/reading_math_2013.

National Governors Association Center for Best Practices & Council of Chief State School Officers. (2010). *Common Core State Standards for English language arts and literacy in history/social studies, science, and technical subjects*. Washington, DC: Authors.

RAND Reading Study Group (2002). *Reading for Understanding. Towards an R&D Program in Reading Comprehension*. Santa Monica, CA: RAND Corporation.

Rasool, J. M., & Royer, J. M. (1986). Assessment of reading comprehension using the Sentence Verification Technique: Evidence from narrative and descriptive texts. *The Journal of Educational Research*. Vol 79(3), Jan-Feb 1986, 180-184.

Reynolds, C. R., Livingston, R. B., & Willson, V. (2009). *Measurement and assessment in education* (2nd ed.). Upper Saddle River, NJ: Pearson Education International.

Royer, J. M. (1990). The Sentence Verification Technique: A new direction in the assessment of reading comprehension. In S. M. Legg & J. Algina (Eds.), *Cognitive Assessment of Language and Math Outcomes* (pp. 144-191). Norwood, NJ: Ablex.

Royer, J. M. (2001). Developing Reading and Listening Comprehension Tests Based on the Sentence Verification Technique (SVT). *Journal of Adolescent & Adult Literacy, 45*(1), 30-41.

Royer, J. M. (2004). Uses for the sentence verification technique for measuring language comprehension. *Progress in Education*. Retrieved online at: http://www.readingsuccesslab.com/publications/Svt%20Review%20PDF%20version.pdf.

Royer, J. M., Abranovic, W. A., & Sinatra, G. (1987). Using entering reading performance as a predictor of course performance in college classes. *Journal of Educational Psychology*, 79, 19-26.

Royer, J. M., & Carlo, M. S. (1991a). Assessing the Language Acquisition Progress of Limited English Proficient Students: Problems and New Alternative. *Applied Measurement in Education*, *4*(2), 85-113.

Royer, J. M., & Carlo, M. S. (1991b). Transfer of comprehension skills from native to second language. *Journal of Reading*, 34, 450-455.

Royer, J. M., Carlo, M. S., Carlisle, J. F., & Furman, G. A. (1991). A new procedure for assessing progress in transitional bilingual education programs. *Bilingual Review*, *16*, 3-14.

Royer, J. M., Carlo, M. S., & Cisero, C. A. (1992). School-based uses for the Sentence Verification Technique for measuring listening and reading comprehension. *Psychological Test Bulletin*, *5*(1), 5-19.

Royer, J. M., & Cunningham, D. J. (1978). On the theory and measurement of reading comprehension. Technical Report No. 91. *Center for the Study of Reading*. U S. Department of Health, Educational Welfare; National Institute of Education.

Royer, J. M., & Cunningham, D. J. (1981). On the theory and measurement of reading comprehension. *Contemporary Educational Psychology*, *6*(3), 187-216.

Royer, J. M., & Greene, B. A. (1990). *The computer-based assessment of cognitive reading skills in Belize: Final Report*. Arlington, VA: Institute for International Research.

Royer, J. M., Greene, B. A., & Anzalone, S. J. (1994). Can U.S. developed CAI work effectively in a developing country? *Journal of Educational Computing Research, 10*, 41-61.

Royer, J. M., Greene, B. A., & Sinatra, G. M. (1987). The Sentence Verification Technique: A practical procedure teachers can use to develop their own reading and listening comprehension tests. *Journal of Reading, 30*, 414-423.

Royer, J. M., Hastings, C. N., & Hook, C. (1979). A sentence verification technique for measuring reading comprehension. *Journal of Literacy Research*, *11*(4), 355-363.

Royer, J. M., Kulhavy, R. W., Lee, J. B., & Peterson, S. E. (1986) The Sentence Verification Technique as a measure of listening and reading comprehension. *Educational & Psychological Research*, 6(4), 299-314.

Royer, J. M., Lynch, D. J., Hambleton, R. K., & Bulgareli, C. (1984). Using the sentence verification technique to assess the comprehension of technical text as a function of subject matter expertise. *American Educational Research Journal*, *21*(4), 839-869.

Royer, J. M., Marchant, H., Sinatra, G., & Lovejoy, D. (1990) The prediction of college course performance from reading comprehension performance: Evidence for general and specific prediction factors. *American Educational Research Journal*, 27, 158-179.

Royer, J. M., & Sinatra, G. M. (1988). *Using the Sentence Verification Technique to Assess Storage and Retrieval Processes*. Massachusetts University: Amherst.

Royer, J. M., & Sinatra, G. M. (1994). A cognitive theoretical approach to reading diagnostics. *Educational Psychology Review*, *6*(2), 81-113.

Royer, J. M., Sinatra, G. M., Greene, B. A., & Tirre, W. C. (1989). Assessment of on-line comprehension of computer-presented text. *The Journal of Educational Research*, 348-355.

Royer, J. M., & Sinatra, G. M, & Schumer, H. (1990). Patterns of individual differences in the development of listening and reading comprehension. *Contemporary Educational Psychology, 15*, 183-l96.

Shapiro, E. S., Keller, M. A., Lutz, J. G., Santoro, L. E., & Hintze, J. M. (2006). Curriculum-Based Measures and Performance on State Assessment and Standardized Tests Reading and Math Performance in Pennsylvania. *Journal of Psychoeducational Assessment*, *24*(1), 19-35.

Silberglitt, B., & Hintze, J. M. (2007). How Much Growth Can We Expect? A Conditional Analysis of R—CBM Growth Rates by Level of Performance. *Exceptional Children*, *74*(1), 71-84.

Stecker, P. M., Fuchs, L. S., & Fuchs, D. (2005). Using Curriculum-Based Measurement to Improve Student Achievement: Review of Research. *Psychology in the Schools*, *42*(8), 795-819.

Swets, J. A. (2014). *Signal detection theory and ROC analysis in psychology and diagnostics: Collected papers*. New York: Psychology Press.

Swets, J. A., Tanner Jr, W. P., & Birdsall, T. G. (1961). Decision processes in perception. *Psychological review*, *68*(5), 301.

Tabachnick, B. G., & Fidell, L. S. (2013). *Using multivariate statistics* (6th ed.). Boston, MA: Pearson Education, Inc.

Tichá, R., Espin, C. A., & Wayman, M. M. (2009). Reading Progress Monitoring for Secondary-School Students: Reliability, Validity, and Sensitivity to Growth of Reading-Aloud and Maze-Selection Measures. *Learning Disabilities Research & Practice*, *24*(3), 132-142.

Tindal, G. A. (2013). Curriculum-based measurement: A brief history of nearly everything from the 1970s to the present. *ISRN Education*.

Tindal, G. A., Fuchs, L. S., Fuchs, D., Shinn, M. R., Deno, S. L., & Germann, G. (1985). Empirical validation of criterion-referenced tests. *The Journal of Educational Research*, 203-209.

Tindal, G. A., & Germann, G. (1991). Mainstream consultation agreements in secondary schools. In G. Stoner, M. R. Shinn, & H. M. Walker (Eds.), *Interventions for achievement and behavior problems* (pp. 495-518). Bethesda, MD: National Association of School Psychologists.

Tindal, G. A., Parker, R., & Germann, G. (1990). An analysis of mainstream consultation outcomes for secondary students identified as learning disabled. *Learning Disability Quarterly, 13*(3), 220-229.

Tsikalas, K. E. (2012). *Effects of video-based peer modeling on the question asking, reading motivation and text comprehension of struggling adolescent readers* (Doctoral dissertation, City University of New York, New York).

Tucker, J. (1987). Curriculum-based assessment is not a fad. *The Collaborative Educator, 1*, 4, 10.

Twyman, T., & Tindal, G. A. (2007). Extending curriculum-based measurement into middle/secondary schools: The technical adequacy of the concept maze. *Journal of Applied School Psychology, 24*(1), 49-67.

Ulusoy, M., & Cetinkaya, C. (2012). The use of sentence verification technique for measuring reading and listening. *Hacettepe Üniversitesi Eğitim Fakültesi Dergisi (H. U. Journal of Education) 43,*460-471.

United States Census Bureau, State and County Quick Facts, retrieved from: http://quickfacts.census.gov/qfd/states/22/22047.html.

Vannest, K. J., Parker, R., & Dyer, N. (2011). Progress monitoring in grade 5 science for low achievers. *The Journal of Special Education, 44*(4), 221-233.

Wallace, T., Espin, C. A., McMaster, K., Deno, S. L., & Foegen, A. (2007). CBM progress monitoring within a standards-based system. *The Journal of Special Education, 41*(2), 66-67.

Watson, A., Kehler, M., & Martino, W. (2010). The problem of boys' literacy underachievement: Raising some questions. *Journal of Adolescent & Adult Literacy, 53*(5), 356-361.

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

APPENDIX A: LOUISIANA STATE UNIVERISITY INSTITUTIONAL REVIEW BOARD PROJECT DESCRIPTION

College of Human Sciences & Education

Research Description for Parents
Reliability and Validity Studies for General Outcome Measures

The present research targets reliability and/or validity questions surrounding online general outcome measures of content knowledge and reading comprehension. A sample of up to 200 public school students from Grades 4 through 6 who assent and whose parents consent to participate in the study will be administered a series of reading and/or content tests. Additionally, traditional demographic (e.g., age, gender, grade, race/ethnicity, socioeconomic status) and school achievement data (e.g., statewide accountability test results) will be collected. Analysis will involve comparisons among the measures.

The following measures may be administered to students as part of the study:

- Sentence Verification Technique (SVT): a set of grade-level reading passages that are accompanied by a series of test sentences that students read to determine whether each is the same or different than a sentence in the story, with the score consisting of the number of sentences marked correctly.
- Maze (MZ): a grade-level passage that has been adapted so that every 7th word is replaced by a multiple-choice selection of 3 choices, with students expected to choose the word that best fits the sentence and scoring consisting of the number of correct selections in 3 minutes.
- Critical Content Monitoring (CCM): An online test that presents multiple-choice questions in which the stem includes a definition and the choices include vocabulary content terms, with students expected to select the correct choice and scoring consisting of the number of correct selections in 5 minutes. Five forms of the test will be administered.
- Stanford Achievement Test (10th edition)(SAT-10) online abbreviated form (Comprehension, Vocabulary, Science, Social Studies): An online standardized achievement test for students in Grades 3 through 12, the SAT-10 is group administered, with the grade-level reading vocabulary and comprehension tests lasting 44 minutes and the science and social studies tests each taking 19 minutes. https://pearsonassessments.com/haiweb/cultures/en-us/productdetail.htm?pid=SAT10Online

Demographic and achievement test data will also be collected from school officials. Achievement data will consist of the student's scores from the 2015 statewide grade-level assessments in English language arts (ELA), reading, science, and social studies.

Louisiana State University
225 Peabody Hall
Baton Rouge, LA 70803

O: 225-578-5667
F: 225 578 9135
www.lsu.edu

APPENDIX B: LOUISIANA STATE UNIVERSITY INSTITUTIONAL REVIEW BOARD APPROVAL FORM AUGUST 2014

ACTION ON EXEMPTION APPROVAL REQUEST

Institutional Review Board
Dr. Dennis Landin, Chair
130 David Boyd Hall
Baton Rouge, LA 70803
P: 225.578.8692
F: 225.578.5983
irb@lsu.edu | lsu.edu/irb

TO: Paul Mooney
Education

FROM: Dennis Landin
Chair, Institutional Review Board

DATE: August 26, 2014

RE: **IRB#** E8269

TITLE: Technical adequacy studies for Critical Content Monitoring

New Protocol/Modification/Continuation: Modification

Brief Modification Description: Adding Renee Lastrapes as Co-PI

Review date: 8/25/2014

Approved X **Disapproved** _____

Approval Date: 8/25/2014 **Approval Expiration Date:** 4/7/2016

Re-review frequency: (three years unless otherwise stated)

LSU Proposal Number (if applicable): 41470

Protocol Matches Scope of Work in Grant proposal: (if applicable) _____

By: Dennis Landin, Chairman _____

PRINCIPAL INVESTIGATOR: PLEASE READ THE FOLLOWING –
Continuing approval is CONDITIONAL on:

1. Adherence to the approved protocol, familiarity with, and adherence to the ethical standards of the Belmont Report, and LSU's Assurance of Compliance with DHHS regulations for the protection of human subjects*
2. Prior approval of a change in protocol, including revision of the consent documents or an increase in the number of subjects over that approved.
3. Obtaining renewed approval (or submittal of a termination report), prior to the approval expiration date, upon request by the IRB office (irrespective of when the project actually begins); notification of project termination.
4. Retention of documentation of informed consent and study records for at least 3 years after the study ends.
5. Continuing attention to the physical and psychological well-being and informed consent of the individual participants including notification of new information that might affect consent.
6. A prompt report to the IRB of any adverse event affecting a participant potentially arising from the study.
7. Notification of the IRB of a serious compliance failure.
8. SPECIAL NOTE:

**All investigators and support staff have access to copies of the Belmont Report, LSU's Assurance with DHHS, DHHS (45 CFR 46) and FDA regulations governing use of human subjects, and other relevant documents in print in this office or on our World Wide Web site at http://www.lsu.edu/irb*

APPENDIX C: LOUISIANA STATE UNIVERSITY INSTITUTIONAL REVIEW BOARD APPROVAL FORM JANUARY 2015

ACTION ON EXEMPTION APPROVAL REQUEST

Institutional Review Board
Dr. Dennis Landin, Chair
130 David Boyd Hall
Baton Rouge, LA 70803
P: 225.578.8692
F: 225.578.5983
irb@lsu.edu | lsu.edu/irb

TO: Paul Mooney
Education

FROM: Dennis Landin
Chair, Institutional Review Board

DATE: January 23, 2015

RE: IRB# E8269

TITLE: Reliability and Validity Studies for General Outcome Measures

New Protocol/Modification/Continuation: Modification

Brief Modification Description: Revised Targeted Teaching Behaviors checklist

Review date: 1/23/2015

Approved X **Disapproved** _____

Approval Date: 1/23/2015 **Approval Expiration Date:** 4/7/2016

Re-review frequency: (three years unless otherwise stated)

LSU Proposal Number (if applicable): 41470

Protocol Matches Scope of Work in Grant proposal: (if applicable) _____

By: Dennis Landin, Chairman _____

PRINCIPAL INVESTIGATOR: PLEASE READ THE FOLLOWING –
Continuing approval is CONDITIONAL on:

1. Adherence to the approved protocol, familiarity with, and adherence to the ethical standards of the Belmont Report, and LSU's Assurance of Compliance with DHHS regulations for the protection of human subjects*
2. Prior approval of a change in protocol, including revision of the consent documents or an increase in the number of subjects over that approved.
3. Obtaining renewed approval (or submittal of a termination report), prior to the approval expiration date, upon request by the IRB office (irrespective of when the project actually begins); notification of project termination.
4. Retention of documentation of informed consent and study records for at least 3 years after the study ends.
5. Continuing attention to the physical and psychological well-being and informed consent of the individual participants including notification of new information that might affect consent.
6. A prompt report to the IRB of any adverse event affecting a participant potentially arising from the study.
7. Notification of the IRB of a serious compliance failure.
8. SPECIAL NOTE:

**All investigators and support staff have access to copies of the Belmont Report, LSU's Assurance with DHHS, DHHS (45 CFR 46) and FDA regulations governing use of human subjects, and other relevant documents in print in this office or on our World Wide Web site at http://www.lsu.edu/irb*

APPENDIX D: PARENT CONSENT FORM TO TAKE THE ASSESSMENT

College of Human Sciences & Education

PARENTAL CONSENT FORM FOR PARTICIPATION
Institutional Review Board (IRB)#_____

Title: Reliability and Validity Studies for General Outcome Measures
Performance Sites: Louisiana public school classrooms in Grades 5 through 8
Contact: Renee Lastrapes, (225) 578-2360, rlastr2@tigers.lsu.edu. Available: 8 a.m.-5 p.m.
Dr. Paul Mooney, (225) 578-2360, pmooney@lsu.edu. Available: 8 a.m.-5 p.m.
IRB Contact Information: This study has been approved by the LSU IRB. For questions about participants' rights, please contact the chair, Dr. Dennis Landin, 578-8692, or irb@lsu.edu.

Purpose: The study examines validity and reliability statistics for scores from a series of measures of reading comprehension, science knowledge, and social studies knowledge.

Participants: Public school students in Grades 4 through 6.

Research Procedures: Participating students will be administered a series of formal and informal achievement tests during school time. Students will either be asked to read brief passages, answer multiple-choice questions, or write brief summaries of material that they read. Testing will take place over multiple days so that no more than one hour per day of instructional time is utilized. Testing is not expected to last more than two hours per student. Following each day of testing, participants will receive candy treats for their best efforts. Researchers will also collect participant demographic and statewide accountability test result data.

Potential Benefits: There are believed to be no immediate benefits to students beyond their receipt of candy treats for their best effort on a given day of testing.

Potential Risks: There are believed to be no apparent risks to your youth.

Right to Refuse: Participation is voluntary. You or your youth can choose not to participate in the study. Also, your youth can quit the study at any time without penalty. You or your youth's relationship with the school, investigators, or LSU will not be damaged in any way if you choose for your youth not to participate in the study or if your youth decides to quit during the study.
Privacy: The confidentiality of your reply will be ensured. Names will only be released to research team members (i.e., investigators). Documents will be maintained in a locked file cabinet when not being gathered. Entered data will not include student names and will remain on the office computer of the primary investigator or statistician.
Financial Information: There will be no financial compensation for participating.
Signature: "I have been fully informed of the above-described procedure, its possible benefits and risks, and I give my permission for my youth to participate in the study."

_____ _____ _____
Parent Signature Youth's Name (Please Print) Date

Louisiana State University
223 Peabody Hall
Baton Rouge, LA 70803

LOVE PURPLE
LIVE GOLD

P: 225-578-6867
F: 225-578-9135
www.lsu.edu

APPENDIX E: YOUTH ASSENT FORM TO TAKE THE ASSESSMENT

College of Human Sciences & Education

YOUTH ASSENT FORM FOR PARTICIPATION
Institutional Review Board (IRB)#_____

Title: Reliability and Validity Studies for General Outcome Measures

I have spoken with my parent(s) and teacher(s) about the study. I know that I will be completing a series of informal and formal tests of reading comprehension and science and social studies knowledge during school hours and that testing should not take more than two hours to complete. I also know that my demographic and statewide accountability test data will be released to the research team. I agree to do all the activities of the study that I have been told about by my teacher(s). I know that I can talk to my parent(s) or teacher(s) if I have concerns about the study activities. I also know that I can quit the study at any time without penalty.

_____ _____ _____
Youth Signature Youth's Name (Please Print) Date

Louisiana State University
223 Peabody Hall
Baton Rouge, LA 70803

P 225-578-6809
F 225-578-7995
www.lsu.edu

APPENDIX F: SVT PASSAGES AND TEST ITEMS

SEPTEMBER

SVT Inquiry Passage 1
Grade level 9.4

Science means, "having knowledge." Science is a way of seeing, studying, and thinking about things in your world to gain knowledge. Many observations cannot be explained easily. When people cannot explain things, they ask questions. Science tries to answer questions and solve problems to better understand the world. Every time you attempt to find out how and why things look and behave the way they do, you are performing science.

The scientific method is a process that people use to investigate and answer questions. The scientific method helps scientists to explain how things happen in the natural world. Scientists do not always follow the steps of the scientific method in order, but they do make sure that they and others can repeat their procedures.

Scientists conduct controlled experiments to determine a cause-and-effect relationship among the factors, called variables, which are changed in the experiment. The variable that is changed in the experiment is called an independent variable. The variable that is being measured is called the dependent variable. Scientists try to keep every other variable constant, or unchanged, in the experiment. A controlled experiment must have two groups, the experimental group where the independent variable is changed and a control group where it is unchanged. Both the experimental and control groups must include the same factors under the same conditions. Once completed, the results of the experiment are analyzed and explained.

Item	Type	
1	P	Scientists use the scientific method to help them better understand things that happen naturally in our world.
2	MC	Science is a way to study the things in your world to make them different.
3	D	Science is used to discover diseases.
4	O	Every time you attempt to find out how and why things look and behave the way they do, you are performing science.
5	MC	Scientists cannot answer questions or solve problems to better understand the world.
6	O	The scientific method is a process that people use to investigate and answer questions.
7	O	Many observations cannot be explained easily.
8	MC	When people cannot explain things, they make up answers.
9	O	A controlled experiment must have two groups, the experimental group where the independent variable is changed and a control group where it is unchanged.
10	MC	Scientists control experiments by making sure that there are no variables or factors that are changed.
11	P	The results of the experiment are analyzed and explained when the experiment is finished.
12	D	Some scientific investigations cost a lot of money to do.
13	P	The same factors and conditions must be used in both the experimental and

		control groups.
14	P	The dependent variable is the one that is measured.
15	D	When you place things with similar properties into groups, you are classifying.
16	P	The independent variable is the one that is changed in the experiment.

SVT Inquiry Passage 2
Grade level 11.8

Different types of scientific questions call for different types of investigations. A scientific investigation is a way of answering a scientific question. Questions that ask about the effects of one factor on another are often tested by performing an experiment. A controlled experiment is a scientific investigation that involves changing one factor and observing its effects on another factor while keeping all other factors constant. Sometimes scientists have questions that cannot be answered with a laboratory experiment.

Scientists often attempt to answer questions that cannot be answered through laboratory exploration by observing the natural world. A field study is an investigation in which scientists make observations and collect information outside of the laboratory. Sometimes making a model is an effective way to answer a scientific question. A model is a representation of an object or an event that is used as a tool for understanding the natural world. Models are often made when the investigation involves elements that are difficult to observe or understand. Models are useful, but they are not exact and often lack detail. If learning about your world begins with asking questions and making observations, can science provide answers to these questions? Science can answer a question only with the information available at the time. Any answer to a scientific question is uncertain because people will never know everything about the world around them. Some observations might force scientists to think of new explanations. Science can only provide possible explanations.

Item	Type	
16	P	A controlled experiment is a type of scientific study where one element or characteristic is changed and the effects it has on other factors is monitored while all other elements are kept the same.
17	D	Scientific investigation takes a long time and many skills.
18	D	A control group allows scientists to determine if changes observed in an experiment are due to changes in the dependent variable or changes in some other variable.
19	MC	A scientific investigation is not a way to answer a scientific question.
20	D	A field study is an investigation in which scientists do experiments and collect information inside of the laboratory.
21	O	Scientists often attempt to answer questions that cannot be answered through laboratory exploration by observing the natural world.
22	D	Scientists ask questions about the planets.
23	P	Various scientific investigations are driven by specific types of research questions.
24	MC	Scientists never have questions that cannot be answered with a laboratory experiment.

25	P	Scientific inquiry is only able to provide cause-and-effect suggestions
26	O	Models are useful, but they are not exact and often lack detail.
27	MC	Science is not able to answer questions with information that is available at that time.
28	O	If learning about your world begins with asking questions and making observations, can science provide answers to these questions?
29	MC	Some observations might not push scientists to think of new experiments.
30	P	A model is a depiction of a thing or phenomenon that is used as a way to comprehend nature.
31	D	You might observe that the days get shorter in winter.
32	O	Any answer to a scientific question is uncertain because people will never know everything about the world around them.

OCTOBER

Passage 1
Grade level 6.8

Everything around us is made of matter—your clothes, the trees, even the water you drink! We divide matter into four major categories, which are called the four states of matter: liquid, gaseous, solid, and plasma; but we will focus on the first three. Whatever the state of matter may be, all matter is made of tiny particles called atoms. These particles are too tiny to see with the naked eye; they're even too small to see with a regular microscope. If you line up a million atoms next to each other, they will be as thick as a single piece of human hair. We can only look at atoms through very powerful tools such as a microscope. We can easily see liquids and solids around us, but most gases aren't visible. We can't see the air around us, but it is still made of atoms that constantly move around freely in space. When we pump air into a balloon, it visibly inflates. Gaseous matter fills the balloon and takes up space. The more air we blow into the balloon, the bigger it gets. We can observe the way gas moves around space. Inflatable pool toys also fill with air so that they can float on water. When we fill the plastic shells of pool toys with air, the toys take shape. Since air is lighter than water, the pool toys can rest on the water without sinking. When something inflates, we can see that even something like air, which is a gas, takes up space.

Sentences for Passage 1

Item	Type	
1	P	We need very powerful tools, such as a microscope, to be able to see atoms.
2	D	The temperature at which a liquid becomes a solid is called the freezing point.
3	P	Atoms in the air, even though we cannot see them, are constantly moving freely around us.
4	D	Solids can change to liquids if you add enough heat.
5	O	Whatever the state of matter may be, all matter is made of tiny particles called atoms.
6	MC	Some things like your clothes are made of matter, but some things like water are not made of matter.

7	O	We can easily see liquids and solids around us, but most gases aren't visible.
8	MC	You do not need a microscope to see atoms; you can see atoms with your eyes alone.
9	O	When something inflates, we can see that even something like air, which is a gas, takes up space.
10	MC	If inflatable pool toys are filled with air, then they cannot float.
11	D	A water molecule is made of two hydrogen atoms and one oxygen atom.
12	MC	Toys get their shapes from filling plastic with air.
13	P	When we blow more air into a balloon, it gets bigger.
14	D	A change of state is a physical change.
15	O	We can observe the way gas moves around space.
16	P	We can see a balloon inflate when we pump air into it.

Passage 2
Grade Level 7

Atoms are constantly moving. Atoms move at different speeds within different states of matter. Atoms move slower in solids than they do in liquids. Because atoms in solids are tightly packed, and there is less space to move around freely. The atoms in gas move the fastest. Since the atoms move more freely in liquids and gases, they can undergo a process called diffusion. Solids can diffuse as well, although it's a much longer process. Diffusion is the movement of particles from a higher concentration to a lower concentration. When you spray perfume in a corner of a room, you will eventually smell it on the other side of the room. The atoms from the perfume diffuse through the air. Because of this diffusion, the perfume scent is spread. The difference between gases, liquids and solids has to do with the space between atoms. When the atoms are far apart from each other and are acting independently, they are gases. When the atoms are closer together than in gases, they are liquids. Because of the space between atoms in a liquid, liquids take the shape of whatever container they are in. The atoms in solids are tightly packed, usually in a regular pattern, which is why they keep their own shape.

Item	Type	Match the sentence number to the number of the line in the text, then write an item like the item type that was selected.
1	P	Though it takes longer, atoms in solids can also diffuse.
2	D	Atoms can bond together with other atoms to form molecules.
3	D	Evaporation occurs when liquid water becomes a gas.
4	MC	All atoms move at the same speed in all states of matter.
5	O	The atoms in gas move the fastest.
6	MC	Atoms in solids have more space to move around freely because they are not as close together.
7	O	Atoms move slower in solids than they do in liquids.
8	P	Atoms are always in motion.
9	O	When you spray perfume in a corner of a room, you will eventually smell it on the other side of the room.

10	MC	Atoms in a solid are not packed closely together, which allows them to change shapes.
11	P	The atoms in liquids are closer together than in gases.
12	MC	Atoms in perfume stay in one place in the air.
13	D	A solution is a mixture in which one substance dissolves into another substance.
14	D	Physical properties of matter include the shape, color, and texture of an object.
15	O	Because of the space between atoms in a liquid, liquids take the shape of whatever container they are in.
16	P	Perfume can spread through air because of diffusion.

NOVEMBER

Properties and Changes
Grade Level 5.7

All matter has physical and chemical properties. Physical properties are what an object looks and feels like. You can observe many physical properties with your five senses. You use your senses to observe things like color, shape, smell, taste and size. Other physical properties must be measured. Length, mass, and density are some things that can be measured. The state of matter is a physical property, too. An object's state can be solid, liquid, or gas. The temperatures at which a substance boils or freezes are also physical properties. Chemical properties, on the other hand, have more to do with what the matter is made up of. All matter is made up of atoms. Water is made up of one oxygen and two hydrogen atoms. Chemical properties also deal with how substances react with each other. Different substances have a different way of reacting to things like water, air, and fire. When you add fire to paper, it will burn. When you add fire to metal, it will heat up but will not burn.

Item	Type	Match the sentence number to the number of the line in the text, then write an item like the item type that was selected.
1	P	Some physical properties, such as length, must be measured.
2	P	The states of matter are solid, liquid, and gas.
3	MC	Chemical properties are what an object looks and feels like.
4	O	You use your senses to observe things like color, shape, smell, taste and size.
5	MC	You cannot measure length, mass, or density.
6	O	The state of matter is a physical property, too.
7	D	A solid has a definite mass and a definite shape.
8	D	The particles in a solid are packed tightly together.
9	D	Magnets are materials that attract pieces of iron or steel.
10	MC	Physical properties deal with how substances react to other substances.
11	O	Chemical properties, on the other hand, have more to do with what the matter is made up of.
12	P	Metal can be heated, but it will not burn.
13	O	All matter is made up of atoms.
14	P	Two hydrogen atoms and one oxygen atom make up water.
15	D	The north pole of one magnet will repel, or push away, the north pole of

		another magnet.
16	MC	The boiling point and freezing point of a substance are some of its chemical properties.

Charges and Electricity
Grade level 7.1

Atoms are the basic building blocks of matter. Atoms are made up of protons, neutrons, and electrons. The nucleus is the center of the atom. The nucleus contains the protons and neutrons. The electrons orbit outside of the nucleus.

Protons and electrons each carry an electrical charge. The charges they carry are opposite to each other. Protons carry a positive charge. Electrons carry a negative charge. Neutrons are neutral. This means that neutrons carry no charge at all.

Electricity is the flow of electrons from one place to another. Materials that electricity can move through easily are called conductors. Most metals are good conductors. Other materials, such as rubber, wood, and glass, block the flow of electricity. Materials that block the electric flow are called insulators.

Item	Type	
1	MC	Protons, neutrons, and electrons are made of smaller particles called atoms.
2	O	Protons and electrons each carry an electrical charge.
3	P	The center of an atom is called the nucleus.
4	O	Atoms are the basic building blocks of matter.
5	P	Electrons circle around the outside of an atom’s nucleus.
6	MC	Protons carry a negative charge.
7	D	Hitting or heating a magnet can change the strength of the magnet.
8	D	Electromagnets are magnets made by electricity.
9	P	Insulators block the flow of electricity.
10	D	Friction is a force that opposes motion.
11	O	Materials that electricity can move through easily are called conductors.
12	MC	The flow of protons from one place to another is called electricity.
13	D	If you apply force to an object, you may change its energy.
14	MC	Electrons have no charge; they are neutral.
15	P	Neutrons do not carry any charge at all.
16	O	Most metals are good conductors.

JANUARY

Newton's Laws of Motion
Grade Level 6.8

Sir Isaac Newton was one of the greatest scientists and mathematicians who ever lived. He worked on developing calculus and physics at the same time. During his work, Newton came up with the three basic ideas that are applied to the physics of most motion. The ideas have been tested and verified so many times over the years, that scientists now call them Newton's Three Laws of Motion.

The first law says that an object at rest tends to stay at rest, and an object in motion tends

to stay in motion, with the same direction and speed. The first law is also known as the law of inertia. Motion (or lack of motion) cannot change without an unbalanced force acting. If you are not in motion, and no force acts on you, nothing will happen. On the other hand, if you are going in a specific direction, unless a force acts on you, you will always go in that direction at that same speed.

When we observe videos of astronauts in space, we can see some objects floating without falling to the ground. Astronauts can just place their tools in space, and the tools will stay in one place. There is no force in space that will cause the tools to change position. Likewise, if an astronaut throws something at the camera, that object will not stop moving unless a force stops it. If they threw something when doing a spacewalk, that object would continue moving in the same direction and with the same speed into space unless a force interfered with it. On Earth, when we throw an object, air resistance will slow the object down. Gravity on Earth causes objects to drop to the ground.

Item	Type	
1	P	Nothing will happen if you are staying still, and no force acts on you.
2	MC	The ideas have been tested and verified so many times over the years, that scientists now call them Newton's Three Ideas of Motion.
3	MC	The first law is also known as the law of energy.
4	D	The gravitational pull of the Sun is one type of force.
5	O	During his work, Newton came up with the three basic ideas that are applied to the physics of most motion.
6	D	Forces are acting everywhere in the universe at all times.
7	O	The first law says that an object at rest tends to stay at rest, and an object in motion tends to stay in motion, with the same direction and speed.
8	P	Unless there is an unbalanced force, motion cannot change.
9	O	Astronauts can just place their tools in space, and the tools will stay in one place.
10	MC	When we observe videos of astronauts in space, we can see that if an astronaut lets go of an object, it will fall to the ground.
11	P	On a spacewalk, if an object is thrown, that object will continue to move in the same direction at the same speed until a force acts on it.
12	P	If you are going a specific direction, you will continue going in that direction at the same speed until a force acts on you.
13	O	Gravity on Earth causes objects to drop to the ground.
14	MC	If an astronaut throws something at the camera in space, that object will probably be stopped by friction and fall to the ground because of gravity.
15	D	Energy is used to do work.
16	D	When you apply force to lift an object, you have added energy to the object and have done work.

Passage Two
Grade Level 7.8

Newton's second law states that acceleration is produced when a force acts on a mass. The greater the mass (of the object being accelerated), the greater the amount of force needed (to accelerate the object). This law is easy to understand, because it explains that heavy objects are harder to move, so they require more force. Lighter objects are easier to move, so they require less force. For example, it would require more force to move a truck that just ran out of gas than it would require to throw a baseball. Newton's Second Law can also be expressed in a mathematical equation. The equation is force = mass x acceleration.

Newton's third law says that for every action (force) there is an equal and opposite reaction (force). Forces are found in pairs. For example, when you sit in a chair, your body exerts a force downward. That chair needs to exert an equal force upward, or the chair will collapse. Acting forces encounter other forces in the opposite direction. For example, when a cannonball is fired through the air (by the explosion), the cannon is pushed backward. The force pushing the ball out was equal to the force pushing the cannon back. The effect on the cannon is less noticeable because the cannon has a much larger mass than the cannonball. That is to say that whenever an object pushes another object, it gets pushed back in the opposite direction equally hard.

Item	Type	
1	P	Lighter objects are harder to move, so they require more force.
2	O	Newton's Second Law can also be expressed in a mathematical equation.
3	Mc	Newton's Law is hard to understand because it can't explain why some things are harder to move than other things.
4	D	Velocity is the rate of motion in a specific direction.
5	Mc	The second law states that objects with mass can speed up without any force acting on it.
6	O	The third law says that for every action (force) there is an equal and opposite reaction (force).
7	P	Force = mass x acceleration is the equation for Newton's Second Law.
8	D	The study of thermodynamics has to do with the study of heat and thermal energy.
9	P	Your body exerts a force downward on a chair whenever you sit down.
10	P	Each acting force is met with another force in the opposite direction.
11	O	That chair needs to exert an equal force upward, or the chair will collapse.
12	D	Heat naturally moves from high to low temperatures.
13	MC	The force pushing the ball out out of a cannon is much stronger the force pushing the cannon back.
14	Mc	That is to say that whenever an object pushes another object, one object always pushes harder than the other.

15	D	Heat can move by conduction, convection, or radiation.
16	O	Forces are found in pairs.

FEBRUARY

Litter is Pollution You Can See
Grade level 7.6

We pollute when we add things to the environment that are harmful. Pollution can affect the air, soil, or water of an ecosystem. Pollution can make people or animals sick. Pollution can cause diseases or death. Trash on the ground is called litter. Littering is one way that people pollute. Litter harms both plants and animals. Litter can kill plants and entangle animals. Old fishing lines and nets kill many aquatic animals. Some animals might eat litter and get sick. Sea turtles are especially harmed when clear plastic sandwich bags are thrown into the ocean. The sea turtle thinks the bag is a jellyfish, which is what they eat. The bag gets stuck in the turtle's stomach. If a bag is stuck in an animal's stomach, it will eventually starve. Fishing lines, balloons, and plastic bags last a long time in the ocean. These plastic materials are not biodegradable, which means that they do not break down easily.

Item	Type	Statement
1	M	Litter is good for plants and animals.
2	D	Normally all ecosystems are naturally balanced.
3	D	Non-native species introduction can be a threat to ecosystems.
4	O	Trash on the ground is called litter.
5	P	Adding harmful things to the environment is called pollution.
6	O	Pollution can affect the air, soil, or water of an ecosystem.
7	P	Plants and animals can get sick because of pollution.
8	M	Litter could injure animals, but it does not affect plants.
9	D	If a species does not have a predator, it will grow out of control.
10	O	Old fishing lines and nets kill many aquatic animals.
11	M	Plastic materials are biodegradable, because they break down easily.
12	P	Plastic bags can get stuck in the stomach of a sea turtle.
13	P	A sea turtle will eat a plastic bag, because it looks like a jellyfish.
14	O	Some animals might eat litter and get sick.
15	D	Invasive earthworms eat different things than native earthworms.
16	M	Fishing lines, balloons, and plastic bags dissolve quickly in the ocean.

Other Ways to Pollute
Grade level 7.1

Putting harmful things into the air is called air pollution. Cars, trains, buses, and airplanes put lots of pollution into the air. Coal power plants and some factories also pollute the air. Air pollution can travel many miles and affect many different living things. Smog is one form of air pollution in cities. Putting harmful things into the soil is called soil pollution. Landfills are one source of soil pollution. Throwing away chemicals can pollute the soil. Also, soil pollutants are washed into water systems during rain. Any type of soil pollution that dissolves in water can be washed from soil into water systems. Putting harmful things into the

water is called water pollution. Dumping trash and chemicals into water pollutes it. Rain carries soil pollutants into water. So, runoff from a rainstorm can also pollute water. One form of water pollution increases nutrients in water systems, which causes an overgrowth of plants and algae. An overgrowth of algae in water can be harmful to the environment.

Item	Type	Statement
1	O	Landfills are one source of soil pollution.
2	D	A population is a group of living things in an area.
3	M	Putting harmful things in the soil is called air pollution.
4	M	Pollution does not travel through the air; it stays in one place.
5	D	Ecosystems have limited resources.
6	P	Soil can be polluted when we throw away chemicals.
7	P	One type of air pollution in cities is called smog.
8	O	Coal power plants and some factories also pollute the air.
9	O	Rain carries soil pollutants into water.
10	M	Rain helps to clean all of the pollutants away, making the soil and water less polluted.
11	D	Carrying capacity is the number of organisms an ecosystem can support.
12	P	Water gets polluted when we dump trash and chemicals into it.
13	M	Runoff from a rainstorm can NOT pollute the water.
14	P	Water can be polluted by an increase in nutrients that causes an overgrowth of algae.
15	D	A swamp can support a large population of ducks.
16	O	An overgrowth of algae in water can be harmful to the environment.

VITA

Renée Lastrapes has been a special education teacher for 16 years. She has worked with students with learning disabilities and emotional and behavioral disorders for 13 years as a middle school resource specialist in California and as an elementary mild/moderate teacher for 3 years in Louisiana. At LSU, she became interested in quantitative research methodology and pursued a doctorate in educational research with a minor in applied statistics. She enjoys teaching statistics, and will be employed in fall at the University of Houston-Clear Lake as an Assistant Professor in Educational Research and Assessment.