A PSYCHOMETRIC ANALYSIS OF TEACHER-MADE BENCHMARK ASSESSMENTS IN ENGLISH LANGUAGE ARTS

by

Andrea Milligan

A Dissertation Submitted to the
Faculty of the College of Graduate Studies at
Middle Tennessee State University
in Partial Fulfillment
of the Requirements for the Degree of
Doctorate of Philosophy
in Literacy Studies

Middle Tennessee State University
May 2017

Dissertation Committee:

Dr. Jwa K. Kim, Chair

Dr. Amy M. Elleman

Dr. Cyrille Magne

ABSTRACT

The implementation of the Common Core State Standards (CCSS) has placed increased accountability for outcomes on both students and teachers. To address the current youth literacy crisis in the United States, the CCSS call for students to read increasingly complex informational and literary texts. Since teachers are held accountable for students' mastery of the standards, reliable benchmark tests aligned to the CCSS are crucial to both student and teacher evaluation. The purpose of the study was two-fold. First, classical test theory (CTT) was used to glean information about the reliability and validity of the tests along with basic item analyses for each item. Then, exploratory factor analysis was conducted to confirm the item structure and to ensure the data were suitable for item response theory (IRT) analysis. If exploratory factor analysis revealed a unidimensional structure, IRT was applied to evaluate the strength and weakness of each item. Archival data from tenth grade students enrolled in a public high in North Alabama were used for the analysis. Data from the October, December, and March regular and honor's course benchmarks tests were analyzed. Cronbach's alpha indicated that tests were generally reliable even though honor's course benchmark test scores were less reliable than regular course benchmark test scores. In addition, exploratory factor analysis partially supported a three-factor solution. Finally, items were generally strong based on CTT and IRT calibrations.

TABLE OF CONTENTS

LIST OF TABLES

CHAPTER I: INTRODUCTION

While American schools do a fine job of imparting literacy skills to students in grades K through three, the literacy skills of the nation's late elementary through high school students have only recently been gaining attention in the research and few reading programs address the needs of students in middle and high school (Kirk, 2000). For high school students with reading comprehension problems, the consequences can be dire, and researchers have called for increased attention to the plight of literacy skills for adolescents (Alexander & Fox, 2011; Alvermann, 2001; Biancarosa & Snow, 2004; Heller & Greenleaf, 2007; Kirk, 2000; Moore, Bean, Birdyshaw & Rycik, 1999).
In addition, technological innovations and a global economy place unique literacy demands on students in the United States, and many secondary students do not have the literacy skills needed to be successful in post-secondary education and the workforce (Fang & Schleppergrell, 2010).

In response to increased literacy demands and the current adolescent literacy crisis, many states have adopted the Common Core State Standards in English Language Arts (CCSS ELA). A result of the *No Child Left Behind Act* (Klein, 2015), the CCSS ELA have increased the rigor and expectations of literacy for future generations of students. The standards are applied to prepare students for the workforce, college, and careers. The CCSS ELA standards were developed by education professionals and governors in 48 states. Currently, forty-five states and the District of Columbia have adopted the standards which delineate what students should be able to do in mathematics and English language arts/literacy.

To gauge student success and growth with the CCSS ELA, teachers in a Northern Alabama school district created a benchmark test to measure student mastery of the standards. The benchmark tests are administered at regular intervals during the school year to measure student progression on the CCSS ELA. Test items assess three strands of the CCSA ELA: language, reading for information, and reading literature. Tests items are reportedly aligned with the national CCSS ELA. Tenth grade students enrolled in honor's and regular courses complete a benchmark test in October, December, and March each academic year. These tests were written by several teachers in the district and administered to students without rigorous psychometric validation of the created test items. Even so, teachers are held accountable for student performance on the test. The current study was concerned with the benchmark assessment of the CCSS ELA strand for tenth grade students which includes competencies in language, reading literature, and reading informational text.

Purpose of the Study

The benchmark tests used in measuring students' mastery of the standards had not been psychometrically analyzed prior to the present study. The purpose of the study was two-fold. First, classical test theory (CTT) was used to glean information about the reliability of the tests along with basic item analyses for each item. Then, exploratory factor analysis (EFA) was conducted to confirm the item structure and to ensure that the data were suitable for item response theory (IRT) analysis. If factor analysis revealed a unidimensional structure, IRT was applied to evaluate the strength and weakness of each item. Archival data from tenth grade students in a public high in North Alabama were

used for the analysis. Data from the October, December, and March benchmarks tests were analyzed.

Research Questions

Through this psychometric validation study, the researcher examined the psychometric properties of the assessments using classical test theory (CTT) and item response analysis (IRT) and made recommendations on item and test selection to educators and administrators. The following questions were addressed:

1. Do the items for the honor's course benchmark test scores and the regular course benchmark test scores show a three-factor solution to match the reading for information, reading literature, and language conventions strands as described in the CCSS ELA?
2. Do the tests show strong reliability based on CTT?
3. Do the items demonstrate strong item characteristics based on CTT and IRT?

CHAPTER II: REVIEW OF THE LITERATURE

The National Literacy Crisis

Since the United States offers a free public education to all its citizens, it would seem logical that it could boast the most literate population in the world. However, this is not the case. In fact, there is cause for much concern for the literacy skills of adults in the United States. For example, The Literacy Project Foundation (2016) reports that 50% of adults in the United States are unable to read on an eighth-grade level and that 44% of adults do not read even one book per year. Also, the United States ranks twelfth out of twenty in a study of literacy among "high income" countries (Literacy Project Foundation, 2016).

Many societal problems are correlated with low literacy skills. For example, 75% of welfare recipients are functionally illiterate and three out of five prisoners in America cannot read. In fact, "to determine how many prison beds were needed in future years, some states base part of their projection on how well current elementary students are performing on reading tests" (Literacy Project Foundation, 2016, para. 2). Sadly, 85% of juvenile offenders also have trouble reading (Literacy Project Foundation, 2016). Adults who cannot read are 50% more likely to have an income below the poverty level. The cost of illiteracy to the American taxpayer is around $20 billion per year while "school dropouts cost our nation $240 billion in social service expenditures and lost tax revenues" (Literacy Project Foundation, 2016, para. 2).

These low literacy levels impact all members of society, not just those who struggle with reading. To ameliorate these dismal findings, the nation's high schools need to improve their efforts to graduate literate students (Hayes, 2011). Research has

shown that many middle and high school students lack the literacy skills necessary to be successful in educational endeavors after high school graduation (Hayes, 2011).

Literacy Crisis and the Nation's Youth

The plight of the nation's youth literacy crisis gained national attention with the publication of *A Nation at Risk* in 1983. For the first time, stakeholders in American public education had to come to grips with the fact that high schools were not graduating students prepared to meet the demands of a global economy (Hayes, 2011). Business and industry leaders report that they must spend millions of dollars on training and remediation programs since graduates do not possess basic reading, writing, spelling, and mathematical skills (*A Nation at Risk*, 1983). Critical findings of *A Nation at Risk* are:

- Functional illiteracy among minority youth may be as high as 40%.
- Many high school students do not possess the higher order thinking skills necessary to draw inferences from texts they read.
- Less than one-half of students can write a persuasive essay.

Even though *A Nation at Risk* was published over 30 years ago, anxiety for the nation's literacy skills is prevalent with much work to be done in the middle and high school grades.

Since the publication of *A Nation at Risk*, the youth literacy crisis in the United States has been given national attention with much of concern being placed on early reading instruction and outcomes in the elementary grades. In kindergarten through third grade, reading instruction makes up the bulk of daily instruction and classroom activities (Heller & Greenleaf, 2007). Because of this increased focus, reading achievement for fourth grade students has steadily increased over the years with the strongest gains being

seen in minority students and students in poverty (Heller & Greenleaf, 2007). The growth is a result of the increased resources and attention given to early literacy development (National Center for Education Statistics, 2015). While this is a welcome report, students in the high school grades have not fared so well.

The early elementary years of literacy instruction are spent teaching children *how* to read. In the late elementary grades, however, students are expected to read in order to learn (Hayes, 2011). While American fourth graders score among the best in the world with regard to literacy achievement, by tenth grade, American students "place close to the bottom among developed nations" (Hayes, 2011, p. 10). According to the National Center for Education Statistics (2015), reading scores for high school seniors have remained relatively flat. Heller (2016) reports that "for more than three decades now, many of the nation's secondary school students have failed to demonstrate the expected competence in reading and writing, and only a handful of students -- 3 percent of 8th graders in 2007 -- have been found to read at an advanced level" (para. 4).

According to the Nation's Report Card, reading scores in 2015 were actually *lower* than those in 1992 when the initial reading assessment was given. Only 37% of high school seniors scored at or above *Proficient* on the national reading assessment in 2015 (Nation's Report Card, 2015). Results from American College Test Incorporated (ACT) (2014) report that only 44% of students taking the ACT are college-ready while 86% of students indicated plans to continue their education beyond high school. More students are planning to continue their education beyond high school, but many are simply not prepared to meet the demands of higher education.

The sobering fact is that as much as 60% of high school seniors will need some kind of remediation course upon entering college or university. Such courses award no credit toward graduation (National Center for Public Policy and Higher Education, 2010). According to Wise (2009), "Because too many students are not learning the skills they need to succeed in college or work while they are in high school, the nation loses more than $3.7 billion a year in costs associated with college remediation" (p. 372). The importance of literacy in the secondary grades cannot be overstated because students with only basic literacy skills will not be prepared to meet the demands of college course work or a competitive job market (Heller & Greenleaf, 2007). On a more positive note, the crisis in adolescent literacy is receiving more attention in the research without the "reading wars" that have surrounded the teaching of literacy skills in the early grades (Heller, 2016).

Transition to the Common Core State Standards in English Language Arts

In response to increased literacy demands and the current adolescent literacy crisis, many states have adopted the Common Core State Standards in English Language Arts (CCSS ELA). A result of the *No Child Left Behind Act* (Klein, 2015), the CCSS ELA have increased the rigor and expectations of literacy for future generations of students. The standards are applied to prepare students for the workforce, college, and careers. The CCSS ELA were developed by education professionals and governors in 48 states. Currently, forty-five states and the District of Columbia have adopted the standards which delineate what students should be able to do in mathematics and English language arts/literacy.

The goal of the CCSS is to standardize learning outcomes and expectations across the nation to prepare students in kindergarten through 12th grade for college-level credit courses and entry into the workforce (Common Core State Standards Initiative, 2016). Because of these new, rigorous standards, more is required from the students, and teachers must create innovative ways to teach and assess mastery of the standards (Kibler, Walqui, & Bunch, 2015).

Prior to the implementation of the CCSS ELA, states mandated their own standards. Students were primarily required to read texts and recall basic facts about the text. However, the CCSS ELA calls for students to do much more than simply recall facts or provide a summary. With the new CCSS ELA expectations for both narrative and informational texts, anchor standards require students to identify key ideas and details, explain the structure of the text contributes to the overall meaning and/or theme, integrate their own knowledge and ideas, and read from range of reading levels and text complexity. In other words, students must not only read the text, but they must use the information to form new opinions through inferencing while supporting their opinions with specific evidence from the text. The following expectations are emphasized in the CCSS ELA. In addition to reading for information and reading literature, the language strand of CCSS ELA requires students to master the conventions of standard English as required in college-level coursework.

Expectation 1. Students are to receive regular practice with <u>complex texts</u> and their academic language. Key to this practice is requiring students to interact with more informational text starting in kindergarten. Exposure to informational text is paramount to building background knowledge and academic vocabulary that are necessary as

students progress in school. As students encounter progressively more difficult texts, they will also be expected to learn and apply more difficult academic vocabulary. Academic vocabulary is expected to increase through reading, direct instruction, and speaking (Common Core State Standards, 2016).

Expectation 2. Students are expected to read, write, and speak using grounded evidence from informational and literary texts. The objective of this standard is to encourage students to read texts carefully to answer questions and to form arguments based on evidence from the text. Students must answer text-dependent questions rather than relying on background knowledge (Common Core State Standards Initiative, 2016)**.** For both narrative and expository texts, students are expected to "cite strong and thorough textual evidence to support analysis of what the text says explicitly as well as inferences drawn from the text" (Common Core State Standards Initiative, 2016). Students are to think critically about a text's craft and structure, integrate knowledge and ideas, make inferences, and provide objective summaries when reading a range of both narrative and expository texts. Since many states have adopted the CCSS, students are expected to do more with text than simply answer text-based comprehension questions or write a summary (Common Core State Standards Initiative, 2016).

Expectation 3. Students are expected to build knowledge through content-rich nonfiction. While students are still expected to read and comprehend fiction, the CCSS also emphasizes nonfiction texts. Traditionally, literacy instruction in grades 6-12 has focused primarily on fiction. However, the CCSS stipulates that literacy instruction should consist of a 50-50 balance of fiction and nonfiction texts.

Issues in Assessing the CCSS ELA

While the CCSS ELA delineate the skills a student should acquire prior to high school graduation, assessing these skills is problematic. Educators assess a child's reading comprehension and language skills to monitor progress and to identify students who are having difficulties mastering the standards (Cain & Oakhill, 2006; Torgesen & Miller, 2009). Cain and Oakhill (2006) reiterate the importance of effective reading assessments by claiming that "the accurate assessment of reading comprehension ability is crucial for empirical research, the development of our theoretical understanding of the reading process, and to ensure appropriate identification of and intervention for children with reading comprehension impairments" (p. 704). Identifying students with comprehension difficulties is paramount so that the correct interventions can be implemented.

Resource allocation decisions are often made based on student assessments, so the correct assessment is crucial if students are to receive the instruction and resources needed to improve (Torgesen & Miller, 2009). Ideally, assessments help teachers identify any weaknesses a student may have and allow them to adjust instruction accordingly so that students are able to meet the demand for high-level literacy skills necessary for success in the workplace and post-secondary education (Torgesen & Miller, 2009).

Educators have a daunting task when it comes to choosing the correct assessment since there are many published tests available. Unfortunately, due to time and budget constraints in school settings, tests that are easy and inexpensive to administer and score are often used to assess comprehension (Keenan & Meenan, 2014). Also, educators may

believe that all comprehension tests are created equally and measure the same construct. In an ideal world, a child's reading comprehension would be assessed using a battery of tests instead of relying on a single test since studies have shown that comprehension assessments are not interchangeable (Cutting & Scarborough, 2006; Eason et al., 2012; Keenan & Meenan, 2014; Nation & Snowling, 1997). Without accurate, reliable, and valid measurements, any further attempts to predict student performance on other variables and to enhance student reading are futile.

Benchmark Tests

When the *No Child Left Behind* act was implemented in 2002, schools faced increased accountability for standards-based student achievement and teacher effectiveness (Klein, 2015). In an effort to monitor both student achievement and teacher effectiveness, many schools adopted the practice of benchmark testing (Abrams, McMillan & Wetzel, 2015; Bancroft, 2010). Educators must consider whether to choose a norm-referenced test or a criterion-referenced test since each type of test yields different information. For example, norm-referenced tests measure a child's performance against his or her peers "even though the entire population was not tested" (Kirk & Vigeland, 2014, p. 365). Such tests are typically used for achievement measures (Behuniak & Tucker, 1992; Bell & McCallum, 2008). Conversely, a criterion-referenced test assesses students on a given set of standards and is used primarily to guide instruction and curriculum choices (Behuniak & Tucker, 1992; Bell & McCallum, 2008). A benchmark test is generally a criterion-referenced test.

Benchmark tests are tests which are administered at regular intervals during the school year (usually quarterly) to measure student progression in state or national

education standards (Bell & McCallum, 2008). According to Bancroft (2010), "systems of regular, intermittent benchmark tests have become increasingly utilized as a means to have greater surveillance of teaching and learning, with the ultimate goal of closing achievement gaps" (p. 99). Also known as interim assessments, schools administer such benchmark tests to evaluate curriculum and plan instruction (Abrams, McMillan, & Wetzel, 2015; Reed, 2015; Shapiro, Hilt-Panahon, Gischlar, Semeniak, Leichman, & Bowles, 2012). Marsh, Pane, and Hamilton (2006) report that 80% of superintendents and 80% of principals believe that benchmark assessments are helpful when making instructional decisions.

The efficacy of using benchmark data to drive instruction is uncertain. Henderson, Petrosino, Guckenburg, and Hamilton (2007) found that students who participated in a mathematics benchmark exam did not perform any better on an end-of-year assessment than students who did not complete a benchmark exam. In addition, Reed (2015) found that teachers did not use benchmark data to guide instruction. There may be several reasons for this. First, teachers may have lacked the necessary resources to implement new instructional strategies. Also, the teachers reported that they had little confidence in the assessment. Furthermore, teachers reported that students did not take the tests seriously and were administered too often (Reed, 2015).

Bancroft (2010) reports that benchmark data for students who are below grade level do not provide the necessary information to remediate student weaknesses. Such standards-based assessments may be "insensitive to the instructional needs of many struggling readers who continue to have difficulties with word-level skills in middle and

high school" (Torgesen & Miller, 2009, p. 9). In other words, if a student lacks basic literacy skills, he or she will not be able to meet the demands of a benchmark assessment.

Teachers maintain that the number of standards students are required to master is too broad. In the state of Alabama, students in the tenth grade are required to master over forty standards (with multiple skills in each standard) in the areas of language, speaking and listening, reading literature, reading informational text, and writing (Bice, 2016). The standards assume that students already possess the foundational reading skills necessary for success in high school.

On the other hand, benchmark tests can be a valuable tool to aid instruction if used correctly (Black & Wiliam, 1998). Abrams, McMillan, and Wetzel (2015) found that teachers who perceived that the test items were of high quality, who had timely access to benchmark data, and could discuss results with peers were more likely to adjust their instructional methods and pace. In their meta-analysis of interim evaluation, Fuchs and Fuchs (1986) found that teachers who received the most support in interpreting and using the data resulted in higher student achievement.

While benchmark tests can be useful, teacher-made district-wide benchmark tests may prove more problematic if such assessments are not subjected to the same rigor as standardized tests. This issue is of extreme importance since student test scores may be tied to teacher pay raises and tenure decisions. Now, more than ever, teachers are held responsible for student outcomes on tests. Since the reading benchmark tests discussed in this study had not been psychometrically analyzed, teachers were concerned with the tests' reliability along with item calibration and strength. Like the teachers in Reed's (2015) study, teachers had little confidence in the tests. With merit pay and even careers

at stake, sound reading benchmark assessments are paramount. As it is true in every discipline, an accurate measurement of a given trait or construct is one of the fundamental components of science. Psychometric theories provide important information for practitioners regarding the efficacy of teacher-made benchmark tests.

Psychometric Theories

Testing is ubiquitous. Thousands of tests exist to measure achievement, personality traits, intelligence, attitudes, and so on. From the time a student enters school, he or she is subject to a myriad of tests, both academic and psychological. Results of such tests can decide a child's educational future. Results can determine whether child will receive special education services or other instructional modifications to help him or her succeed (Furr & Bacharach, 2014; Raykov & Marcoulides, 2007).

Tests can also tell educators if a student has made adequate progress in a particular area across the school year (Furr & Bacharach, 2014). These tests may also determine the amount of scholarship money high school graduates will receive as well as which college they will attend. Results of certain tests can impact a child's future in numerous ways.

Given the importance of testing, practitioners must be vigilant to ensure such tests are valid and reliable and truly measure the construct in question. To investigate if test items are valid and test scores are reliable practitioners and educators must understand the underlying theories of testing and measurement. There are two theoretical approaches in psychometrics: classical test theory (CTT) and item response theory (IRT).

Classical test theory. Through symbolic representation**,** test theories and models allow test makers to measure the factors that influence observed test scores. These test theories and models are illustrated by various assumptions (Allen & Yen, 2001). Classical test theory (CTT) is the most well-known and widely used method in test construction and validation. CTT, also known as *weak* or *true-score* theory, allows the psychometrician to examine how the error of measurement influences the observed scores (Allen & Yen, 2001). The CTT model is composed of assumptions regarding the observed score, the true score, and the error of measurement (Furr & Bacharach, 2014). Allen and Yen (2001) describe six assumptions related to CTT.

*Assumption 1***.** The first assumption in CTT posits that every student has a true score. A student's true score is defined as how much of an ability or aptitude a student possesses. For example, teachers might wish to know the student's true score in reading comprehension, spelling, vocabulary, or mathematics (Furr & Bacharach, 2014; Gall, Gall, & Borg, 2007; Hambleton & Jones, 1993; Lord, 1980).

An important consideration in CTT is to differentiate between the observed score and the true score. The observed scores (X) provide a frequency distribution from which the mean (expected value) is derived. The mean of this frequency distribution is also called the *true score (T)* which is assumed to have fixed value (Allen & Yen, 2001; Furr & Bacharach, 2014; Lord, 1980). Allen and Yen (2001) explain the *true score (T)* is "the mean of the theoretical distribution of X scores that would be found in repeated independent testings of the same person with the same test" (p. 57). In other words, a student's true score is equal the average of his observed scores if he were able to take a

test an infinite number of times. Since this is impossible in a real-world setting, the *T* is the theoretical construct (Allen & Yen, 2001). The observed score (*X*) is noted as

$$X = T + E \quad (1)$$

where *E* equals the measurement of error (Allen & Yen, 2001; Raykov & Marcoulides, 2007). Raykov and Marcoulides (2007) describe this equation as classical test theory decomposition. In a perfect assessment, the observed score (*X*) would equal the true score (*T*) which would result in a measurement error of 0. It is impossible to have a measurement error of 0 in a real-world testing situation since the true score cannot truly be known. The mathematical model for CTT measurement of error is

$$E = X - T \quad (2)$$

In CTT, a student's true score (*T*), which theoretically measures ability, is of most concern. As Lord explains (1980), "When a job applicant leaves the room where he was tested, it is *T*, not *X*, that determines his capacity for future performance" (p.5). Since no assessment is perfect, the second assumption in CTT is that every test will have some measurement error which are randomly distributed. For example, if 100 essays were graded, some raters will assign high scores while other raters may assign lower scores. In this way, the high and low scores would counterbalance each other, resulting in the random distribution of the errors of measurement (Gall, Gall, & Borg, 2007).

Assumption 2: The second assumption of CTT states that

$$\varepsilon(X) = T \quad (3)$$

in which the expected value of *X* is equal to the true score *T*, a theoretical construct since T is derived from a student taking a test an infinite number of times. Based on this assumption, Allen and Yen (2001) state that (theoretically) each time a student takes a

test, this will not impact the score on following tests. In summary, "each testing has no influence on subsequent testing" (Allen & Yen, 2001, p. 57).

Assumption 3. Assumption three states that a population's true scores and error scores on a single test are uncorrelated. (Allen & Yen, 2001, Furr & Bacharach, 2014; Raykov & Marcoulides, 2007). As Allen and Yen (2001) explain, "This assumption implies that examinees with high true scores do not have systematically more positive or negative errors of measurement than examinees with low scores" (p. 58). For example, if a student with low mathematics ability cheats on a test and does very well, then his or her true score and error score would be negatively correlated.

Assumption 4. Like Assumption 3, Assumption 4 relates to correlation but with regards to measurement error of multiple testing situations instead of a single test. Assumption 4 indicates that the measurement errors from two tests are uncorrelated. However, this assumption may not be valid if the testing environments for the two tests differ significantly from each other. Fatigue, the student's mood, the environment, or practice effects could impact the scores on a second test (Allen & Yen, 2001; Raykov & Marcoulides, 2007)

Assumption 5. Like Assumptions 3 and 4, Assumption 5 relates to the correlation of scores. Based on Assumption 5, the error scores on one test and the true scores on the other tests are not correlated with each other (Allen & Yen, 2001; Furr & Bacharach, 2014; Raykov & Marcoulides, 2007).

Assumption 6. This assumption is concerned with the existence of parallel tests. Parallel tests are defined as two tests with the same observed score (X), true score (T) and error variance, σ_E^2.

Reliability. The true score and measurement error provide information on test reliability. According to Gall, Gall, and Borg (2007), in CTT, "the reliability of a test refers to the degree to which measurement error is absent from the scores yielded by the test. (Note in this definition that reliability is a property of test *scores*, not of the test itself)" (p. 200, emphasis added). Furr and Bacharach (2014) describe reliability as the ratio of true score variance to observed score variance. The reliability index represents the consistency of test scores across different times.

As stated earlier, both the true score and the measurement of error are hypothetical constructs that are estimated by statistical means which yield a reliability coefficient. The coefficient varies from .00 to 1.00. The higher the reliability coefficient, the more reliable the test. Allen and Yen (2001) assert that a reliable test has a strong correlation between true and observed scores. "Or reliability can also be expressed as a correlation coefficient between observed scores on two parallel tests" (Allen & Yen, 2001, p. 72).

Shortcomings of CTT. Despite its popularity, CTT has several shortcomings. First, statistical information gleaned from a test is dependent upon the sample of students who took the test. For example, if two very different groups took the same test, the measures obtained would be different. Also, the test might be too difficult for some and too easy for others. Thus, the student's true score would be a poor estimate of ability.

In addition, CTT assumes that the measurement of error is the same for all students. While allowing students to take alternate forms of a test may alleviate some of these concerns, in reality, it is impossible to construct a completely parallel test (Gall, Gall, & Borg, 2007; Hambleton, Swaminathan, & Rogers, 1991; Lord, 1980).

Lord (1980) claims that "when two or more 'parallel' tests are published, we usually find that a person obtains different scores on different test forms" (p. 3). In practice, educators simply cannot administer multiple forms of a test because of time and budget constraints (Lord, 1980).

Tests constructed with CTT are also prone to the *ceiling effect* (Allen & Yen, 2001) which is the point in the test that a student is not likely to miss an item (Bell & McCallum, 2008). For example, suppose two students take the same math test worth 100 points. Both students score 100. While it may seem that both students have the same ability in math, the test may have been too easy since both students scored the maximum points. However, due to this ceiling effect, it is impossible to know the true ability of each student (Allen & Yen, 2001; Bell & McCallum, 2008).

Item response theory. In response to the inherent weaknesses of CTT, item response theory (IRT) has become a viable alternate for test makers. Gaining prominence in the 1970s, IRT was first used in the development of standardized tests such at the Scholastic Aptitude Tests (SATs) and the Graduate Record Exam (GRE) (An & Young, n.d.; Yang & Kao, 2014). In contrast to CTT, IRT is mainly concerned with individual test items rather than the test (Baker, 2001). According to Lord (1980), IRT "involves making predictions about things beyond the control of the psychometrician – predictions about how people behave in the real world" (p. 11). Each item represents a single ability. As such, students with varying abilities will perform differently on the item. Finally, the relationship between ability and item performance can be illustrated by a mathematical function called the item response function. The function examines the relationship

between the latent trait represented by θ (unobserved behavior) and ability (Gall, Gall, & Borg, 2007; Hambleton, Swaminathan, & Rogers, 1991; Lord, 1980).

Assumption of local independence. IRT models maintain local independence of test items. In other words, "it is assumed that the [student's] responses to questions are not statistically related to each other, even after the latent trait is taken into consideration or statistically held constant" (Yang & Kao, 2014, p. 172). Local independence assumes unidimensionality or that only one trait is being measured. A student's "response to one item is not contingent on his or her response to another item" (Yang & Kao, 2014, p. 173).

The item response function. For a dichotomous test item, the item response function is simply the probability that the students will get the correct answer represented by $P(\theta)$. The higher the student's ability, the more likely he or she will choose the correct answer (Lord, 1980). Lord (1980) represents a three-parameter logistic function as follows:

$$P = P(\theta) = c + \frac{1-c}{1+ e^{-1.7a(\theta-b)}}. \qquad (4)$$

The *a*, *b*, and *c* are the parameters characterizing an item.

Parameter *a* measures the power of the item to discriminate among students with varying ability levels or different levels of the latent trait θ (An & Young, n.d.; Yang & Kao, 2014; Yu, 2013). As explained by Yang and Kao (2014), "the item discrimination parameter is also called the slope parameter, with steeper slopes at a particular θ level offering better discrimination than less steep slopes, as depicted on the item characteristic curve" (p. 173).

Parameter *b* is the item difficulty index or location parameter. The parameter estimates how difficult it is for a student to achieve a 0.5 probability of answering the item correctly at a given level of θ (Yang & Kao, 2014). Students who struggle for a 50% chance of answering a question correctly may have less ability (or other latent trait) than students who find it easy to achieve a 50% chance of answering an item correctly (Yang & Kao, 2014; Yu, 2014).

Parameter *c* is the guessing parameter or pseudo-chance parameter. Parameter *c* "is the probability that a person completely lacking in ability will answer the question correctly" (Lord, 1980, p. 12). "Students with low ability may guess correctly on a multiple choice test item, which would be accounted for by the guessing parameter or pseudo-chance level" (Yang & Kao, 2014, p. 174).

IRT models. Depending upon the research question, a one-parameter logistic model (1-PL), also known as the Rasch model, can be used in which the item discrimination value is held constant in order to estimate item difficulty (Yang & Kao, 2014; Yu, 2013). In the two-parameter logic model (2-PL), both item discrimination (*a*) and item difficulty (*b*) are estimated. Finally, the three-parameter logic model (3-PL) estimates all three item parameters: item discrimination (*a*), item difficulty (*b*), and the guessing parameter (*c*) (Yang & Kao, 2014; Yu, 2013).

Hambleton and Jones (1993, p. 43) summarize the differences between CTT and IRT in Table 1. Since CTT and IRT provide different information regarding the benchmark tests, applying both will ensure a thorough examination of the assessments.

Table 1

Comparison of CTT and IRT

Area	Classical test theory	Item response theory
Model	Linear	Nonlinear
Level	Test	Item
Assumptions	Weak (easy to meet with test data)	Strong (more difficult to meet with test data)
Item-ability relationship	Not specified	Item characteristic functions
Ability	Test scores or estimated true scores are reported on the test-score scale (or a transformed test-score scale)	Ability scores are reported on the scale $-\infty$ to $+\infty$ (or a transformed scale)
Invariance of item and person characteristics	No-item and person parameters are sample dependent	Yes – item and person parameters are sample independent, if model fit the test data
Item statistics	*p, r*	b, a, and c (for three parameter model) plus corresponding item information functions
Sample size (for item parameter estimation)	200 to 500 (in general)	Depends on the IRT model but larger sample, i.e., over 500, in general are needed

from Hambleton & Jones (1993, p. 43)

CHAPTER III: METHODOLOGY

Participants

Participant data for the study were drawn from tenth grade honor's and regular English classes at a large public high school in Northern Alabama. Tenth graders were chosen for the study to allow time for teachers to implement any necessary changes to instruction and assessment prior to student graduation. Demographic data for the tenth grade students are presented in Table 2. Out of the 1,940 students enrolled in grades 9-12, 25% receive free or reduced lunch (Alabama State Department of Education, 2016). The number of students in grade 10 receiving free or reduced lunch is unavailable. The total reported enrollment for tenth grade is 508 (Alabama Department of Education, 2016); only 457 reported demographic data, and not every student took each benchmark due to absences on test day. The data do not include any information that could lead to the identification of participants.

Table 2

Demographic Data for Tenth Grade

Ethnic Group	Male	Female	Percentage
Caucasian	147	167	68.7
African-American	75	68	31.3
Hispanic	14	16	0.07
Total	206	251	100

Measurement

For the purposes of this study, the researcher examined a series of benchmark tests administered to tenth grade students in a large public high school in North Alabama. The tests were written by English teachers in the district who received no formal training in writing test items. Teachers were volunteers who were compensated for their time. All tests contain multiple choice items with four alternatives. Tests were administered October, December, and March of the 2015-2016 school year. Students took the tests on a computer. Tests were scored automatically, so a student knew his or her score immediately. Not all test items required students to read a passage and answer questions. For some items, students were asked to answer questions about literary terms, grammar, usage, and punctuation.

The tests were purported to align with the Common Core State Standards in English Language Arts in language, reading for information, and reading literature. Standards were assessed based on the district's pacing guides. Therefore, the texts chosen for the tests were aligned with the district's pacing guide.

Students in grade 10 English language arts study early American literature (to 1865), so passages are similar to those students read in class. For example, in the October benchmark tests, student read "Benjamin Franklin's Speech to the Constitutional Convention" and Lincoln's Gettysburg address as part of the reading for information section. For the reading literature portion, students read an excerpt from *The Awakening* by Kate Chopin. For the language portion, students were required to answer questions on sentence structure, subject verb agreement, hyperbole, allusions, and metaphors.

The December and March courses were similar. A copy of the standards assessed for each test can be found in Appendix A.

Procedures

The archival data used in the study contained a total of 1,070 responses from a series of six reading comprehension benchmarks tests broken down into three regular course tests and three honor's course tests. The first test (A) was administered in October, the second test (B) was administered in December, and the third test (C) was administered in March. The archived data were collected during the 2015-2016 academic year as the results for a benchmark assessment. There are no missing data for each assessment item.

The data were analyzed using both CTT and IRT approaches. Means and standard deviations for each item in all six assessments were computed. Test reliability was determined by computing Cronbach's alpha (*a*). Exploratory factor analysis (EFA) was used to determine the alignment of each assessment to the CCSS ELA strands of language, reading for information, and reading literature. EFA was also used to discern the unidimensionality of data for IRT analysis. Based on the results of EFA, IRT analyses were conducted to empirically select the best fitting IRT model. Then, the CTT and IRT results were examined to evaluate strong and weak items. The evaluation results were communicated to the North Alabama School district to improve their item construction and test refinement. The results of this project have the potential to help educators and practitioners create a benchmark test.

A psychometrically validated test will ensure teachers receive correct feedback regarding a student's progress. With this feedback, teachers can implement necessary

instructional strategies to improve student benchmark scores. Since the benchmark results make up a portion of teacher evaluation scores, correct student evaluation is paramount.

CHAPTER IV: RESULTS

Exploratory Factor Analysis for Regular and Honor's English Language Arts Courses

The benchmark tests were designed to measure student mastery of the three primary standards: language, reading for information, and reading literature. Prior to conducting exploratory factor analysis (EFA), Cronbach's alpha was computed for each assessment as measure of reliability. Results for the benchmark tests for the regular and honor's English language arts (ELA) classes are found in Table 3. While there is no universally agreed upon acceptable value for Cronbach's alpha, assessments with a value of 0.7 or greater are generally considered reliable (Bonett & Wright, 2015; Nunnally, 1978). Based on this criterion, all benchmark tests for the regular ELA courses were reliable. The benchmark test scores for honor's courses were not as reliable with the March course being the least reliable of all the tests in the study.

Table 3

Cronbach's Alpha Values for Regular and Honor's Course ELA Benchmark Assessments

Assessment	*N*	*Number of Items*	*Cronbach's Alpha*
October			
Regular Course	129	24	0.803
Honor's Course	218	24	0.616
December			
Regular Course	139	35	0.820
Honor's Course	223	41	0.661
March			
Regular Course	136	27	0.738
Honor's Courses	225	27	0.497

After computing Cronbach's alpha, EFA was conducted with the principal component method to confirm alignment with the language, reading for information, and reading literature strands. The Promax rotation method was used to obtain a more interpretable item structure. Initial EFA for the regular language arts courses yielded a total of 9-12 components with an eigenvalue of greater than 1. Initial EFA for the honor's courses yielded a total of 11-18 components with an eigenvalue greater than 1. Table 4 shows the first three eigenvalues and the percent of variance explained by the components.

The pattern matrix (factor loadings) for each factor analysis was examined to determine which items were loaded on each component based on a factor loading value of .40 or greater. Since the tests measure three strands – reading literature, reading for information, and language – similar items would fall under the same component. However, this was not the case as indicated in Table 5.

Table 4

Eigenvalues and Cumulative Percent of Variance Explained for the First Three Components by Regular and Honor's ELA Class Benchmark Tests

Test	*Component*	*Eigenvalue*	*Cumulative %*
October			
Regular Course	1	4.76	19.81
	2	1.71	26.92
	3	1.48	33.08
Honor's Course	1	2.78	11.65
	2	6.67	18.32
	3	6.19	24.51
December			
Regular Course	1	5.88	16.76
	2	2.14	22.87
	3	1.83	28.11
Honor's Course	1	3.74	9.12
	2	1.83	13.59
	3	1.69	17.71
March			
Regular Course	1	4.54	16.82
	2	1.72	23.17
	3	1.67	29.36
Honor's Course	1	2.85	10.56
	2	1.67	16.77
	3	1.57	22.59

Table 5

Factor Analysis Pattern Matrices for Regular and Honor's ELA Courses

(factor loading ≥ .40)

	Regular				*Honor's*		
	Reading Literature Items	Reading for Information Items	Language Items		Reading Literature Items	Reading for Information Items	Language Items
October				*October*			
Component 1	2	4	0	Component 1	1	4	0
Component 2	1	3	1	Component 2	0	1	5
Component 3	3	1	1	Component 3	2	2	1
December				*December*			
Component 1	3	3	2	Component 1	5	0	1
Component 2	5	1	1	Component 2	2	1	0
Component 3	1	3	1	Component 3	1	1	0
March				*March*			
Component 1	1	5	3	Component 1	0	5	0
Component 2	1	2	4	Component 2	2	1	0
Component 3	1	1	1	Component 3	3	3	0

CTT and IRT Analysis

Tables 7-12 compare the CTT and IRT results for easiest, most difficult, weakest, and strongest items. CTT was implemented to determine item *p*-values and item-test correlations. Item *p*-values indicate the average number of participants who answered the item correctly. *P*-values denote the difficulty of an item. Generally, tests with *p*-values ranging from .30 to .80 are desirable since these values represent a range of difficulty. In other words, a *p*-value below .30 may be too difficult while a value above .80 may be too easy for participants. According to Kehoe (1995), "This point may be summarized by saying that items answered correctly (or incorrectly) by a large proportion of examinees (more than 85%) have markedly reduced power to discriminate. On a good test, most items are answered correctly by 30% to 80% of the examinees" (1995, p. 1). Likewise, item-test correlations of .50 and above are desirable since these items discriminate more clearly among test takers. In summary, "low [item-test correlations] are usually due to an excess of very easy (or hard) items, poorly written items that do not discriminate, or violation of the precondition that the items test a unified body of content" (Kehoe, 1995, p. 3). In addition to *p*-values and item-item test correlations, IRT analysis was also implemented since EFA scree plots indicated that the data were unidimensional.

October Regular Findings

CTT results. Results of both CTT and IRT analyses are presented for all six tests in Tables 7 through 12. *P*-values for the October regular assessment ranged from 0.341 to 0.744. Both the most difficult item (item 3, $p = 0.341$) and the easiest item (item 20, $p = 0.744$) were from the reading for information strand. Based on Kehoe's (1995) assumption, the October regular course assessment presents items is within an acceptable

range of difficulty. The item-test correlations range from 0.024 (item 16, language) to 0.611 (item 10, reading literature). While item 16 has almost no discriminating power, item 10 has a strong predictability of the total scores. Based on the CTT findings the October regular course (Table 7), the reading items (literature and information) are stronger than the language items in terms of power to discriminate.

IRT results. IRT results for the October regular course can be found in Table 8. IRT analysis revealed similar findings to those of CTT based on individual item analysis. The language items were found to be the weakest in both the *a* and *b* parameters while the strongest items were from the reading strand. For example, for IRT parameter *a,* item 1 (language) was the weakest item with a value 0.388. The strongest item in parameter *a* was item 10 (reading literature) with a value 1.230. Parameter c (guessing parameter) ranges from 0.235 to 0.327. Since each item has four alternatives, an average of .25 of the *c*-estimates was expected. Except for item 3, Parameter *c* estimates were all in the range of .25 for the October data.

October Honor's Findings

CTT results. Both the easiest and most difficult items came from the reading strands. Item 9 (reading for information; $p = 0.954$) was the easiest item while item 22 (reading literature; $p = 0.225$) was the most difficult item. Overall, the most difficult items and their respective *p*-values were number 21 ($p = 0.271$), number 22 ($p = 0.225$), and number 23 ($p = 0.284$). Each of these items were from the reading literature strand. Items 3, 4, 7, 9, 10, 11, and 12 showed *p*-values above 0.80 which indicate that these items may be too easy for students. Item – test correlation values ranged from 0.000 (item 21, reading literature strand) to 0.354 (item 12, reading for information).

IRT results. The IRT analysis revealed that the weakest item in parameter *a* is item 5 (a language item) with a value of 0.706 while the strongest item is number 12 (reading for information) with a value of 1.130. Parameter *b* shows that item 9 (reading for information) with a value of -3.031 is the easiest item while the most difficult item is 22 (reading literature) with a value of 3.440.

December Regular Findings

CTT results. *P*-values for the December regular assessment ranged from 0.173 to 0.813. Items with a *p*-value of less than 0.30 may prove too difficult for students. Reading strand items 1, 9, 14, 23, and 26 have a *p*-value of less than 0.30. The easiest items were items 3 and 8 which were both from a reading strand. The item-test correlations range from -0.067 to 0.527. Both lowest and the highest item-test correlations were from the reading literature strand.

IRT results. According to IRT analysis parameter *a*, item 8 (reading literature) was the strongest. Item 22 (language) was the weakest item. For parameter *b*, the most difficult and easiest items were 9 (reading literature) and 3 (reading for information), respectively. Since these assessments provided students with four multiple choice answers, the expected parameter of *c* is 0.25. However, no item reached this level. Parameter *c* values ranged from 0.198 to 0.214.

December Honor's Findings

CTT results. The December honor's course showed *p*-values ranging from 0.157 to 0.987 with the most difficult item coming from the reading for information strand and the easiest item coming from the reading literature strand. Item 10 ($p = 0.157$; reading literature) was the most difficult item. Item 17 ($p = 0.395$; reading for information) was

the next most difficult item. Twelve items had a *p* – value greater than .80, roughly one-third of the test items. Item – test correlation values ranged from -0.101 (language strand) to 0.426 (reading literature).

IRT results. The IRT analysis revealed that the weakest item in parameter *a* is item 2 (reading for information) with a value of 0.335 while the strongest item is number 8 (reading literature) with a value of 0.870. Parameter *b* shows that item 3 (reading for information) with a value of -3.602 is the easiest item and also has the highest value for the pseudo-guessing parameter *c* (0.380). The most difficult item for parameter *b* is item 10 (p = 4.000, reading literature).

March Regular Findings

CTT results. According to CTT, the easiest item for the March regular course was item 5 (language) with a *p* – value of 0.801. Item 16 was the most difficult item (*p* = 0.235, reading for information). Item 10 (*p* = 0.595; reading literature) and item 19 (*p* = -0.105; reading for information) had the highest and lowest item-test correlation.

IRT results. For item response analysis, item 20 (*a* = 1.148; reading for information) had the highest discrimination index. Item 22 (a = 0.455) also a reading for information item, had the lowest discrimination index. For parameter *b*, reading items 5 and 19 had the lowest and highest difficulty index at -0.921 and 3.124 respectively Parameter *c* ranged from 0.210 to 0.263.

March Honor's Findings

CTT results. While the March regular course showed the weakest and strongest items in the two reading strands, the March honor's course CTT analysis revealed that the

most difficult item was number 7, a language item, with a *p*-value of 0.191. This item also showed the lowest item-test correlation (-0.097). The easiest item was 13 (*p* = 0.978). A total of 11 items had *p* – values higher than 0.80. Items Item 3 had the largest item-test correlation value (0.374).

IRT results. For the *a* parameter, item 13 was the highest discriminating item (0.708) while item 15 was the lowest (0.300). Item 13 had the lowest *b* parameter (-3.570). Item 7 is the most difficult item with a *b*-value of 3.951. The *c* parameter ranged from 0.219 to 0.259. Since the students were given four answer choices, this is an acceptable range for parameter *c*, the guessing parameter.

Model Fit Tests

Table 6 lists results of the model-fit tests. Model fit tests results varied. For the October regular course, there was no difference between the 2PLM and 3PLM model. However, considering the results from the other tests based on -2*LL*, each test would logically fit the 3PLM model.

Table 6

Model-Fit Indices of Three Traditional IRT Models for Each Data Set

Test	IRT Model	*df*	*-2LL*	$-2LL_{difference}$
October				
Regular	1PLM	336	3563	
	2PML	312	3445	118
	3PLM	288	3446	-1
Honor's	1PLM	336	5251	
	2PLM	312	5203	48
	3PLM	288	5178	25
December				
Regular	1PLM	490	5404	
	2PLM	455	5304	100
	3PLM	420	5251	53
Honor's	1PLM	574	8929	
	2PLM	533	8841	88
	3PLM	492	8819	22
March				
Regular	1PLM	378	4156	
	2PLM	351	4040	116
	3PLM	324	4030	10
Honor's	1PLM	378	5511	
	2PLM	351	5305	206
	3PLM	324	5274	31

Table 7

October Regular ELA Benchmark Test

		CTT		IRT		
Item	Standard	*P*	*ITC*	*a*	*b*	*c*
1	L41	0.713	0.199	**0.388***	**-0.679+**	0.282
2	L41	0.535	0.198	0.408	0.886	0.277
3	RI11	**0.341-**	0.337	0.789	**1.894-**	0.327
4	RI11	0.364	0.385	0.956	1.232	0.235
5	RI10	0.512	0.327	0.875	0.642	0.249
6	RI10	0.457	0.442	0.984	0.782	0.241
7	RI10	0.411	0.258	0.872	1.235	0.247
8	RI18	0.357	0.262	0.929	1.439	0.241
9	RI11	0.403	0.353	1.082	1.099	0.243
10	RL6	0.682	**0.611****	**1.230****	-0.357	0.240
11	RL6	0.512	0.250	0.728	0.746	0.252
12	RL6	0.481	0.432	1.003	0.670	0.242
13	RL1	0.682	0.509	1.094	-0.310	0.245
14	RL1	0.628	0.339	0.742	0.000	0.249
15	RL1	0.589	0.406	0.830	0.220	0.249
16	L41	0.395	**0.024***	0.743	1.892	0.264
17	RI18	0.504	0.376	0.836	0.603	0.243
18	RI18	0.504	0.380	0.862	0.602	0.244
19	RI18	0.364	0.143	0.895	1.699	0.253
20	RI18	**0.744+**	0.449	1.052	-0.587	0.249
21	RI18	0.636	0.393	0.907	0.003	0.251
22	L37B	0.705	0.391	0.856	-0.418	0.249
23	L37a	0.349	0.281	1.042	1.407	0.240
24	L37c	0.519	0.467	1.024	0.451	0.241

Note: Table 7: $n = 129$; Cronbach's $\alpha = .803$

CTT = classical test theory; P=item difficulty index; ITC=Item-Total Correlation. IRT = Item Response Theory; a=item discrimination; b=item difficulty; c=guessing parameter. Bold faced numbers indicate: + easiest item; - hardest item; * worst discriminating item; **best discriminating item.

Table 8

October Honor's ELA Benchmark Test

		CTT		IRT		
Item	Standard	*P*	*ITC*	*a*	*b*	*c*
1	L11	0.659	0.236	0.804	-0.269	0.240
2	L37	0.770	0.206	0.782	-1.203	0.242
3	L41	0.880	0.279	1.037	-2.007	0.241
4	L41	0.908	0.266	1.007	-2.371	0.242
5	L11	0.578	0.144	**0.706***	0.423	0.244
6	L41	0.775	0.188	0.721	-1.309	0.243
7	RI11	0.807	0.118	0.713	-1.634	0.244
8	RI18	0.509	0.187	0.825	0.848	0.240
9	RI11	**0.954+**	0.226	1.066	**-3.031+**	0.242
10	RI10	0.858	0.091	0.737	-2.149	0.244
11	RI10	0.862	0.245	0.966	-1.889	0.241
12	RI18	0.881	**0.354****	**1.130****	-1.925	0.240
13	RI18	0.651	0.243	0.818	-0.236	0.240
14	RI18	0.546	0.276	0.865	0.526	0.239
15	RI18	0.601	0.257	0.821	0.127	0.239
16	RI18	0.601	0.279	0.930	0.117	0.239
17	RI11	0.399	0.131	0.979	1.705	0.242
18	RI10	0.734	0.319	1.067	-0.747	0.241
19	RI10	0.537	0.261	0.945	0.586	0.240
20	RL6	0.665	0.242	0.784	-0.342	0.241
21	RL6	0.271	**0.000***	1.037	3.048	0.233
22	RL1	**0.225-**	0.047	1.065	**3.440-**	0.222
23	RL10	0.284	0.024	1.050	2.810	0.234
24	RL1	0.734	0.153	0.712	-0.957	0.242

Note: Table 8: $n = 218$; Cronbach's $\alpha = 0.616$

CTT = classical test theory; P=item difficulty index; ITC=Item-Total Correlation. IRT = Item Response Theory; a=item discrimination; b=item difficulty; c=guessing parameter. Bold faced numbers indicate: + easiest item; - hardest item; * worst discriminating item; **best discriminating item.

Table 9

December Regular ELA Benchmark Test

		CTT		IRT		
Item	Standard	*P*	*ITC*	*a*	*b*	*c*
1	RI10	0.230	0.042	1.527	2.559	0.205
2	RI10	0.640	0.375	1.325	-0.127	0.207
3	RI13	0.806	0.382	1.472	**-1.031+**	0.208
4	RI13	0.647	0.415	1.443	0.146	0.206
5	RI11	0.518	0.316	1.513	0.501	0.209
6	RI14	0.612	0.364	1.329	0.025	0.208
7	RI14	0.410	0.233	1.254	1.181	0.210
8	RL1	**0.813+**	0.432	**1.829****	-1.030	0.205
9	RL2	**0.173-**	**-0.078***	1.704	**3.142-**	0.198
10	RL2	0.331	0.362	1.677	1.223	0.198
11	RL1	0.374	0.269	1.625	1.193	0.209
12	RL6	0.626	**0.527****	1.816	-0.079	0.204
13	RL3	0.784	0.510	1.817	-0.791	0.207
14	RL4	0.237	0.029	1.506	2.579	0.207
15	RL1	0.655	0.452	1.511	-0.154	0.207
16	RI14	0.493	0.394	1.651	0.744	0.200
17	RI11	0.626	0.309	1.263	-0.018	0.211
18	RI11	0.748	0.487	1.744	-0.610	0.209
19	L39	0.777	0.418	1.449	-0.886	0.207
20	L39	0.576	0.239	1.134	0.275	0.212
21	L40	0.612	0.235	1.139	0.074	0.214
22	L40a	0.561	0.252	**1.060***	0.291	0.210
23	RL3	0.281	0.161	1.789	1.736	0.202
24	L40	0.475	0.208	1.163	0.972	0.215
25	RL3	0.338	0.334	1.724	1.288	0.200
26	RI13	0.252	0.224	1.707	1.809	0.197
27	L39	0.432	0.324	1.299	0.884	0.204
28	L38c	0.626	0.299	1.155	-0.005	0.212
29	L38a	0.619	0.439	1.802	-0.012	0.206
30	RL6	0.338	0.171	1.516	1.562	0.210
31	RL6	0.331	-0.067	1.461	2.168	0.228
32	RL2	0.525	0.436	1.635	0.406	0.206
33	RL3	0.576	0.309	1.295	0.310	0.212
34	RL4	0.446	0.340	1.637	0.876	0.209
35	L38b	0.547	0.459	1.664	0.258	0.203

Note: Table 9: $n = 139$; Cronbach's $\alpha = 0.820$

CTT = classical test theory; P=item difficulty index; ITC=Item-Total Correlation. IRT = Item Response Theory; a=item discrimination; b=item difficulty; c=guessing parameter. Bold faced numbers indicate: + easiest item; - hardest item; * worst discriminating item; **best discriminating item

Table 10

December Honor's ELA Benchmark Test

		CTT		IRT		
Item	Standard	*P*	*ITC*	*a*	*b*	*c*
1	RI10	0.430	0.140	0.354	2.162	0.257
2	RI10	0.807	0.141	**0.335***	-1.872	0.264
3	RI13	0.937	-0.061	0.361	**-3.602+**	0.380
4	RI13	0.874	0.163	0.471	-2.249	0.250
5	RI11	0.834	0.170	0.615	-1.512	0.248
6	RI14	0.794	0.107	0.407	-1.484	0.253
7	RI14	0.583	0.192	0.535	0.384	0.255
8	RL1	**0.987+**	0.285	**0.870****	-3.505	0.250
9	RL6	0.865	0.281	0.567	-1.894	0.249
10	RL2	**0.157-**	-0.087	0.791	**4.000-**	0.209
11	RL2	0.587	0.132	0.460	0.422	0.257
12	RL1	0.570	0.206	0.607	0.454	0.255
13	RL2	0.628	0.238	0.528	0.042	0.252
14	RL6	0.888	0.211	0.574	-2.094	0.251
15	RL6	0.740	0.115	0.427	-0.887	0.255
16	RL3	0.946	0.282	0.691	-2.721	0.249
17	RL4	0.395	0.188	0.565	1.664	0.246
18	RL1	0.883	0.329	0.644	-1.914	0.250
19	RI14	0.691	0.312	0.611	-0.447	0.246
20	RI11	0.821	0.157	0.467	-1.616	0.252
21	RI11	0.933	0.191	0.585	-2.719	0.251
22	L39	0.955	-0.050	0.497	-3.509	0.251
23	L39	0.641	**-0.101***	0.383	-0.064	0.254
24	L40	0.731	0.276	0.579	-0.703	0.250
25	L40a	0.744	0.265	0.541	-0.854	0.249
26	RL3	0.439	0.166	0.531	1.433	0.250
27	L49	0.641	0.111	0.425	-0.036	0.255
28	RL3	0.596	0.260	0.649	0.214	0.251
29	RI13	0.462	0.287	0.694	0.974	0.245
30	L39	0.659	0.098	0.414	-0.197	0.255
31	L38c	0.446	0.120	0.446	-1.557	0.253
32	L38a	0.550	0.180	0.550	-1.935	0.251
33	RL6	0.552	0.249	0.552	1.268	0.244
34	RL6	0.552	0.057	0.552	2.232	0.260
35	RL2	0.571	0.296	0.571	-0.995	0.248
36	RL3	0.551	0.256	0.551	-0.929	0.250
37	RL1	0.713	**0.426****	0.708	-0.567	0.244
38	RL4	0.798	0.366	0.668	-1.150	0.246
39	RL4	0.605	0.047	0.482	0.277	0.258
40	RL4	0.583	0.168	0.534	0.378	0.254
41	L38b	0.825	0.170	0.509	-1.541	0.253

Note: Table 10: $n = 223$; Cronbach's $\alpha = 0.661$

CTT = classical test theory; P=item difficulty index; ITC=Item-Total Correlation. IRT = Item Response Theory; a=item discrimination; b=item difficulty; c=guessing parameter. Bold faced numbers indicate: + easiest item; - hardest item; * worst discriminating item; **best discriminating item

Table 11

March Regular ELA Benchmark Test

		CTT		IRT		
Item	Standard	*P*	*ITC*	*a*	*b*	*c*
1	L37b	0.733	0.286	0.694	-0.604	0.233
2	L37a	0.304	0.191	0.824	1.973	0227
3	L37a	0.430	0.222	0.781	1.195	0.234
4	L37b	0.316	-0.030	0.855	2.276	0.239
5	L42	**0.801+**	0.436	0.960	**-0.921+**	0232
6	L37a	0.426	-0.066	0.603	1.861	0.251
7	RL4	0.529	0.203	0.691	0.635	0.235
8	RL5	0.500	0.160	0.635	0.895	0.237
9	RL5	0.390	0.306	0.901	1.174	0.224
10	RL4	0.743	**0.595****	1.131	-0.620	0.225
11	RL4	0.566	0.325	0.735	0.339	0.231
12	RL5	0.632	0.364	0.823	-0.035	0.230
13	RI16	0.243	0.223	0.948	2.032	0.214
14	RI16	0.596	0.187	0.603	0.236	0.233
15	RI12	0.551	0.299	0.802	0.450	0.233
16	RI12	**0.235-**	0.181	0.985	2.078	0.214
17	RI15	0.265	0.332	0.985	1.677	0.210
18	RI17	0.684	0.418	0.768	-0.330	0.230
19	RI15	0.243	**-0.105***	0.921	**3.124-**	0.231
20	RI17	0.743	0.570	**1.148****	-0.614	0.225
21	RI15	0.662	0.491	1.100	-0.132	0.229
22	RI16	0.360	0.043	**0.455***	1.976	0.242
23	RI12	0.353	0.054	0.695	1.976	0.263
24	RI17	0.632	0.390	0.856	-0.041	0.229
25	L39a	0.500	0.316	0.718	0.712	0.230
26	L42	0.515	0.462	0.958	0.406	0.219
27	L42	0.691	0.381	0.786	-0.348	0.231

Note: Table: 11: $n = 136$; Cronbach's $\alpha = 0.738$

CTT = classical test theory; P=item difficulty index; ITC=Item-Total Correlation. IRT = Item Response Theory; a=item discrimination; b=item difficulty; c=guessing parameter. Bold faced numbers indicate: + easiest item; - hardest item; * worst discriminating item; **best discriminating item.

Table 12

March Honor's ELA Benchmark Test

		CTT		IRT		
Item	Standard	*P*	*ITC*	*a*	*b*	*c*
1	L39	0.750	0.068	0.368	-1.187	0.251
2	L37b	0.335	0.095	0.560	2.423	0.244
3	L39a	0.808	**0.374****	0.617	-1.341	0.246
4	L42	0.369	0.069	0.506	2.304	0.250
5	L42	0.782	0.318	0.607	-1.134	0.247
6	L37a	0.760	0.135	0.383	-1.269	0.250
7	L37b	**0.191-**	**-0.097***	0.664	**3.951-**	0.219
8	L39a	0.809	0.224	0.505	-1.453	0.252
9	RI12	0.964	0.196	0.598	-3.445	0.250
10	RI15	0.889	0.173	0.470	-2.493	0.250
11	RI17	0.302	-0.037	0.573	3.213	0.251
12	RI12	0.902	0.232	0.539	-2.481	0.249
13	RI17	**0.978+**	0.324	**0.708****	**-3.570+**	0.249
14	RI15	0.791	0.095	0.397	-1.536	0.251
15	L42	0.538	-0.085	**0.300***	1.200	0.259
16	RI16	0.920	0.288	0.598	-2.595	0.248
17	RI16	0.227	0.077	0.613	3.312	0.223
18	RI15	0.920	0.112	0.497	-2.867	0.250
19	RI16	0.924	0.276	0.616	-2.614	0.248
20	RI12	0.907	0.145	0.517	-2.588	0.250
21	RI17	0.662	0.197	0.440	-0.332	0.247
22	RL5	0.742	0.082	0.354	1.139	0.251
23	RL4	0.644	0.200	0.451	-0.171	0.248
24	RL4	0.973	0.245	0.667	-3.485	0.250
25	RL5	0.640	0.184	0.492	-0.098	0.250
26	RL5	0.644	0.122	0.391	-0.129	0.252
27	RL4	0.551	0.084	0.374	0.761	0.253

Note: Table 12: $n = 225$; Cronbach's $\alpha = 0.497$

CTT = classical test theory; P=item difficulty index; ITC=Item-Total Correlation. IRT = Item Response Theory; a=item discrimination; b=item difficulty; c=guessing parameter. Bold faced numbers indicate: + easiest item; - hardest item; * worst discriminating item; **best discriminating item.

CHAPTER V: DISCUSSION

According to the Nation's Report Card, over 70% of students entering high school read below the "proficient" level. Of those 70%, approximately one-half do not exhibit even partial mastery of grade-level subjects. Without the necessary literacy skills, students in the United States will be at a disadvantage compared to other industrialized nations. Students without the literacy skills necessary to compete in a global economy will find fewer opportunities for gainful employment and success in college.

In an effort to combat the problem of low literacy rates, many states have adopted the Common Core State Standards in English Language Arts (CCSS ELA). These standards outline what a student should be able to do upon high school graduation. For instance, students are no longer simply required to memorize and recite facts. They are required to read text critically and form opinions based on textual evidence.

Assessing the new CCSS ELA is paramount to student success. In an effort to measure student success, teachers in a large school district in Northern Alabama developed a series of benchmark tests. The purpose of the study was to address teacher concerns about the item quality, reliability, and validity of these teacher-made benchmark tests in English language arts. Teachers in the district are evaluated on how well their students perform on these tests. However, no information has been available to address these issues. Thus, psychometric validation of each item as well as reliability measures of the tests was a meaningful endeavor and should make a significant contribution to the field of test construction as well as providing a solid theoretical foundation for teacher evaluation with regards to student performance on benchmark tests.

Classical test theory (CTT) was used to glean information about the reliability and validity of the tests along with basic item analyses for each item. Then, exploratory factor analysis was conducted to confirm the item structure and to ensure the data were suitable for item response theory (IRT) analysis. If factor analysis revealed a unidimensional structure, IRT was applied to evaluate the strength and weakness of each item. The study addressed the following research questions:

1. Do the items for the honor's courses and the regular courses show a three-factor solution to match the reading for information, reading literature, and language conventions strands as described in the CCSS ELA?
2. Do the tests show strong reliability and validity based on CTT
3. Do the items demonstrate strong item characteristics based on IRT?

Research Question 1

The EFA partially supported a 3-factor solution. However, the October regular class showed a one factor solution with approximately 20% of the variance explained by the first component. The December and March regular courses indicated a 2-factor solution explaining roughly 23% of the variance for each test. A three-factor solution was found for the October and March honor's course with 24.51 % and 22.59% of the variance explained respectively. The December honor's course factor analysis was inconclusive. Therefore, Research Question 1 was partially confirmed.

Research Question 2

Based on the CTT findings, the assessments for the regular courses had the highest reliability. For all tests, the December regular course had the highest reliability score ($\alpha = 0.820$). The honor's courses, however, did not show such high reliability. In fact, the December honor's course had the lowest reliability of all tests ($\alpha = 0.497$). Therefore, Research Question 2 was confirmed for the regular courses but not the honor's courses.

Research Question 3

Based on the IRT findings, most items do show moderate to strong discrimination power**.** A moderate to high discrimination index ranges from 0.65 to greater than 1.70 (Baker, 2001). For all tests combined, 36% of items score low or very low on the discrimination index. For the December regular test, all items were greater than 1.0 indicating a strong discrimination index. The items for the December honor's course showed that all items scored less 0.70 which indicates that most items had moderate to weak discrimination indices.

For parameter *b*, the item difficulty index, items may range from positive to negative infinity. Items with a high *b* parameter are considered the most difficult. Items for all tests combined scored from -3.602 to + 4.000 indicating that items were geared a number of ability levels. Item *c*, the guessing parameter, ranged from 0.197 to 0.380 (item 26, December regular; item 3, December honor's respectively). Since students must choose from four answers, the *c* parameter should average 0.250. As such the *c* parameter are mostly within the acceptable range. Therefore, Research Question 3 was confirmed.

Implications for Practice

English language arts teachers at a large school district in Northern Alabama are evaluated partly on how well students perform on the benchmark assessments aligned with the CCSS. Prior to completion of this project, teachers were given no information on the tests' validity or reliability. The analyses presented here indicate that the tests are valid and test items are generally strong. There is some concern that the honor's course benchmark test scores have a lower reliability score than the regular course benchmark test scores. The test makers may consider removing or rewriting some of the most difficult and the easiest items.

As a whole, the students in the regular English classes do not perform as well as those in the honor's classes. Generally speaking, students in honor's classes are more motivated to achieve (which is why they take honor's courses in the first place). Therefore, the teachers of honor's classes will naturally show higher test scores than those who teach the regular courses. It appears, then, that teachers of honor's course have an unfair advantage unless a method is put in place to equitably compare performance on the regular and honor's courses.

The majority of test questions come from the two reading strands. Therefore, teachers should spend the majority of class time working with students on reading comprehension skills and reading strategies. Little time should be devoted to the language portion since very few test items were taken from this strand. In addition, test-makers should consider removing items which are too difficult ($p < 0.30$) and too easy ($p > 0.80$).

Limitations of the Study and Suggestions for Further Research

There were several limitations to the study. First, students who took the benchmark were aware that their results had no impact on their course grade, so motivation to do well was a concern. In addition, tests involved reading multiple passages and answering multiple questions about both narrative and informational text. The amount of time required may lead to test fatigue which can be a barrier to student performance. While the tests themselves were not timed, students took the test during their regular 53-minute class period. Thus, students may have made random guesses to finish the test before the class period was over.

An additional limitation of the project is the small sample size; access to the data from the entire district would be beneficial. The regular courses are inclusive, but there was no way to tell if a student was on an individualized education plan, a 504 plan, or a non-native English speaker. Race, ethnicity, and socio-economic status would have been useful in order to glean more information about how these factors impact performance on the tests. Additional studies in which this information is available would help strengthen the test items.

If the district continues to use teacher-made benchmark tests, much thought should be given to the testing theory used to validate the test. If possible, a computer adapted test based on IRT would provide the best information about items, students, and ability levels. Then, instruction could be tailored more closely to the individual needs of students**.**

REFERENCES

Abrams, L., McMillan, J. H., & Wetzel, A. P. (2015). Implementing benchmark testing for formative purposes: Teacher voices about what works. *Educational Assessment, Evaluation & Accountability, 27*, 347-375. doi: 10.1007/s11092-015-9214-9.

ACT Incorporated. (2014). *The condition of college and career readiness*. Retrieved from http://www.act.org/content/act/en/research.html

Alabama State Department of Education. (2016). Data Center. Retrieved from www.alsde.edu

Alexander, P. A., & Fox, E. (2011). Adolescents as readers. In M. Kamil, P. D. Pearson, E. B. Moje, & P.P. Afflerbach (Eds.), *Handbook of reading research: Vol. IV*. (pp. 157-156). New York: Routledge.

Allen, M. J., & Yen, W. M. (2001). *Introduction to measurement theory.* Long Grove, IL: Waveland Press.

Alvermann, D. E. (2001). Effective literacy instruction for adolescents. *Journal of Literacy Research, 34*(2), 189-208.

A nation at risk. (1983). National Commission on Excellence in Education. United States Department of Education. Washington, D.C. Retrieved from http://www2.ed.gov/pubs/NatAtRisk/risk.html

An, X., & Yung, Y. F. (n.d.). *Item response theory: What it is and how you can use the IRT procedure to apply it*. SAS Institute. Cary, NC: Retrieved from https://support.sas.com/resources/papers/proceedings14/SAS364-2014.pdf

Baker, F. B. (2001). *The basics of item response theory*. Washington, DC: ERIC Clearinghouse on Assessment and Evaluation.

Bancroft, K. (2010). Implementing the mandate: The limitations of benchmark tests. *Evaluation and Accountability, 22*, 53-72. doi: 10.1007/s11092-010-9091-1

Behuniak, P., & Tucker, C. (1992). The potential of criterion-referenced tests with projected norms. *Applied Measurement in Education*, *5*(4), 337-353. doi: 10.1207/s15324818ame0504_4

Bell, S. M., & McCallum, R. S. (2008). *Handbook of reading assessment.* Boston, MA: Pearson.

Biancarosa, G., & Snow, C. (2004). *Reading next: A vision for action and research in middle and high school literacy.* Washington, DC: Alliance for Excellent Education.

Bice, T. (2016). English language literacy for college and career readiness. Alabama State Department of Education. http://alex.state.al.us/. Accessed November 1, 2016.

Black, P., & Wiliam, D. (1998). Inside the black box: Raising standards through classroom assessment. *Phi Delta Kappan, 80*(2), 139-149.

Bonett, D. G., & Wright, T. A. (2015). Cronbach's alpha reliability: Interval estimation, hypothesis estimation, and sample size planning. *Journal of Organizational Behavior, 36,* 3-15. DOI: 10.1002/job.1960,

Brown, J. D. (1996). *Testing in language programs.* Upper Saddle River, New Jersey: Prentice Hall Regents.

Brown, J. D. (2000). *What is construct validity?* Retrieved August 9, 2016, from http://jalt.org/test/bro_8.htm

Cain, K., & Oakhill, J. (2006). Assessment matters: Issues in the measurement of reading comprehension. *British Journal of Educational Psychology, 76*, 697-708. doi:10.1348/000709905X69807

Common Core State Standards Initiative. (2016). Retrieved from http://www.corestandards.org

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

Eason, S. H., Goldberg, L. F., Young, K. M., Geist, M. C., & Cutting, L. E. (2012). Reader-text interactions: How differential text and question types influence cognitive skills needed for reading comprehension. *Journal of Educational Psychology, 104*(3), 515-528. doi:10.1037/a0027182

Fang, Z., & Schleppegrell, M. J. (2010). Disciplinary literacies across content areas: Supporting secondary reading through functional language analysis. *Journal of Adolescent & Adult Literacy, 53*(7), 587-597.

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

Furr, M. R., & Bacharach, V.R. (2014). *Psychometrics: An introduction*. Los Angeles, CA: Sage.

Gall, M., Gall, J., & Borg, W. R. (2007). *Educational research: An introduction.* Boston, MA: Pearson.

Hambleton, R. K., & Jones, R. W. (1993, Fall). Comparison of classical test theory and item response theory and their applications to test development. *Educational Measurement: Issues and Practice, 12*(3), 38-47.

Hambleton, R. K., Swaminathan, H., & Rogers, H. J. (1991). *Fundamentals of Item Response Theory.* London: Sage Publications.

Hayes, M. (2011). The federal role in confronting the crisis in adolescent literacy. *Education Digest, 76*(8), 10-15.

Heller, R. D. (2016). *The scope of the adolescent literacy crisis.* Retrieved August 18, 2016, from All about adolescent literacy: http://www.adlit.org/adlit_101/scope_of_the_adolescent_literacy_crisis/

Heller, R., & Greenleaf, C. L. (2007). *Literacy instruction in the content areas: Getting to the core of middle and high school improvement.* Washington, DC: Alliance for Excellent Education.

Henderson, S., Petrosino, A., Guckenburg, S., & Hamilton, S. (2007). *Measuring how benchmark assessments affect student achievement* (Issues & Answers Report, REL 2007–No. 039). Washington, DC: U.S. Department of Education, Institute of Education Sciences, National Center for Education Evaluation and Regional Assistance, Regional Educational Laboratory Northeast and Islands. Retrieved from http://ies.ed.gov/ncee/edlabs

Keenan, J. M., & Meenan, C. E. (2014). Test differences in diagnosing reading comprehension deficits. *Journal of Learning Disabilities, 47*, 125–135. doi: 10.1177/0022219412439326.

Kehoe, J. (1995). Basic item analysis for multiple-choice tests. *Research Assessment and Evaluation, 4*(10), 1-3. Retrieved from http://pareonline.net/getvn.asp?v=4&n=10%20

Kibler, A. K., Walqui, A., & Bunch, G. C. (2015). Transformational opportunities: Language and literacy instruction for English language learners in the common core era in the United States. *TESOL, 6*(1), 9-35.

Kirk, C., & Vigeland, L. (2014). A psychometric review of norm-referenced tests to assess phonological error patterns. *Language, Speech, and Hearing in the Schools, 45,* 365-377.

Kirk, C.A. (2000). A response to the adolescent literacy position statement. *Journal of Adult & Adolescent Literacy, 43*(6), 573-575.

Klein, A. (2015). No child left behind: An overview. *Education Week.* Retrieved from: http://www.edweek.org/ew/section/multimedia/no-child-left-behind-overview-definition-summary.html.

Literacy Project Foundation. (2016). *Staggering illiteracy statistics*. Retrieved from http://literacyprojectfoundation.org/community/statistics/

Lord, F. M. (1980). *Applications of item response theory to practical testing problems.* Mahwah, NJ: Routledge.

Marsh, J. A., Pane, J. F., & Hamilton, L. S. (2006). *Making sense of data-driven decision making in education.* Santa Monica, CA: Rand Corporation.

Moore, D. W., Bean, T. W., Birdyshaw, D., & Rycik, J. A. (1999). Adolescent literacy: A position statement. *Journal of Adolescent & Adult Literacy, 43*(1), 97-112.

Nation, K., & Snowling, M. (1997). Assessing reading difficulties: The validity and utility of current measures of reading skill. *British Journal of Educational Psychology, 67*, 359-370.

National Center for Education Statistics. (2015). *The Nation's Report Card: Reading 2009 (NCES 2101-458).* Institute of Education Sciences, U.S. Department of Education, Washington, DC. Retrieved from http://www.nationsreportcard.gov/reading_math_g12_2015

National Center for Public Policy and Higher Education. (2010). *Beyond the rhetoric: Improving college readiness by improving state policy.* Retrieved from http://www.highereducation.org/reports/college_readiness/gap.shtml

Nation's Report Card. (2015). Retrieved from National Center for Educational Statistics: http://www.nationsreportcard.gov/reading_math_g12_2

Nunnally, J. C. (1978). *Psychometric theory*. (2nd ed.) New York: McGraw Hill.

Raykov, T., & Marcoulides, G. A. (2007). *Introduction of psychometric theory.* New York, NY: Routledge.

Reed, D. K. (2015). Middle level teachers' perceptions of interim reading assessment: An exploratory study of data-based decision making. *Research in Middle Education, 38*(6), 1-11.

Shapiro, E., Hilt-Panahon, A., Gischlar, K., Semeniak, K., Leichman, E., & Bowles, S. (2012). An analysis between team decisions and reading assessment data within an RTI model. *Remedial & Special Education, 33*, 335-347. doi: 10.1177/0741932510397763

Torgesen, J. K., & Miller, D. H. (2009). *Assessments to guide adolescent literacy instruction.* Florida Center for Reading Research, Florida State University. Retrieved from centeroninstruction.org

Wise, B. (2009). Adolescent literacy: The cornerstone of student success. *Journal of Adolescent & Adult Literacy, 52*(5), 369-375. doi:10.1598/JAAL.52.5.1

Yang, F. M., & Kao, S. T. (2014). Item response theory for measurement validity. *Shanghai Archives of Psychiatry, 26*(3), 171-177. doi: 10.3969/j.issn.1002-0829.2014.03.010

Yu, C. H. (2013). A simple guide to the item response theory (IRT) and rasch modeling. Retrieved from: http://www.creative-wisdom.com

APPENDICES

APPENDIX A: Standards and Items Assessed

CCSS ELA Language Strand Total Items for Regular and Honor's English

	Standard	October Benchmark		December Benchmark		March Benchmark	
Conventions of Standard English	ID	Reg	Hon	Reg	Hon	Reg	Hon
Demonstrate command of the conventions	L37	1	3				
Use parallel structure.	L37a	1				1	1
Use various types of phrases (noun,	L37b	1				1	2
Demonstrate command of the conventions	L38						
Use a semicolon (and perhaps a conjunctive adverb) to link two or more closely related independent clauses	L38a			1	1		
Use a colon to introduce a list or	L38b			2	1	1	
Spell correctly.	L38c			1	1		
Knowledge of Language							
Apply knowledge of language to understand how language functions in different contexts, to make effective choices for meaning or style, and to comprehend more fully when reading or listening.	L39						1
Write and edit work so that it conforms to the guidelines in a style manual (e.g., *MLA Handbook*, Turabian's *Manual for Writers*) appropriate for the discipline and writing type.	L39a					3	2
Vocabulary Acquisition and Use							
Determine or clarify the meaning of unknown and multiple-meaning words and phrases based on *grades 9-10 reading and content*, choosing flexibly from a range of strategies.	L41			3	2		
Demonstrate understanding of figurative	L41a	3	3				
Acquire and use accurately general academic and domain-specific words and phrases, sufficient for reading, writing, speaking, and listening at the college and career readiness level; demonstrate independence in gathering vocabulary knowledge when considering a word or phrase important to comprehension or expression.	L42					3	3

CCSS ELA Reading Literature Strand Total Items for Regular and Honor's English Class

	Standard	October Benchmark		December Benchmark		March Benchmark	
Key Ideas and Details	ID	Reg	Hon	Reg	Hon	Reg	Hon
Cite strong and thorough textual evidence to support analysis of what the text says explicitly as well as inferences drawn from the text.	RL1	3	3	3	4		
Determine a theme or central idea of a text and analyze in detail its development over the course of the text, including how it emerges and is shaped and refined by specific details; provide an objective summary of the text.	RL2			3	4		
Analyze how complex characters (e.g., those with multiple or conflicting motivations) develop over the course of a text, interact with other characters, and advance the plot or develop the theme.	RL3			5	4		
Craft and Structure							
Determine the meaning of words and phrases as they are used in the text, including figurative and connotative meanings; analyze the cumulative impact of specific word choices on meaning and tone (e.g., how the language evokes a sense of time and place; how it sets a formal or informal tone).	RL4			2	4	3	3
Analyze how an author's choices concerning how to structure a text, order events within it (e.g., parallel plots), and manipulate time (e.g., pacing, flashbacks) create such effects as mystery, tension, or surprise.	RL5					3	3
Analyze a particular point of view or cultural experience reflected in a work of literature from outside the United States, drawing on a wide reading of world literature.	RL6	3	3	3	5		

CCSS ELA Reading Informational Text Strand Total Items for Regular and Honor's English Course

	Standard	October Benchmark		December Benchmark		March Benchmark	
Key Ideas and Details	ID	Reg	Hon	Reg	Hon	Reg	Hon
Cite strong and thorough textual evidence to support analysis of what the text says explicitly as well as inferences drawn from the text.	RI-10	3	3	2	2		
Determine a central idea of a text and analyze its development over the course of the text, including how it emerges and is shaped and refined by specific details; provide an objective summary of the text.	RI-11	3	3	3	3		
Analyze how the author unfolds an analysis or series of ideas or events, including the order in which the points are made, how they are introduced and developed, and the connections that are drawn between them.	RI-12					3	3
Craft and Structure							
Determine the meaning of words and phrases as they are used in a text, including figurative, connotative, and technical meanings; analyze the cumulative impact of specific word choices on meaning and tone (e.g., how the language of a court opinion differs from that of a newspaper).	RI-13			2	3		
Analyze in detail how an author's ideas or claims are developed and refined by particular sentences, paragraphs, or larger portions of a text (e.g., a section or chapter).	RI-14			3	3		
Determine an author's point of view or purpose in a text and analyze how an author uses rhetoric to advance that point of view or purpose.	RI-15					3	3
Integration of Knowledge and Ideas							
Analyze various accounts of a subject told in different mediums (e.g., a person's life story in both print and multimedia), determining which details are emphasized in each account.	RI-16					3	3
Delineate and evaluate the argument and specific claims in a text, assessing whether the reasoning is valid and the evidence is relevant and sufficient; identify false statements and fallacious reasoning.	RI-17					3	3
Analyze seminal U.S. documents of historical and literary significance (e.g., Washington's Farewell Address, the Gettysburg Address, Roosevelt's Four Freedoms speech, King's "Letter from Birmingham Jail"), including how they address related themes and concepts.	RI-18	6	6				

APPENDIX B: IRB Letter

IRB
INSTITUTIONAL REVIEW BOARD
Office of Research Compliance,
010A Sam Ingram Building,
2269 Middle Tennessee Blvd
Murfreesboro, TN 37129

MIDDLE TENNESSEE STATE UNIVERSITY

IRBN007 – EXEMPTION DETERMINATION NOTICE

Friday, January 06, 2017

Investigator(s): Andrea D. Milligan (Student PI) and Jwa Kim (FA)
Investigator(s') Email(s): *adm5y@mtmail.mtsu.edu; jwa.kim@mtsu.edu*
Department: Literacy/CoED

Study Title: A psychometric analysis of teacher-made bench mark test in English language arts
Protocol ID: **17-1116**

Dear Investigator(s),

The above identified research proposal has been reviewed by the MTSU Institutional Review Board (IRB) through the **EXEMPT** review mechanism under 45 CFR 46.101(b)(2) within the research category *(4) Study involving existing data*. A summary of the IRB action and other particulars in regard to this protocol application is tabulated as shown below:

IRB Action	EXEMPT from further IRB review***	
Date of expiration	NOT APPLICABLE	
Participant Size	Not Applicable	
Participant Pool	**Not Applicable - Existing data collected from academic records**	
Mandatory Restrictions	No active data collection can be performed	
Additional Restrictions	NONE	
Comments	NONE	
Amendments	**Date**	**Post-Approval Amendments**
		NONE

***This exemption determination only allows above defined protocol from further IRB review such as continuing review. However, the following post-approval requirements still apply:

- Addition/removal of subject population should not be implemented without IRB approval
- Change in investigators must be notified and approved
- Modifications to procedures must be clearly articulated in an addendum request and the proposed changes must not be incorporated without an approval
- Be advised that the proposed change must comply within the requirements for exemption
- Changes to the research location must be approved – appropriate permission letter(s) from external institutions must accompany the addendum request form
- Changes to funding source must be notified via email (irb_submissions@mtsu.edu)
- The exemption does not expire as long as the protocol is in good standing
- Project completion must be reported via email (irb_submissions@mtsu.edu)

- Research-related injuries to the participants and other events must be reported within 48 hours of such events to compliance@mtsu.edu

The current MTSU IRB policies allow the investigators to make the following types of changes to this protocol without the need to report to the Office of Compliance, as long as the proposed changes do not result in the cancellation of the protocols eligibility for exemption:

- Editorial and minor administrative revisions to the consent form or other study documents
- Increasing/decreasing the participant size

The investigator(s) indicated in this notification should read and abide by all applicable post-approval conditions imposed with this approval. Refer to the post-approval guidelines posted in the MTSU IRB's website. Any unanticipated harms to participants or adverse events must be reported to the Office of Compliance at (615) 494-8918 within 48 hours of the incident.

All of the research-related records, which include signed consent forms, current & past investigator information, training certificates, survey instruments and other documents related to the study, must be retained by the PI or the faculty advisor (if the PI is a student) at the sacure location mentioned in the protocol application. The data storage must be maintained for at least three (3) years after study completion. Subsequently, the researcher may destroy the data in a manner that maintains confidentiality and anonymity. IRB reserves the right to modify, change or cancel the terms of this letter without prior notice. Be advised that IRB also reserves the right to inspect or audit your records if needed.

Sincerely,

Institutional Review Board
Middle Tennessee State University

Quick Links:

Click here for a detailed list of the post-approval responsibilities.
More information on exmpt procedures can be found here.