COMMON CORE STATE STANDARDS BENCHMARK ASSESSMENTS:

ITEM ALIGNMENT TO THE SHIFTS IN TENNESSEE

by

Melissa Stugart

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
March, 2016

Dissertation Committee:

Dr. Jwa K. Kim, Chair

Dr. Amy M. Elleman

Dr. James Herman

ACKNOWLEDGMENTS

I would like to express my deepest gratitude to my advisor, Dr. Jwa K. Kim, for his incredible guidance, care, encouragement, and patience. Without his persistent help, this dissertation would not have been possible. I would also like to thank my committee members, Dr. Amy Elleman and Dr. James Herman, for their suggestions and advice throughout the dissertation process. I could not have finished my dissertation without the dedicated guidance of my committee members.

I would like to thank my husband, Reuben, and my mother, Susan, for never giving up on me and providing eternal support.

The more you read, the more you know. The more you learn, the more places you will go.

~ Dr. Seuss

ABSTRACT

Our nation is in the midst of one of the largest education reforms in decades centered on the adoption of the Common Core State Standards (CCSS) and aligned assessments. In an era of rising accountability measures and declining literacy proficiency, it is vital to ensure that educational resources, such as benchmark assessments, are appropriately aligned to state education reform movements. The purpose of this study was to use exploratory factor analysis (EFA), classical test theory (CTT), and item response theory (IRT) to consider if factors aligned to the three instructional shifts of CCSS can be confirmed within benchmark assessments designed to measure student progress across three grade levels: 4^{th}, 8^{th}, and 10^{th}. Data samples were specific to a test administered to Tennessee students during the fall and winter of the 2014-2015 school year. The researcher hypothesized it would be more likely that the benchmark items would align more strongly with a four-factor solution because the tests were designed to assess four strands of the CCSS (Language, Reading Informational Text, Reading Literature, and Writing). However, EFA revealed a stronger alignment with a three-factor solution after removal of misfit items using CTT and IRT. Overall, the results were inconclusive and additional study is required to determine if benchmark assessments are being designed to assess not just the CCSS, but the theoretical underpinnings of the standards. Benchmark assessments must align with the CCSS in order to provide the best possible information to aid both student learning and teacher development.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER I: INTRODUCTION

Our nation is in the midst of the largest education reforms in decades centered on the adoption of the Common Core State Standards (CCSS) and aligned assessments. The CCSS were developed through a series of initiatives aimed at the support of the movement towards standards-based instruction at the heart of our country's education reform for the past 25 years. The CCSS represent an attempt to address the lowering of academic content standards that occurred in many states during the No Child Left Behind (NCLB) era (Darling-Hammond, 2007; Liebtag, 2013; Conley & Gaston, 2013). Compared to most existing state standards, the CCSS are considered more complex and more demanding (Conley & Gaston, 2013). Also, unlike many former state educational standards that progress forward, like those in Tennessee and Utah, the CCSS were back-mapped from college- and career-readiness anchor standards at the 12th-grade level through learning progressions that extend to the third grade (Tennessee Department of Education, 2013; Utah State Office of Education, 2010; CCSSI, 2014). The standards focus on a sound foundation of key content knowledge, requiring that students not only obtain a rote command of operations and techniques but deep understanding of the concepts involved.

Forty-four states, the District of Columbia, and four territories originally adopted the CCSS. Although most participating states adopted the standards between 2010 and 2011, over the past few years the reform effort has seen political pushback from various constituencies across the country (Cristol & Ramsey, 2014). During both the 2014 and 2015 Tennessee legislative sessions, dozens of bills related to the repeal or delay of the

CCSS and aligned assessments were proposed. While none of these measures successfully repealed the use of the CCSS, the legislature did opt to pull out of the Common Core-aligned test consortia known as the Partnership for Assessment of Readiness for College and Career (PARCC) in favor of releasing a Request for Proposal (RFP) for a new test vendor (Wagner, 2014).

Anti-CCSS legislative action is not unique to Tennessee; nearly every Common Core state legislature saw bills of this nature over the past few years. Jochim and Lavery (2015) coded 11,785 CCSS-related bills introduced in state legislatures between 2011 and September 2014 and identified 238 bills that were negative in tone. Indiana, a state that originally adopted the CCSS, even withdrew from the Common Core State Standards Initiative (CCSSI) in April 2014, opting instead to develop their own standards (Ballentine, 2014).

While many opponents have voiced concerns about student data and federal infringement upon state education rights, others have long questioned the amount of time teachers have spent on test prep since the implementation of No Child Left Behind (2001). The Tennessee Educator Survey Report (2015) reveals that the majority of Tennessee teachers say they have spent too much time on test prep. For the second year in a row, over 60 percent of all teachers believe that they have spent too much time "helping students prepare for statewide assessments." Additionally, almost half of teachers say they spend more than 20 class periods preparing for state assessments through activities like practice tests. In fact, 61 percent of teachers report using educational technology to diagnose student learning needs at least once a week. As

teachers spend increasing amounts using technology to administer diagnostic benchmark assessments, it is more critical than ever to ensure resources are aligned with the CCSS.

The Common Core State Standards for English/language arts (ELA) and literacy are of particular concern in a state with severely low rates of proficiency in reading. The 2013 National Assessment of Educational Progress (NAEP) results show that 66% and 67% of Tennessee 4th and 8th graders, respectively, are reading below a proficient level. In 2015, Tennessee students made little progress on the NAEP, maintaining 66% and 67% proficiency levels for 4th and 8th grade, respectively (National Center for Education Statistics, 2015). As teachers across the state begin to implement more rigorous practices, it is important that ELA teachers are fully prepared with resources, like benchmark assessments, designed to effectively teach and measure the shifts expected with the new standards.

The ELA CCSS were designed around three key shifts, or expectations: 1) regular practice with complex text and its academic language; 2) reading, writing, and speaking grounded in evidence from text, both literary and informational; and 3) building knowledge through content-rich nonfiction. Each of these shifts is critical to successful implementation of CCSS and described more fully in the next chapter.

Benchmark assessments are designed to measure the achievement of standards. The fundamental purpose of benchmark assessment is to provide information that teachers can use to guide instruction over time as a set of check-ins over the course of the year to ensure students are ready for summative, or final, assessments (Bergan, Bergan, & Burnham, 2009, p. 2). Many districts choose to purchase sets of assessments from for-

profit companies to compare progress over the year and be sure students will be able to pass end of year exams, such as the assessment we will analyze in this study.

Purpose of the Study

The purpose of this study is to use exploratory factor analysis (EFA), classical test theory (CTT), and item response theory (IRT) to consider if factor structure of items aligned to the three instructional shifts of the ELA CCSS can be confirmed within benchmark assessments designed to measure student progress across three grade levels: 4^{th}, 8^{th}, and 10^{th}. Benchmark assessments must align with the CCSS to provide the best possible information to aid both student learning and teacher development.

Research Questions

1. Do the assessments have sound psychometric properties, such as validity and reliability?
2. Do the assessments show three- or four-factor solutions through factor analysis?
3. Are the benchmark assessments used by TN schools/students aligned to the shifts or strands of the ELA CCSS?

Hypotheses

Hypothesis 1: Because the assessments were developed by a large test company and administered to a large sample size, the researcher hypothesizes that the assessments will have sound psychometric properties.

Hypothesis 2: Because the assessments have been developed to measure the ELA CCSS which are based on three key instructional shifts, the researcher hypothesizes that the assessments will show a three-factor solution.

Hypothesis 3: Because of the anticipated three-factor solution, the researcher anticipates that the factors will align to the shifts of CCSS.

Significance of the Study

Change has been difficult in the era of accountability of No Child Left Behind and adoption of CCSS. Because so many districts are now utilizing purchased benchmark assessments to measure student progress towards mastery on end-of-course assessments, it is imperative that we ensure the tests fully align to the intention of the CCSS.

Assumptions, Delimitations, and Limitations

An assumption exists that the responses of the participants will be an accurate and honest reflection of their knowledge, especially because the data to be evaluated was collected during the first half of a school year. The researcher will only examine the data from the first and second of a series of benchmark assessments administered during the 2014-2015 school year to students across Tennessee.

It must also be noted that participant responses to items will likely show a strong degree of correlation based on the overlapping nature of the ELA CCSS. The researcher was not provided access to the language of the items, only the strand and standard the items were designed to assess.

Finally, there are also CCSS for Math and Literacy in the Content Areas. The researcher has identified ELA instruction as the focus for this study to target understanding of benchmark assessment alignment in a particular area.

Definition of Key Terms

Informational Text – Text with the primary purpose of expressing information about the arts, sciences, or social studies. This text ranges from newspaper and magazine articles to digital information to nonfiction trade books to textbooks and reference materials (Young, 2012).

Text Complexity – The inherent difficulty of reading and comprehending a text combined with consideration of reader and task variables; in the Standards, a three-part assessment of text difficulty that pairs qualitative and quantitative measures with reader-task considerations (CCSS, pp. 31, 57; Reading, pp. 4–16).

Text-dependent Question – Questions that require students to read carefully and produce evidence in their verbal and written responses (Fisher and Frey, 2012).

CHAPTER II: REVIEW OF THE LITERATURE

In Tennessee, teachers have been asked to learn new education standards twice in the past five years. In 2007, after receiving an "F" for "Truth in Advertising" about student proficiency from the U.S. Chamber of Commerce (Tennessee Department of Education, 2013), Tennessee developed a new set of standards in alignment with the American Diploma Project. These new standards were first implemented during the 2009-2010 school year. Now, Tennessee has adopted the CCSS for the 2014-2015 school year. The challenge of assessing student progress in the constantly changing landscape of education reform is of concern to both proponents and opponents of the national standards movement. Those in favor of the new standards are determined to ensure students are prepared for the more rigorous expectations of the new standards, while those opposed voice concerns that constant change makes it impossible to measure student progress.

To best understand the issue of student assessment within new reform efforts, this literature review will examine the development of the CCSS in Tennessee. It is of equal importance to examine the key instructional shifts for ELA around which the CCSS were designed, as well as investigate the benefits of benchmark assessments as resources in today's classrooms.

History of Common Core Adoption and Education Reform in Tennessee

Since No Child Left Behind (2001), all states have been required to adopt and monitor student success on education standards, or learning expectations, for each course

of study. In Tennessee, the State Board of Education is the governing and policy making body that regulates the state standards in all subjects for elementary and secondary instruction. The board is composed of nine members who are appointed by the governor for a five-year term.

The idea for the CCSS was developed over a series of initiatives aimed at the support of standards-based education reform across states. In 2004, the American Diploma Project released a report entitled, "Ready or Not: Creating a High School Diploma that Counts," that demonstrated a lack of college and career readiness upon graduation from most high schools, which started a national movement towards aligning education standards with postsecondary and workforce expectations (Achieve, 2004). Then, in 2007, Tennessee created the Tennessee Diploma Project to align our state's education standards with others as part of the American Diploma Project Network. This work was led by the Tennessee Alignment Committee, which consisted of a panel of state and local government officials, business, higher education, and K-12 leaders (Tennessee Department of Education, 2013).

The CCSS initiative was officially launched by the National Governors Association (NGA) and Council of Chief State School Officers (CCSSO) in 2008. The standards were developed and validated with a focus on alignment with expectations for college and career success; clarity; consistency across all states; inclusion of content and the application of knowledge through higher-order skills; and improvement upon current state standards and demands of top-performing nations. In June 2009, Tennessee Governor Phil Bredesen and Education Commissioner Tim Webb joined the CCSSI,

along with 47 other states (Rothman, 2011). Then, in January 2010, the state applied for the federal Race to the Top grant, in which the Tennessee Department of Education (TDOE) promised to adopt CCSS if awarded funding (U.S. Department of Education, 2010). After receiving the award in July 2010, the Tennessee State Board of Education unanimously voted to adopt the standards with implementation to begin in the 2014-2015 school year, despite having already revised the Tennessee academic standards as part of the American Diploma Project in 2007.

Two major test consortia were also developed out of federal Race to the Top funding in order to provide appropriately aligned assessments for the CCSS: the Partnership for Assessment of Readiness for College and Careers (PARCC) and Smarter Balanced. Initially, each state that adopted the CCSS chose which consortium to join. The Tennessee State Board of Education originally agreed to be a PARCC governing state, which meant that state representatives voted on all PARCC decisions and participated in the 2013-2014 pilot. However, during the 2014 legislative session, state lawmakers decided to delay implementation of the PARCC assessment in order to consider other options (Tennessee General Assembly, 2014). During the 2015 Tennessee legislative season, more bills were put forward to repeal the CCSS, ultimately resulting in the formal adoption of Governor Haslam's standards review process to conclude by the end of 2015 (Tennessee Government, 2014). In addition to codifying a formal review process, HB1035 required the Tennessee State Board of Education (TSBOE) to cancel any memoranda of understanding concerning the CCSS causing a return to calling the standards the Tennessee State Standards (Tennessee General Assembly, 2015).

Currently, the review process has ended and committee recommendations were presented to the TSBOE for a first reading in January, 2016 with a final reading scheduled for April, 2016 (Tennessee Government, 2016). It is important to note that despite the changing political winds around standard adoption, the CCSS-aligned benchmark assessment data analyzed in this study was collected during the 2014-2015 school year.

While Tennessee has demonstrated commitment to successfully implementing the CCSS, Cohen and Bhatt (2012) warn about the importance of considering infrastructure development when introducing any new reform movement aimed at improving literacy. They ask whether a reform restricted to standards and assessments can deeply change schools, while suggesting that in order to successfully put the standards into practice, states will need to build an educational infrastructure consisting of examinations, curricular frameworks, teacher education, inspection systems for improving instruction, and an academically successful teaching force.

For the recent CCSS reform movement, the TDOE utilized $15 million of the $500 million Race to the Top grant to provide state-wide professional development through an effort called TNCore for teachers preparing to implement the CCSS (TNCore, n.d.). Using a train-the-trainer approach to professional development, the TDOE recruited and trained regionally-based "Core Coaches" to provide extensive professional development across the state. During the summer of 2012, over 11,000 math instructors in grades 3-8 received 3 days of training in math CCSS. The following summer, 29,146 Tennessee teachers received between 2-4 days of training in K-8 math, high school math, K-12 English/language arts, and 6-12 literacy (science, social students, and career and

technical education). In addition to training teachers, TNCore developed a series of leadership courses in CCSS for administrators. Between the spring of 2013 and 2014, over 5,700 Tennessee administrators were trained in CCSS implementation (N. Roberts, personal communication, April 28, 2014).

While the TDOE did provide professional development to teachers around the shifts of CCSS, they did not provide mandated benchmark assessments to districts. There is no data to indicate how many districts in Tennessee have purchased or developed their own benchmark tests. Thus, it is necessary to evaluate the factor structure of test items used in Tennessee schools to ensure that assessments are effectively measuring what they intend to measure.

Overview of the English/Language Arts Common Core State Standards

Because the ELA CCSS are multifaceted, using them is a complex task. The Common Core State Standards are founded on the College and Career Readiness (CCR) standards. As noted in the CCSS for English Language Arts, "The CCR standards anchor the document and define general, cross-disciplinary literacy expectations that must be met for students to be prepared to enter college and workforce training programs ready to succeed" (NGA & CCSSO, 2010, p. 4). The CCR Anchor Standards provide a foundation for the CCSS and specify what students should know and be able to do to by the end of 12th grade to succeed in college and the workplace. The CCR Standards are broad while the CCSS represent more specific benchmarks that buttress each anchor standard. The grade-specific CCSS connect to the College and Career Readiness Standards as benchmarks of what students in each grade level, K–12, should know and be

able to do to reach the College and Career Readiness Standards by the time the students graduate from high school.

The ELA CCSS are structured into six strands: Reading Foundational Skills, Reading Literature, Reading Informational Text, Writing, Speaking and Listening, and Language. With the exception of the Reading Foundational Skills, each of these strands includes standards that have been back-mapped from the CCR anchor standards and show a clear alignment between grade levels. These standards are further divided within each strand into categories that group them according to similarity. For example, within the Grade 6 Reading Literature strand, the category Key Ideas and Details contains three standards related to citing textual evidence, determining a central theme, and describing plot development. Another category in the same strand, Craft and Structure, includes three standards related to analyzing how an author uses language to develop more nuanced meaning in the text. The ELA CCSS also include three lengthy appendices. Appendix A provides research supporting the key elements of the standards and a glossary of terms. Appendix B provides text exemplars and performance tasks. Appendix C provides samples of student writing.

According to the ELA CCSS, students are expected to identify key ideas and details in complex literary and informational texts, integrate knowledge and ideas, and read a range of texts of varying complexity. They need to be able to write in multiple genres and use research in their writing to support their points of view; demonstrate a mastery of the conventions of the English language; have a vocabulary sufficient to express themselves according to standard scholarly conventions; understand what is said

to them, to engage in effective and fruitful conversation with others; and present their ideas orally in ways that a wide range of audiences can understand. With such demanding expectations, it is easy to see why the majority of teachers in Tennessee predict the CCSS will improve their students' abilities, despite recognizing the incredible challenge it represents for students and teachers alike (Pepper, Burns, Kelly & Warach, 2013).

If all students are to be ready for college and/or career by the end of high school, teachers need to go beyond teaching skills and also consider the texts to which students apply their skills. The ELA CCSS emphasize the nature, complexity, and rigor of the literary and informational texts that students should read at each grade level. The previous Tennessee state standards stressed the skills and strategies students use as they read but did not stipulate what students should actually read (Tennessee Department of Education, 2009a-l). The CCSS put emphasis on the texts themselves, specifying readability levels and the proportions of classroom time to be devoted to both informational and literary texts.

The ELA CCSS are founded upon three key instructional shifts for teachers to consider when implementing the standards: 1) regular practice with complex text and its academic language; 2) reading, writing, and speaking grounded in evidence from text, both literary and informational; and 3) building knowledge through content-rich nonfiction. Each of these shifts is described more fully below as they will be utilized as a foundation for analyzing the alignment of benchmark assessments to the CCSS within this study.

Shift One: Regular Practice with Complex Text and Its Academic Language

The ELA CCSS are designed to highlight the growing complexity of texts students must read to be ready for the demands of college and career. Closely related to text complexity, the standards also emphasize academic vocabulary, calling for students to grow their vocabularies through a combination of speaking and listening, reading, and writing to texts. They ask students to determine word meanings, appreciate the nuances of words, and steadily expand their range of words and phrases.

The 2006 ACT, Inc. report entitled, “Reading Between the Lines," highlighted the major distinctions between students who met the benchmark for college readiness (21 out of 36) and those who did not. Unpredictably, ACT, Inc. did not find the strongest differentiating factor to be critical thinking or higher order reading skills. Instead, they found the most significant factor was the complexity of the texts. Students scoring below benchmark from all population subgroups performed no better than chance on passages considered "complex." This finding underscores the importance of teaching students with grade appropriate materials to best prepare students for college and career.

Research has indicated that text complexity in high school has been steadily declining. Hayes and Ward (1992) found that only Advanced Placement high school courses met the equivalent of newspapers at the time while more recently Williamson (2006) found there to be a 350L (Lexile) gap between the end of high school and college level reading. This gap represents a 1.5 standard deviation. Milewski, Johnson, Glazer and Kubota (2005) similarly found that college professors are assigning more challenging texts than high school teachers while Stenner, Koons, and Swartz (2009) found that even

workplace reading exceeds the current 12th grade reading complexity. Beyond this current gap in text complexity, college professors also expect their students to read more independently (Pritchard, Wilson, & Yamnitz, 2007) while high school teachers were found to rarely hold students accountable for independent reading (Heller & Greenleaf, 2007).

Additional research indicates that teachers have not been selecting reading assignments with text complexity in mind. Shanahan and Duffett (2013) conducted a survey with 1,154 public school instructors of English/language arts and reading from each grade band in 46 states during the spring of 2012 to determine how teachers were selecting and utilizing texts in their classrooms. Teachers were first asked how they selected texts. Responses showed that while 64% of elementary teachers made a substantial effort to match students with books presumed to align with their instructional reading levels, only 40% of middle school teachers and 25% of high school teachers selected texts this way. Additionally, 77% of all teachers reported that they assigned novels that all students were expected to read and that roughly 33% of their students were reading below grade level. The researchers pointed out that, "… in these classrooms, when a single text (such as a novel) is used with all students, they may be asked to read easier-than-grade-level texts, no matter what their individual reading proficiencies, since teachers would aim for a classroom average reading level" (p. 21).

The first foundational shift of ELA CCSS seeks to address the concerns outlined above. There exist vast gaps in reading level expectations between K-12, college, and career that must be addressed for students to be successful beyond high school.

Additionally, teachers need guidance on how to be more intentional in text selection to differentiate instruction for students and progress all students to proficiency.

While the ELA CCSS identify several measures through which teachers can determine the appropriate text complexity for students, they do not prescribe texts. Appendix B highlights exemplar texts and suggested readings for each grade band, but it is expected that teachers will utilize the text complexity measures outlined in Appendix A to make their own curricular choices. The appendix outlines a three-part model for measuring text complexity: qualitative, quantitative, and reading and task considerations.

The qualitative dimension refers to aspects of text complexity that are only measurable by a human reader. Specifically, the document includes four qualitative factors to consider when measuring text complexity: levels of meaning (literary) and purpose (informational), structure, language conventionality and clarity, and knowledge demands. Texts with a single level of meaning or explicit purpose are considered to be less complex than texts with multiple meanings or implicit purposes. Texts with well-marked, conventional structures are considered to be less complex than those with implicit, unconventional structures. Texts that rely on literal language conventions are considered to be less complex than those that employ figurative or ambiguous language conventions. Finally, texts that make few assumptions about students' cultural and content knowledge are considered less complex than those that make many assumptions.

The quantitative dimension refers to the aspects of text complexity, such as word and sentence length that are too challenging to be evaluated by a human reader. This dimension is typically measured by computer software. Many formulas exist to assist

teachers to identity the complexity of text, such as the Flesch-Kincaid Grade Level test, which uses word and sentence length. Another is the Dale-Chall Readability Formula, which utilizes word frequency as the key factor in determining text complexity. Another commonly used formula is the Lexile Framework for Reading, developed by MetaMetrics, Inc., which uses both word frequency and sentence length to determine an exact Lexile measurement of a texts complexity.

The final dimension, reader and task considerations, refers to the professional judgment of text appropriateness while considering both qualitative and quantitative factors. Teachers are recommended to consider their students' interests and abilities in making text selections to effectively differentiate and tailor instruction.

Academic vocabulary instruction is also a large part of this key shift of the ELA CCSS. Vocabulary has been widely and inextricably linked to reading comprehension (National Institute of Child Health and Human Development, 2000), and research has historically demonstrated that disparities in vocabulary knowledge lead to disparities in academic achievement (Baumann & Kameenui, 1991; Becker, 1977; Stanovich, 1986). Therefore, one cannot consider the significance of increasing the complexity of text without also considering the implications of vocabulary instruction.

Appendix A of the standards specifically addresses recommendations for vocabulary instruction and acquisition. The ELA CCSS emphasize that vocabulary should not be taught in isolation, but rather through repeated, incremental exposure to words in a variety of contents for students to make multiple connections and build meaning (p. 32). The standards also place heavy emphasis on speaking and listening

skills, recognizing that students must first acquire vocabulary through oral language development.

The ELA CCSS borrow the three-tier model of conceptualizing categories of words from Beck, McKeown, and Kucan (2002, 2008). Tier One words are the words of everyday speech generally learned in the early grades. These words occur in high frequency in the English language and are not considered to be challenging to the average native speaker. Tier Two words, or academic language, appear more frequently in the written language than speech and are therefore more challenging to acquire. Tier Three words, or domain-specific words, are specific to a domain or field of study. These words are more common in informational text than literary text.

Appendix A does not make many specific recommendations for the instruction of vocabulary, but it suggests that Tier Two and Tier Three words need to be taught explicitly in text. Specifically, the document says, "teachers need to be alert to the presence of Tier Two words and determine which ones need careful consideration" (p. 33). Additionally, Tier Three vocabulary development occurs most effectively through an integrated and coordinated curriculum across disciplines.

Despite the research-based rationale for increasing text complexity and improving vocabulary, there has been little research to indicate that the ELA CCSS have had an impact on student outcomes in reading proficiency. This is due in large part to the fact that implementation has only just begun and it is difficult to study the impact of an intervention before the intervention has been fine tuned. Wixson and Valencia (2014) warn that in implementing the three-part model of text complexity teachers may not fully

connect texts to tasks appropriately. They caution that it is important not to merely increase the complexity of the text, but consider how it relates to the difficulty of the associated task to ensure students do not get frustrated.

Furthermore, it is worth noting that not all researchers and practitioners agree with the shift towards increased text complexity. Fang and Pace (2013) state that an overemphasis on grade bands and reader-text match may inhibit teachers from making student-centered reading selections. Additionally, they find the descriptions of the four qualitative measures of text complexity to be too vague to be helpful to teachers.

Regardless, the Tennessee Department of Education has developed a series of resources related to CCSS implementation on their TNCore website, with a page specifically devoted to text complexity. The page is home to an "online learning series" with modules that provide guidance around selecting texts for grade bands 2-5, 6-8, and 9-12. Each module includes resource materials, a video, and an assessment for teachers to ensure understanding. Additionally, the site provides an eight page training module with resources and activities related to assessing text complexity and developing academic language for ELA 3-12. Some of these resources include a text complexity analysis worksheet, rubrics for assessing qualitative measures for both literary and informational texts, and a reader and task considerations guide.

Overall, the shift towards increasing student expectations to read complex text and receive explicit instruction in academic language represent a reversal from the recent trend towards less complex text and limited vocabulary instruction past the middle grades. In considering how to assess students, test-designers will need to be knowledgeable about best practices in selecting complex text.

Shift Two: Reading, Writing, and Speaking Grounded in Evidence from Text

Beyond the shift towards increasingly complex text, students and teachers are now moving towards using evidence-based reading, writing, and speaking. The ELA CCSS emphasizes using evidence from texts to deliver clear summaries, well-considered analyses, and strong claims. Historically, forms of writing in K–12 have drawn heavily from student experience and opinion, which alone will not prepare students for the demands of college and career. For example, in Tennessee, the previous writing assessments given in the 3rd, 8th, and 11th grade years required students to write a persuasive essay on a topic given without any reference texts. The expectation was that students write only about their own experiences as evidence to support their claims, which does not align with interdisciplinary college or workplace writing expectations.

Research shows that schools are not spending enough time on writing instruction. When Applebee and Langer (2006) evaluated the 2005 National Assessment of Educational Progress NAEP scores and surveys, they found 48% of students reported spending 11-40% of class time on writing instruction, with 11% spending even less (p. 5). They also found, unsurprisingly, that scaled scores for higher-order writing tasks such as analysis, summary, and research are directly correlated to time spent on instruction. The more writing instruction students received, the more likely they were to perform well on the writing tasks.

More critically, research shows that writing improves understanding of a text. In their landmark meta-analysis study, *Writing to Read,* Steve Graham and Michael Hebert (2010) state, "Writing about a text proved to be better than just reading it, reading and

rereading it, reading and studying it, reading and discussing it, and receiving reading instruction" (p. 12). They additionally outline the many ways that reading and writing are linked both cognitively and functionally. From their extensive research, they have identified three recommendations for improving reading through writing: 1) have students write about what they read, 2) teach students writing skills, and 3) increase how much students write. All three of these recommendations correlate to the ELA CCSS shift towards evidence-based writing.

The ELA CCSS require students to answer questions that depend on reading texts carefully and using the information to form their answers. The instructional focus for teachers is to begin providing students with "text-dependent" questions. Fisher and Frey (2012) have defined text-dependent questions as: "effective questions about literature and nonfiction texts [that] require students to delve into a text to find answers" (p. 70). Rather than asking students questions about their own experiences or how they relate to the text, teachers are now expected to ask questions that require students to refer to what they've read to answer. For any given grade level, the ELA CCSS include standards related to utilizing evidence every strand except language. The standards set the expectation that in reading literature and informational text, students will be able to identify and utilize evidence to support their claims while both speaking and writing.

Teachers must also consider the types of text-dependent questions, which Fisher and Frey (2012) identify in a progression of difficulty as general understanding; key details; vocabulary and text structure; purpose; inferences; and opinions, arguments, and

intertextual connections. These types of questions promote critical thinking and reading skills when used by teachers and/or students.

Appendix A of the ELA CCSS outlines three types of writing students are expected to perform: argument, informational/expository, and narrative. The standards put special emphasis on argument, which is defined as a "reasoned, logical way of demonstrating that a writer's position, belief, or conclusion is valid" (p. 23). Being able to write sound arguments is critical to college and career readiness because being able to argue effectively forces a writer to evaluate the strengths and weaknesses of multiple perspectives. The 2009 ACT National Curriculum Survey of postsecondary instructors supports this supposition. The report found that writing to argue or persuade was the most important type of writing for incoming college students. Informational/expository writing conveys information. The purpose of this form of writing is to describe or explain something to a reader. Narrative writing conveys experience and can be used for multiple purposes. The structure of narrative writing can take the form of creative stories, poetry, or personal reflections and often include dialogue and figurative elements. Beyond text types and purposes, the ELA CCSS writing standards also focus on the production and distribution of writing; research to build and present knowledge, and the development of a range of writing.

The TNCore site hosts resources related to text-dependent questioning and using close, analytic reading as a strategy with students, including a guide to creating text-dependent questions, a checklist for evaluating question quality, and guides to close reading. They also provide several resources produced through a partnership with the

Institute for Learning, including a resource for developing text-based questions that outlines how to build questions over multiple readings of a text. The TNCore site does not include a targeted web page for writing instruction. The site does include some writing rubrics and sample prompts for teachers preparing students for the Tennessee Comprehensive Assessment Program (TCAP) writing assessment. Additionally, embedded within the summer training resources is a module devoted to the analysis of a writing research simulation task.

The shift towards developing students' ability to read for evidence to support arguments will require assessment companies to change the structure of their items to include a much stronger focus on writing, particularly evidence-based argumentative writing. Test-designers will need to demonstrate proficiency in developing text-dependent questions to support students' close, analytic reading of more complex text than they have previously.

Shift Three: Building Knowledge through Content Rich Nonfiction

The final shift of the ELA CCSS includes a movement towards increasing the amount of non-fiction, or informational text, students read both in and out of the classroom based on the premise that students should be exposed to information about the world around them to develop the knowledge and vocabulary required for college and career. Informational reading includes appropriately complex nonfiction text in history/social studies, sciences, technical studies, and the arts. In K-5, the standards require a 50-50 balance between informational and literary reading. In grades 6-12, the expectation is that students will read informational text 70% of the time, through both

their ELA and content area courses, such as history/social students, science, and technical subjects. One of the greatest shifts in the ELA CCSS is the shared responsibility of reading instruction between all disciplines; expecting that reading, writing, speaking, and listening span the school day from K-12 as integral parts of every subject.

Traditionally, students in the early and middle grades have not been expected to read much informational text. In fact, as little as 7 and 15 percent of elementary and middle school instructional reading, for example, is expository (Yopp & Yopp, 2006). A study of basal readers showed that only 20 percent of the selections were informational texts, resulting in few opportunities for students to engage with informational texts and become familiar with their structures and features (Moss & Newton, 2002). However, informational text makes up the vast majority of reading required college and the workplace. The 2007 Achieve report, *Closing the Expectations Gap*, outlines steps required for a better alignment between K-12 and postsecondary institutions. In a survey of 29 states, they discovered that K-12 standards and courses tended to emphasize literature while most of the reading students will encounter in college or workplace is informational (p. 9). Furthermore, research has found that informational text is harder for most students to read than literary text (Heller & Greenleaf, 2007; Shanahan & Shanahan, 2008). Put together, the lack of focus on informational text in K-12, the gap in alignment between K-12 and postsecondary institutions, and the additional challenge informational text presents have resulted in a generation of students who are unprepared for college and career expectations – a challenge the ELA CCSS hopes to address with a strong emphasis on content-rich nonfiction text.

There is also a growing body of research that suggests that informational text can be more motivating and engaging for young readers than its literary counterpart. Stien and Beed (2004) conducted a study with 3rd graders which paired a non-fiction text with a fiction text and found that students became more invested in reading and would breeze through fictional text to get to the non-fiction. Correia (2011) studied the choices of kindergartners over a 19 week period regarding library books signed out as fiction vs. nonfiction and found that nonfiction was chosen over fiction in 14 out of 19 weeks. Others, (Maloch & Bomer, 2013; Dreher, 2003), have similarly found that young children are often motivated by informational texts, and their use in classrooms result in more engaged children.

However, not all research has demonstrated that inclusion of informational text improves student outcomes. Baker, et al. (2011) conducted a longitudinal informational text intervention study with low-income second to fourth graders. Over a period of two years, they provided students with more access to informational text through their classrooms and libraries. Their findings suggest that their informational text infusion has little impact on student reading comprehension or reading engagement.

The TNCore site currently has few resources related to informational text. The most relevant resource available to Tennessee teachers is a review instrument that assists districts in choosing textbooks for their K-3 grades that contain the recommended ratio of informational and literary text.

The shift towards utilizing more informational text is significant for companies designing benchmark assessments to measure student progress. Test-designers may need guidance in selecting text and incorporating them into benchmark assessments over the course of the year.

Benchmark Assessments

The post No Child Left Behind educational climate has been focused on political and public demand for accountability of student learning in schools. Benchmark assessments are now a typical weapon in the arsenal of schools and districts hoping to prevent failure. However, benchmark assessments owe their roots to formative assessments, which aim to provide interim feedback to teachers about students' progress toward meeting standards that will be measured and assessed on high-stakes summative state tests (Burke, 2010; Popham, 2008).

The contemporary use of the term "formative assessment" can be traced to Michael Scriven's 1967 trailblazing essay about educational evaluation in which he first contrasted "formative" and "summative" to indicate the differences in goals for collecting evaluation information and how that information is used (Scriven, 1967). British researchers Paul Black and Dylan Wiliam later published a meta-analysis that extensively reviewed more than 250 empirical research studies focused on classroom assessment and discovered that, when properly employed, formative assessment helped students learn substantially better (Black & Wiliam, 1998). However, despite the growing emphasis on formative assessment since Black and Wiliam's 1998 study, Leung and Mohan (2004) have indicated that formative assessment has eluded proper study in more recent times.

"Formative assessment's status as an ethereal construct has further been perpetuated in the literature due to the lack of an agreed upon definition" (Dunn & Mulvenon, 2009, p. 2).

Formative assessment supports benchmarking, the process of comparing learning outcomes goals to selected standards for the purpose of overall improvement. Insight into whole-class and individual progress gained through continually measuring understanding helps both the teacher and students identify strengths, points of confusion, and the additional skill and knowledge development that will further progress toward mastery (Greenstein, 2010). Regular use of benchmark assessments, particularly those aligned with state standards, is widely perceived as having the potential to improve student performance. According to a survey of school superintendents in 2005, approximately 70% of school districts used benchmark tests as a component of their assessment programming (Olson, 2005).

Yet, studies on the impact of utilizing benchmark assessments to improve student outcomes are inconclusive. For example, in 2007, Henderson, Petrosino, Guckenburg, and Hamilton reviewed the data for quarterly benchmark assessments in Massachusetts and found no statistically significant differences between those who took the assessments and those who did not. Yet, Faria, et al (2012) analyzed benchmark assessment data from more than 100 schools in four districts and reported small and slightly significant effects on student achievement for those who took the benchmarks.

Herman, Osmundson, and Dietal (2010) identified several criteria that schools should consider when selecting or developing benchmark tools. Validity is the all-

encompassing concept that defines the quality of an educational measurement. It defines the extent to which an assessment measures what it is intended to measure and provides comprehensive information supporting the purposes for which it is being used. Consequently, benchmark assessments themselves are not valid or invalid; instead, the validity rests on the underlying evidence for the benchmark assessment's specific use (Herman et al., 2010).

As a result, Herman et al. (2010) concluded that benchmark assessments must "be aligned with district and school learning goals and intended purposes" (p.6). Alignment refers to the extent that what is being assessed complements what is being taught. Aligned assessments should capture both the depth and breadth of learning standards, signify the most important concepts and skills being taught, reflect the consistency and sequence of the local and state curriculum (Herman et al., 2010; Popham, 2010). For the purposes of this study, it is important to consider the importance of this definition of benchmark assessment validity. These assessments are only valid to the extent to which they measure the framework they are designed to assess.

Data Analysis

Several data analysis methods were utilized in this study to answer the research questions. Exploratory factor analysis (EFA) was utilized to consider confirmation of alignment to either the three shifts or four tested strands of CCSS ELA. Additionally, both classical test theory (CTT) and item response theory (IRT) were utilized to consider internal test consistency, item difficulty, and item discrimination. The subsections below more fully describe the mathematical theories behind each of these methods.

Exploratory Factor Analysis (EFA)

Factor analysis is a multivariate statistical procedure that can be used to reduce a large number of variables into a smaller set of variables (or factors); establish underlying dimensions between measured variables and latent constructs; and provide construct validity evidence of self-reporting scales (Williams, Onsman, & Brown, 2010). Exploratory factor analysis (EFA) is utilized when a researcher does not have a set number of factors expected to result from data analysis and is thus exploratory in nature.

Because making meaning from the results often relies upon the researcher, using factor analysis has been criticized as a statistical approach due to its subjective nature (Tabachnick & Fidell, 2007; Preacher, Zhang, Kim & Mels, 2013). However, it is still often employed for its usefulness. Kerlinger and Lee (2000) describe sample size and factor extraction as key features of factor analysis.

To have meaningful factors, the data must be meaningful; therefore, sample size is important. However, Williams, Onsman, and Brown (2010) also note that researchers have not come to agreement on the sample size necessary for factor analysis, though the general consensus is that 300 is good.

There are several methods for extracting factors from data sets. The most common are principal-axis factoring (PAF), principal components analysis (PCA), and maximum likelihood (ML) method. In a literature review of EFA practice in research, Conway and Huffcutt (2003) found that PCA was utilized 39.6% of the time; PAF was used 22.4% of the time; and ML was used 3.8% of the time. For the purpose of this

study, PCA was utilized to consider confirmation of factor alignment to either the three shifts or four strands of CCSS.

Classical Test Theory (CTT)

There are several fundamental differences between CTT and IRT. CTT relies predominantly on the total test score while IRT relies on individual responses to items (de Ayala, 2009). Researchers utilizing CTT are primarily interested in investigating the properties and relationships between respondents' total scores on an instrument while researchers utilizing IRT are primarily interested in understanding how the individual items are related to the underlying factor and each respondent's ability on each of these factors. In IRT, estimations of the properties of items, such as their levels of difficulty and their capacity to discriminate between respondents of varying levels of ability, can be used to obtain estimates of the abilities of the respondents.

Spearman first developed CTT over a century ago (Hambleton & van der Linden, 1982). CTT's primary focus is on test-level information, although the CTT can also be useful for item statistics, such as item difficulty and item discrimination. CTT functions by collectively considering a sample of examinees and examining their success rate on each item of a test. The CTT can be expressed as

$$X_{ip} = T_{ip} + E_{ip} \quad (1)$$

where, X_{ip} is the observed score for item i and person p, T_{ip} is the true scores for item i and person p, and E_{ip} is the error score for item i and person p.

CTT can provide several pieces of information that help ensure reliability and validity of a test. First CTT provides descriptive statistics, such as mean and variance, which can inform researchers about the variability on items. For example, if an item has low variance, it may not be a useful item (Kline, 2005). Second, CTT can provide information on an item's difficulty level. The success rate of a particular sample of test-takers on an item, known as the p-value of the item, is considered the item's difficulty. P-values vary between 0 to 1 and mid-range values are considered to be the strongest. Third, CTT can be used to determine the capacity of an item to discriminate between higher ability test-takers and lower ability test-takers. The discrimination index (*D*) is calculated using the following formula:

$$d_i = U_i/n_{iU} - L_i/n_{iL} \quad (2)$$

where,

U_i: # of people in the upper group who have the item i correct,

n_{iU}: # of people in the upper group,

L_i: # of people in the lower group who have the item i correct, and

n_{iL}: # of people in the lower group.

Fourth, CTT can be used for item-to-total correlations conducted through Pearson product-moment item-to-total correlation coefficient, which analyzes how responses to items relate to the total text score.

The two predominant psychometric properties that must be considered for a measurement to be psychometrically sound are validity and reliability according to CTT. Validity is concerned with the extent to which an instrument measures what it is intended

to measure, and reliability is concerned with the ability of an instrument to measure consistently (Tavaskol & Dennick, 2011). Cronbach's alpha is an "internal consistency measurement that describes the extent to which all the items in a test measure the same concept or construct" (Tavaskol & Dennick, 2011, p. 53) and is commonly used to measure reliability.

Though CTT has been the primary form of test measurement for over 100 years, CTT has recently been eclipsed by IRT as the stronger model of a test measurement framework, specifically regarding test construction research (Hambleton & Jones, 1993). According to Fan (1998), the main limitation of CTT is its circular dependency in which "(a) The person statistic (i.e. observed score) is (item) sample dependent, and (b) the item statistics (i.e. item difficulty and item discrimination) are (examinee) sample dependent" (p. 2). This circular dependency limits CTT to only examining the test as a whole, and not individual items, preventing psychometricians from identifying specific items that may need modifications. An example of this circular dependency is that the CTT cannot account for the test taker's abilities; therefore, item validation results could be skewed by a person's ability. Hambleton, Swaminathan, and Rogers (1991) identify another primary limitation of CTT as parallel test assumption. According to this assumption, all tests from a battery should have identical true scores and error variances, which presents a challenge in practice because it is usually not possible for the performance on two tests to be identical so that the true score and error variance will be equal from two test scores.

Item Response Theory (IRT)

Though IRT was developed around the mid-20th century, only in the past 15 years have computer programs made IRT analysis possible for psychometricians. Whereas the focus of CTT is typically on one test score, treating each item as though they were equal, IRT focuses on the patterns of responses that the examinee makes to the set of items. The fundamental assumption of IRT is that there is a connection between an item on the test and the characteristic being assessed by the test, known as a latent trait (Kline, 2005). IRT is useful in eliminating most of the problems associated with CTT (Hambleton, Swaminathan, & Rogers, 1991).

The most basic model in IRT is the one-parameter logistic model (1PL), which is often referred to as the Rasch model after the man who developed it in 1960 (Kline, 2005). This model estimates the difficulty item parameter. The next model is the two-parameter logistic model (2PL) which estimates both the difficulty and item discrimination parameters. The three-parameter logistic model (3PL) will be utilized in the current study. This model estimates the difficulty, item discrimination, and pseudo-chance parameters.

The basic 3-parameter IRT model can be described as,

$$P_i(\theta) = c_i + (1-c_i)\frac{e^{Da_i(\theta-b_i)}}{1+e^{Da_i(\theta-b_i)}} \quad (3)$$

where $P_i(\theta)$ is the probability to get an item correct for given θ,

θ is the latent trait (ability or proficiency),

a_i is Item discrimination parameter,

b_i is Item difficulty parameter, and

c_i is pseudo-chance parameter.

The formulas of IRT can be more intuitive than the formulas of CTT. For example, CTT relies upon a single reliability while IRT utilizes local reliabilities resulting in more opportunities to discover complex information about a test. Additionally, CTT has no invariance in parameter estimates, but IRT asserts that items are not concurrently found more and less difficult, regardless of the population used in a study (Hambleton & Jones, 1993), which is beneficial when deciding whether to retain or delete items from a test. The main advantage of IRT over CTT is that IRT models are falsifiable models with empirical data, which is not possible with CTT (Kim & Nicewander, 1993). IRT models can be tested with actual data for the fit of the models to real data sets.

Summary

The Common Core State Standards Initiative is a major undertaking, requiring coordination of many moving parts within multiple agencies, and is possibly the most intense education reform movement our nation has ever seen. Though Tennessee is now

five years into the process of implementation, only this year, the first official year of CCSS-aligned assessment implementation, will state education leaders be able to truly measure the success of their endeavors.

The actual impact of CCSS implementation will depend much less on the standards themselves than on how they are used. The three key instructional shifts of the ELA CCSS represent the major, measurable expectations for teachers to be able to implement the CCSS with fidelity. Teachers require quality professional development and time to be prepared for this level of change.

Mathis (2012) highlights two factors that will be particularly crucial in measuring successful implementation: "The first is whether states invest in the necessary curricular and instructional resources and supports, and the second concerns the nature and use of CCSS assessments developed by the two national testing consortia" (p. 1). The second factor listed is problematic because Tennessee, as a state, has yet to administer a CCSS-aligned assessment to measure student achievement and growth. Therefore, any measurement of successful implementation will rely on the primary determining factor, state investment in necessary curricular and instructional resources, such as benchmark assessments.

The purpose of this study is to use exploratory factor analysis (EFA to consider if factor structure of items aligned to the three instructional shifts of the ELA CCSS can be confirmed within benchmark assessments designed to measure student progress across three grade levels: 4^{th}, 8^{th}, and 10^{th}. Additionally, classical test theory (CTT), and item response theory (IRT) are utilized to remove misfit items and further hone the data set.

These statistical procedures were selected for analysis of Tennessee benchmark assessments to answer the following research questions:

1. Do the assessments have sound psychometric properties, such as validity and reliability?
2. Do the assessments show three- or four-factor solutions through factor analysis?
3. Are the benchmark assessments used by TN schools/students aligned to the shifts or strands of the ELA CCSS?

CHAPTER III: METHODOLOGY

Participants

The archival data contained a total of 8,621 participants for a series of two 4th grade, 8th grade, and high school English/Language Arts assessments in Tennessee. The first test (A) in the series was administered between August and September of 2014. The second test (B) in the series was administered between November and December of 2014. The sample contained 1,343 4th grade participants (678 on Test A and 665 on Test B), 1,089 8th grade participants (574 on Test A and 515 on Test B), and 6,189 10th grade participants (2,957 on Test A and 3,232 on Test B). The archived data were collected during the 2014-2015 academic year as the results for a benchmark assessment. There are no missing data for each assessment item or the total scores from the test; however some demographic variables have missing data including gender, race, English as a Second Language (ESL), Special Education (SpEd), and Free and Reduced Lunch. These demographic variables were not analyzed because the current study mainly seeks to analyze item-level information. The data did not include any information that could lead to the identification of participants; however, all data points were assigned subject ID numbers.

Measurement

The test items for the benchmark test series analyzed in this study were developed by a for-profit testing company in the United States. For the purposes of this study, the researcher examined the first and second of a series of tests designed to be administered to 4th, 8th, and 10th grade students over the course of the 2014-2015 school year.

The series of tests were designed to measure student proficiency on the ELA CCSS and each item was aligned to specific standards. The researcher did not have access to the actual questions, but was able to access which standards to which items were aligned. The tests only measured proficiency for four of the six strands of ELA CCSS: Language (L), Reading Informational Text (RI), Reading Literature (RL), and Writing (W). The tests did not include items aligned to Speaking and Listening or Reading Foundational Skills. All tests contain multiple choice items with four alternatives.

Both 4th grade tests analyzed contained 30 items. Test 4A contained 5 Language items, 9 Reading Informational Text items, 7 Reading Literature items, and 9 Writing items. Test 4B contained 7 Language items, 7 Reading Informational Text items, 8 Reading Literature items, and 8 Writing items. Test 4A was administered between August 8, 2014 and November 24, 2014. Test 4B was administered between December 4, 2014 and March 4, 2015.

Both 8th grade tests contained 34 items. Test 8A contained 6 Language items, 10 Reading Informational Text items, 9 Reading Literature items, and 9 Writing items. Test 8B contained 6 Language items, 10 Reading Informational Text items, 9 Reading Literature items, and 9 Writing items. Test 8A was administered between August 21, 2014 and October 22, 2014. Test 8B was administered between November 17, 2014 and December 23, 2015.

Both 10th grade tests contained 32 items. Test 10A contained 7 Language items, 9 Reading Informational Text items, 10 Reading Literature items, and 6 Writing items.

Test 10B contained 6 Language items, 9 Reading Informational Text items, 10 Reading Literature items, and 7 Writing items. Test 10A was administered between August 21, 2014 and October 31, 2014. Test 10B was administered between October 21, 2014 and March 25, 2015.

Procedures

Means and standard deviations for each item in all six assessments were computed. Then, to confirm reliability of the benchmark tests, Cronbach's alpha (*a*) was computed for each assessment as an internal consistency reliability index. Additionally, to confirm concurrent validity across each grade level series, Pearson's correlation coefficient (*r*) was computed using total raw test scores for students who took both A and B tests for each grade level. Pearson's correlation coefficient is a statistical measure of the strength of a linear relationship between paired data.

Exploratory factor analysis (EFA) was then utilized to consider confirmation of alignment to either the three shifts or four tested strands of CCSS ELA. The EFA results for all six data sets were unclear but met the unidimensionality assumption to apply IRT analyses. Item and test analyses using both CTT and IRT were conducted to identify misfit items to be removed from each data set. In CTT, the mean value of each item is identical to the *p*-value, or item difficulty index. Item-test correlation were also computed as part of CTT analyses as an item discrimination index. Three IRT models (1, 2, and 3PLM) were tested with the data. The best-fitting model was selected, and the three comparison criteria (-2LL, AIC, and BIC) were recorded. IRT analyses included the following indices: estimation of slope (a-parameter) location (b-parameter), and

pseudo-chance (c-parameter) indices as well as item and test information function for each factor.

To identify and remove misfit items before conducting a second round of EFA, two index statistics were considered: item-test correlations in CTT (*ITC)* and discriminant coefficients (*a*-parameter) in IRT, which both indicate an item's ability to discriminate between test taker abilities. Items with item-test correlations below .19 are considered poor items and should be removed (Ebel & Frisbie, 1986). For the purposes of this study, a stricter standard of 0.25 was used to eliminate items. Items with discriminant coefficients with 0.5 or below are generally considered low in their ability to discriminate. For the purposes of this study, a stricter standard of 0.6 was used to eliminate items. Once misfit items were removed, EFA was conducted once more to reconsider confirmation of alignment to either the three shifts or four tested strands of ELA CCSS.

CHAPTER IV: RESULTS

Reliability and Validity

Before conducting exploratory factor analysis (EFA), tests for reliability and validity were performed. Cronbach's alpha was computed for each assessment confirm reliability of the benchmark tests as an internal consistency reliability index. Nunnaly (1978) has indicated 0.7 to be an acceptable reliability coefficient and all six tests exceeded that limit: 4A (a = 0.803), 4B (a = 0.851), 8A (a = 0.839), 8B (a = 0.881), 10A (a = 0.825), and 10B (a = 0.804).

Additionally, to confirm validity across each grade level series, Pearson's correlation coefficient (r) was computed using total raw test scores for students who took both A and B tests for each grade level. Evans (1996) suggested for the absolute values of r between 0-.19 are "very weak," .20-.39 are "weak," .40-.59 are "moderate," .60-.79 are "strong," and .80-1.0 are "very strong." Based on the results of the analysis, student performance on test 4A is strongly related to student performance on test 4B (r = .68, N= 475, p < .05). Based on the results of the analysis, student performance on test 4A is strongly related to student performance on test 4B r = .68, N= 475, p < .05. Based on the results of the analysis, student performance on test 4A is strongly related to student performance on test 4B, r = .68, N= 475, p < .01; student performance on test 8A is strongly correlated to test 8B, r = .76, N = 483, p < .01; and student performance on test 10A is strongly correlated to test 10B, r = .63, N = 2565, p < .01.

Exploratory Factor Analysis (EFA)

In this study, EFA was utilized to identify the number of factors within six dichotomous data sets including a series of two 4^{th} (N = 678; 665), 8^{th} (N = 574; 515), and 10^{th} (N = 2957; 3232) grade reading comprehension benchmark assessments based on the ELA CCSS. The standards the benchmarks were designed to assess if items are anchored in three key instructional shifts, but are also divided into four tested strands. Our interest was to consider confirmation of alignment of items to either the three shifts or four tested strands of ELA CCSS.

The EFA results for all six data sets were unclear depending on the method utilized for determining factors. For example, when examining the eigenvalues, all data sets exhibited at least seven components with the eigenvalues greater than 1 (William, Brown, & Onsman, 2012). However, the scree plots demonstrate a tendency towards one-factor solutions because the slope of the curves reach a steady rate from the second factor. Table 1 shows the first eigenvalues and percent of variance explained by the first four components for each data set and Figure 1 shows the scree plots. Deeper inspection of the factor loadings additionally shows that for each data set, more items loaded on factors when testing for a three-factor solution than when testing for a four-factor solution.

Because the results were unclear, but demonstrated a promising tendency towards a three-factor solution, CTT and IRT analysis were employed to remove any misfit items with the expectation that removal of bad items from each data set would produce a clearer understanding of which factor-solution is best aligned to the benchmark assessments.

Table 1 Eigenvalues and Cumulative Percent of Variance Explained for First Four Components by Test

Test	*Component*	*Eigenvalue*	*Cumulative %*
4A	1	4.70	15.7
	2	1.30	20.0
	3	1.22	24.1
	4	1.20	28.1
4B	1	5.93	19.8
	2	1.41	24.5
	3	1.20	28.5
	4	1.18	32.4
8A	1	5.71	16.8
	2	1.48	21.2
	3	1.24	24.8
	4	1.21	28.4
8B	1	7.11	20.9
	2	1.39	25.0
	3	1.34	29.0
	4	1.16	32.4
10A	1	5.24	16.4
	2	1.43	20.8
	3	1.19	24.6
	4	1.09	28.0
10B	1	4.72	14.8
	2	1.33	18.9
	3	1.12	22.4
	4	1.08	25.8

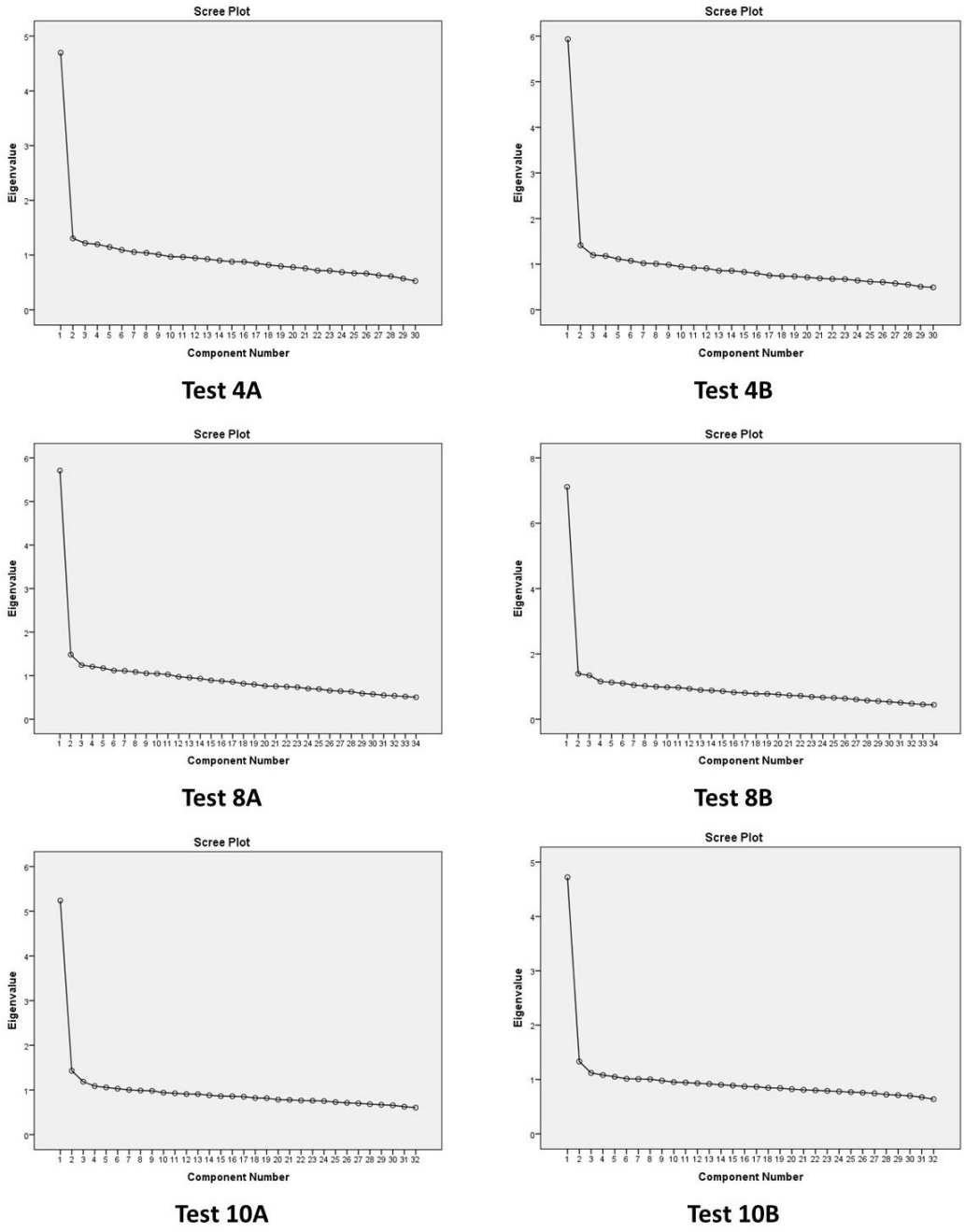

Figure 1: Scree plot of eigenvalues by test

CTT and IRT Analysis

In this study, CTT and IRT models were applied to all six data sets in order to identify misfit items to be removed before reanalyzing the data with a second round of EFA. Based on the scree plots generated by EFA, we concluded that the data met the unidimensionality assumption. Therefore, the application of an IRT model to the data was justified.

For the item and test analyses in CTT, *p*-values and item-test correlations were computed. In CTT, the *p*-value, or item difficulty, is identical to the mean of participants who answered each item correctly. A range of *p*-values, or difficulty levels, between .30 and .70 on a test is desirable because it demonstrates that the test is of varying difficulty. Very high or very low *p*-values will restrict the range of answers and generally correspond to low item-test correlations, though it is not always the case. Additionally, positive item-test correlations in the midrange (about 0.50 and above) are desirable because they show that each item is correlated to the overall test score, which shows how well the item is able to discriminate between test takers.

For item and test analyses in IRT, we analyzed estimation of location (*b*-parameter), slope (*a*-parameter), and pseudo-chance (*c*-parameter) indices for each test item. The *a*-value, or item discrimination index, of an item indicates the slope of the item characteristic curves (*ICC*s) at the point of inflection for each item and suggests the extent to which item responses vary as theta levels change. The *b*-value of an item, or item difficulty index, is the average of the *b*-parameters for that item. The *c*-value indicates the pseudo-chance parameter, which may include guessing component along with others factors for each item. As shown in Table 2, the model-fit indices of three

IRT models indicated that 3PLM fit the given data for all six sets best when 3PLM was compared with 1PLM and 2PLM.

Tables 3-5 show the CTT and IRT analysis results for each grade level. Test 4A demonstrates a range of *p*-values (0.33-0.76). In CTT analysis, the hardest items are 24, 23, and 18 because they have the lowest *p*-values. The easiest items are 4, 13, and 3. The average item-test correlation for Test 4A is close to the mid-range at 0.39. The strongest discriminating items are 15, 28, and 17 and the weakest items are 26, 7, and 9. In IRT analysis, the average *b*-value, or difficulty, for Test 4A is 0.82. We expect the average *b*-value to be zero, so we can interpret that this test is harder than usual. The hardest items are 26, 30, and 7 and the easiest items are 4, 13, and 8. The average *a*-value, or discrimination level, for Test 4A is 0.84. We would expect an average *a*-value of one, so we can interpret that this test has a low ability to discriminate between test-takers. The strongest discriminating items are 24, 23, and 15 and the weakest items are 9, 26, and 8. Both CTT and IRT analyses show that items 4 and 13 are the easiest, that item 15 is the strongest at discriminating, and that items 26 and 9 are the weakest at discriminating.

Test 4B demonstrates a range of *p*-values (0.32-0.82). In CTT analysis, the hardest items are 9, 20, and 30 and the easiest items are 2, 3, and 5. The average item-test correlation for Test 4B is close to the mid-range at 0.43. The strongest discriminating items are 6, 15, and 18 and the weakest items are 27, 9, and 30. In IRT analysis, the average *b*-value, or difficulty, for Test 4B is 0.37, so we can interpret that this test is slightly harder than usual. The hardest items are 27, 9, and 30 and the easiest items are 2, 1, and 5. The average *a*-value is 0.93, so we can interpret that this test has an

average ability to discriminate between test-takers. The strongest discriminating items are 6, 15, and 29 and the weakest items are 1, 23, and 2. Both CTT and IRT analyses show that items 2 and 5 are the easiest, while items 9 and 30 are the most difficult. Additionally, they show that items 6 and 15 are the strongest discriminating items.

Test 8A demonstrates a range of *p*-values (0.30-0.87). In CTT analysis, the hardest items are 17, 27, and 23 and the easiest items are 2, 1, and 3. The average item-test correlation for Test 8A is close to the mid-range at 0.40. The strongest discriminating items are 19, 32, and 26 and the weakest items are 23, 8, and 2. In IRT analysis, the average *b*-value, or difficulty, for Test 8A 4A is 0.18. We expect the average *b*-value to be zero, so we can interpret that this test is slightly harder than usual. The hardest items are 17, 23, and 13 and the easiest items are 2, 1, and 9. The average *a*-value, or discrimination level, for Test 8A is 0.87. We would expect an average *a*-value of one, so we can interpret that this test has a slightly low ability to discriminate between test-takers. The strongest discriminating items are 19, 18, and 4 and the weakest items are 2, 1, and 8. Both CTT and IRT analyses show that items 2 and 1 are the easiest, while items 17 and 23 are the most difficult. Additionally, they show that item 19 the strongest discriminating item, while 8 and 2 are the weakest.

Test 8B demonstrates a range of *p*-values (0.34-0.88). In CTT analysis, the hardest items are 28, 32, and 23 and the easiest items are 13, 3, and 4. The average item-test correlation for Test 8B is close to the mid-range at 0.45. The strongest discriminating items are 11, 25, and 20 and the weakest items are 14, 3, and 18. In IRT analysis, the average *b*-value, or difficulty, for Test 8B is -0.44. We expect the average *b*-value to be zero, so we can interpret that this test is easier than usual. The hardest

items are 14, 28, and 32 and the easiest items are 3, 13, and 33. The *a*-value, or discrimination level, for Test 8B is 1.19. We would expect an average *a*-value of one, so we can interpret that this test has a strong ability to discriminate between test-takers. The strongest discriminating items are 13, 4, and 9 and the weakest items are 14, 18, and 3. Both CTT and IRT analyses show that items 13, and 3 are the easiest, while item 28 and 32 are the most difficult. Additionally, they show that items 14, 18, and 3 are the weakest discriminating items.

Test 10A demonstrates a range of *p*-values (0.19-0.79). In CTT analysis, the hardest items are 32, 31, and 25 and the easiest items are 2, 22, and 3. The average item-test correlation for Test 10A is close to the mid-range at 0.39. The strongest discriminating items are 22, 19, and 28 and the weakest items are 32, 31, and 27. In IRT analysis, the average *b*-value, or difficulty, for Test 10A is 0.98. We expect the average *b*-value to be zero, so we can interpret that this test is harder than usual. The hardest items are 27, 26, and 32 and the easiest items are 2, 22, and 8. The *a*-value, or discrimination level, for Test 10A is 0.92. We would expect an average *a*-value of one, so we can interpret that this test has an average ability to discriminate between test-takers. The strongest discriminating items are 22, 17, and 16 and the weakest items are 5, 27, and 23. Both CTT and IRT analyses show that items 2 and 22 are the easiest, while items 32 is of the most difficult. Additionally, they show that item 27 is of the least discriminating.

Test 10B demonstrates a range of *p*-values (0.24-0.72). In CTT analysis, the hardest items are 28, 27, and 24 and the easiest items are 1, 10, and 3. The average item-test correlation for Test 10B is close to the mid-range at 0.37. The strongest

discriminating items are 25, 19, and 30 and the weakest items are 15, 29, and 8. In IRT analysis, the average *b*-value, or difficulty, for Test 10B is 1.16. We expect the average *b*-value to be zero, so we can interpret that this test is harder than usual. The hardest items are 15, 24, and 27 and the easiest items are 1, 10, and 3. The *a*-value, or discrimination level, for Test 10B is 0.85. We would expect an average *a*-value of one, so we can interpret that this test has a low ability to discriminate between test-takers. The strongest discriminating items are 28, 25, and 10 and the weakest items are 8, 1, and 15. Both CTT and IRT analyses show that items 1, 10, and 3 are the easiest, while items 24 and 27 are the most difficult. Additionally, they show that item 25 is the strongest discriminating item, while 15 and 8 are the weakest.

To identify and remove misfit items from the second round of EFA, two index statistics were considered: item-test correlations (*ITC*) in CTT and discriminant coefficients (*a*-parameter) in IRT, which both indicate an item's ability to discriminate between test taker ability. Items with item-test correlations below .19 are considered poor items should be removed (Ebel & Frisbie, 1986). For the purposes of this study, a strict standard of 0.25 was used to eliminate items. Items with discriminant coefficients with 0.5 or below are generally considered low in their ability to discriminate. For the purposes of this study, a strict standard of 0.6 was used to eliminate items. Table 6 show the items removed from EFA analysis based on their item-test correlations and discriminant coefficients.

Table 2 Model-fit Indices of Three Traditional IRT Models for Each Data Set

Test	IRT model	*NP*	*df*	*-2LL*	$-2LL_{difference}$
4A	1PLM	30	676	24332	
	2PLM	60	518	23729	603
	3PLM	90	360	23521**	208
4B	1PLM	30	659	23182	
	2PLM	60	510	22097	1085
	3PLM	90	360	21399**	698
8A	1PLM	34	573	22142	
	2PLM	68	491	21018	1124
	3PLM	102	408	20480**	538
8B	1PLM	34	507	21692	
	2PLM	68	458	20792	900
	3PLM	102	408	20480**	312
10A	1PLM	32	2953	112672	
	2PLM	64	1669	106864	5808
	3PLM	96	384	105024**	1840
10B	1PLM	32	3225	126087	
	2PLM	64	1805	119276	6811
	3PLM	96	384	118464**	812

Table 3 4th Grade Test A and Test B CTT and IRT

4th Grade Test A

	CTT		IRT		
Item	*P*	*ITC*	*a*	*b*	*c*
1	0.658	0.357	0.597	-0.011	0.307
2	0.606	0.471	0.949	0.176	0.280
3	0.671	0.462	1.034	0.081	0.372
4	**0.757-**	0.405	0.771	**-0.764-**	0.251
5	0.474	0.354	0.688	1.005	0.253
6	0.624	0.450	0.755	-0.041	0.239
7	0.389	0.260	0.660	1.833	0.262
8	0.655	0.363	0.587	-0.201	0.248
9	0.481	0.274	**0.491***	1.199	0.255
10	0.586	0.429	0.809	0.282	0.258
11	0.502	0.398	0.798	0.740	0.252
12	0.382	0.430	0.969	1.199	0.223
13	0.704	0.458	0.886	-0.404	0.251
14	0.642	0.443	0.746	-0.098	0.247
15	0.516	**0.503****	1.037	0.486	0.235
16	0.398	0.356	0.863	1.337	0.245
17	0.485	0.474	0.999	0.740	0.248
18	0.344	0.374	1.024	1.489	0.228
19	0.422	0.420	1.003	1.098	0.245
20	0.435	0.324	0.738	1.305	0.260
21	0.369	0.360	0.848	1.466	0.233
22	0.447	0.466	0.919	0.856	0.228
23	0.338	0.406	1.037	1.411	0.217
24	**0.333+**	0.307	**1.085****	1.757	0.247
25	0.454	0.423	0.941	0.946	0.248
26	0.410	**0.196***	0.530	**2.088+**	0.276
27	0.382	0.340	0.865	1.504	0.249
28	0.667	0.493	1.000	-0.197	0.250
29	0.358	0.367	0.972	1.459	0.234
30	0.499	0.392	0.618	1.925	0.252

Note. Table 3.1: *n* = 678; Cronbach's *α* = .803

4th Grade Test B

	CTT		IRT		
Item	*P*	*ITC*	*a*	*b*	*c*
1	0.756	0.354	**0.554***	-0.971	0.234
2	**0.817-**	0.383	0.597	**-1.430-**	0.226
3	0.782	0.468	0.960	-0.682	0.334
4	0.671	0.430	0.749	-0.249	0.247
5	0.776	0.455	0.863	-0.842	0.244
6	0.719	**0.614****	**1.472****	-0.448	0.229
7	0.647	0.441	0.771	-0.090	0.250
8	0.621	0.509	0.995	0.022	0.242
9	**0.323+**	0.312	0.929	1.645	0.225
10	0.483	0.468	0.932	0.662	0.231
11	0.705	0.456	0.744	-0.482	0.241
12	0.583	0.490	0.853	0.385	0.230
13	0.427	0.358	0.729	1.172	0.240
14	0.415	0.458	1.016	0.923	0.219
15	0.626	0.595	1.408	-0.010	0.235
16	0.492	0.422	0.913	0.736	0.250
17	0.674	0.468	0.721	-0.333	0.235
18	0.571	0.587	1.310	0.158	0.221
19	0.483	0.485	1.156	0.618	0.231
20	0.334	0.329	1.093	1.465	0.227
21	0.666	0.549	1.146	-0.182	0.242
22	0.535	0.414	0.696	0.489	0.238
23	0.623	0.363	0.567	0.023	0.247
24	0.417	0.439	0.959	0.984	0.225
25	0.376	0.340	0.806	1.469	0.239
26	0.475	0.458	0.887	0.696	0.227
27	0.411	**0.243***	0.610	**1.731+**	0.266
28	0.433	0.387	1.029	1.012	0.247
29	0.672	0.437	1.400	1.203	0.206
30	0.373	0.317	0.907	1.514	0.249

Note. Table 3.2: *n* = 665; Cronbach's *α* = 851

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: + hardest item; - easiest item; * worst discriminating item; ** best discriminating item; () least alpha fit.

Table 4 8th Grade Test A and Test B CTT and IRT

8th Grade Test A

	CTT		IRT		
Item	*P*	*ITC*	*a*	*b*	*c*
1	0.840	0.305	0.502	-1.744	0.264
2	**0.873-**	0.271	**0.467***	**-2.241-**	0.265
3	0.838	0.309	0.791	-1.113	0.364
4	0.834	0.480	1.154	-1.080	0.245
5	0.739	0.412	0.719	-0.701	0.243
6	0.720	0.450	0.797	-0.526	0.247
7	0.749	0.458	0.824	-0.723	0.241
8	0.531	0.268	0.565	0.790	0.266
9	0.815	0.332	0.607	-1.302	0.253
10	0.598	0.434	0.683	0.120	0.240
11	0.671	0.495	1.002	-0.215	0.248
12	0.660	0.466	0.958	-0.131	0.255
13	0.362	0.284	0.740	1.691	0.242
14	0.727	0.505	1.105	-0.462	0.251
15	0.409	0.458	1.120	0.920	0.222
16	0.603	0.329	0.621	0.256	0.265
17	**0.300+**	0.305	0.945	**1.821+**	0.224
18	0.385	0.366	1.222	1.188	0.242
19	0.737	**0.570****	**1.372****	-0.499	0.245
20	0.559	0.410	0.742	0.403	0.251
21	0.382	0.319	0.852	1.422	0.243
22	0.531	0.359	0.788	0.619	0.261
23	0.324	**0.262***	0.967	1.773	0.239
24	0.627	0.517	1.015	-0.020	0.244
25	0.549	0.403	0.680	0.412	0.240
26	0.629	0.521	1.037	-0.050	0.239
27	0.305	0.313	1.099	1.673	0.225
28	0.467	0.393	0.997	0.866	0.254
29	0.589	0.408	0.679	0.190	0.243
30	0.587	0.395	0.750	0.278	0.257
31	0.376	0.369	1.020	1.287	0.237
32	0.570	0.529	1.097	0.197	0.233
33	0.411	0.331	0.818	1.280	0.248
34	0.662	0.448	0.818	-0.164	0.253

Note. Table 4.1: *n* = 574; Cronbach's *α* = .839

8th Grade Test B

	CTT		IRT		
Item	*P*	*ITC*	*a*	*b*	*c*
1	0.699	0.403	0.943	-1.057	0.264
2	0.491	0.511	1.227	0.025	0.265
3	0.800	0.320	0.790	**-1.964-**	0.364
4	0.796	0.507	1.785	-1.154	0.245
5	0.790	0.452	1.412	-1.265	0.243
6	0.598	0.444	1.017	-0.481	0.247
7	0.763	0.488	1.518	-1.083	0.241
8	0.678	0.411	0.952	-0.929	0.266
9	0.765	0.507	1.635	-1.053	0.253
10	0.726	0.448	1.181	-1.050	0.240
11	0.664	**0.540****	1.581	-0.638	0.248
12	0.736	0.513	1.589	-0.935	0.255
13	**0.880-**	0.479	**2.593****	-1.398	0.242
14	0.394	**0.259***	**0.483***	**0.943+**	0.251
15	0.616	0.411	0.906	-0.614	0.222
16	0.540	0.404	0.852	-0.221	0.265
17	0.616	0.407	0.891	-0.621	0.224
18	0.454	0.373	0.778	0.265	0.242
19	0.629	0.458	1.086	-0.610	0.245
20	0.542	0.517	1.287	-0.186	0.251
21	0.443	0.506	1.290	0.224	0.243
22	0.530	0.469	1.130	-0.145	0.261
23	0.388	0.499	1.281	0.458	0.239
24	0.524	0.477	1.121	-0.119	0.244
25	0.588	0.523	1.366	-0.369	0.240
26	0.443	0.430	0.909	0.294	0.239
27	0.573	0.388	0.822	-0.412	0.225
28	**0.336+**	0.440	1.064	0.784	0.254
29	0.596	0.494	1.213	-0.425	0.243
30	0.528	0.504	1.224	-0.132	0.257
31	0.482	0.466	1.106	0.074	0.237
32	0.363	0.424	0.980	0.684	0.233
33	0.785	0.416	1.205	-1.356	0.248
34	0.623	0.494	1.218	-0.544	0.253

Note. Table 4.2: *n* = 515; Cronbach's *α* = .881

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: + hardest item; - easiest item; * worst discriminating item; ** best discriminating item; () least alpha fit.

Table 5 10th Grade Test A and Test B CTT and IRT

10th Grade Test A

	CTT		IRT		
Item	*P*	*ITC*	*a*	*b*	*c*
1	0.479	0.407	0.765	0.743	0.216
2	**0.786-**	0.378	0.728	**-1.060-**	0.218
3	0.692	0.450	1.068	-0.164	0.319
4	0.623	0.410	0.759	0.043	0.254
5	0.526	0.331	**0.495***	0.750	0.246
6	0.441	0.493	1.235	0.762	0.211
7	0.509	0.391	0.866	0.768	0.274
8	0.686	0.425	0.732	-0.385	0.237
9	0.481	0.422	0.765	0.725	0.219
10	0.598	0.497	0.907	0.038	0.211
11	0.575	0.481	0.847	0.149	0.210
12	0.322	0.367	0.857	1.553	0.191
13	0.373	0.309	0.760	1.685	0.249
14	0.581	0.486	0.871	0.129	0.213
15	0.538	0.409	0.618	0.443	0.219
16	0.295	0.451	1.283	1.314	0.162
17	0.327	0.395	1.403	1.375	0.211
18	0.378	0.388	1.040	1.315	0.235
19	0.462	0.527	1.114	0.579	0.185
20	0.401	0.428	1.036	1.111	0.225
21	0.391	0.379	0.616	1.299	0.192
22	0.715	**0.537****	**1.611****	-0.451	0.234
23	0.400	0.294	0.521	1.765	0.240
24	0.393	0.490	1.173	0.936	0.190
25	0.258	0.300	0.840	2.104	0.183
26	0.269	0.285	0.641	2.335	0.184
27	0.309	0.242	0.514	**2.647+**	0.219
28	0.362	0.523	1.197	0.949	0.160
29	0.429	0.358	0.694	1.256	0.243
30	0.313	0.260	0.926	2.099	0.245
31	0.254	0.231	1.268	2.208	0.216
32	**0.192+**	**0.217***	0.611	2.264	0.172

Note. Table 5.1: *n* = 2957; Cronbach's *α* = .825

10th Grade Test B

	CTT		IRT		
Item	*P*	*ITC*	*a*	*b*	*c*
1	**0.724-**	0.360	0.538	**-0.802-**	0.220
2	0.454	0.382	0.582	0.899	0.186
3	0.686	0.413	0.752	-0.223	0.300
4	0.631	0.443	0.800	-0.086	0.229
5	0.385	0.311	0.834	1.619	0.261
6	0.518	0.395	0.690	0.665	0.246
7	0.381	0.399	0.933	1.258	0.215
8	0.550	0.271	**0.370***	0.777	0.258
9	0.485	0.422	0.694	0.667	0.201
10	0.689	0.458	0.889	-0.352	0.243
11	0.487	0.414	0.702	0.678	0.207
12	0.505	0.389	0.678	0.778	0.250
13	0.428	0.476	0.929	0.815	0.185
14	0.328	0.326	0.689	1.822	0.202
15	0.295	**0.186***	0.551	**3.161+**	0.240
16	0.387	0.381	0.749	1.329	0.212
17	0.391	0.402	0.954	1.257	0.228
18	0.296	0.386	1.129	1.536	0.182
19	0.495	0.491	0.936	0.555	0.209
20	0.307	0.351	0.783	1.749	0.187
21	0.378	0.363	0.775	1.433	0.220
22	0.357	0.369	0.862	1.503	0.218
23	0.348	0.435	0.895	1.253	0.174
24	0.285	0.286	0.763	2.180	0.205
25	0.455	**0.544****	1.286	0.627	0.191
26	0.355	0.335	0.905	1.678	0.244
27	0.279	0.290	1.012	2.038	0.211
28	**0.240+**	0.283	**1.480****	1.990	0.192
29	0.322	0.262	1.072	1.998	0.255
30	0.414	0.483	1.214	0.920	0.209
31	0.331	0.384	0.890	1.455	0.187
32	0.299	0.295	0.888	1.955	0.213

Note. Table 5.2: *n* = 3232; Cronbach's *α* = .804

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: + hardest item; - easiest item; * worst discriminating item; ** best discriminating item; () least alpha fit.

Table 6 Items Considered for Removal by Test

Test	*ITC*	*a*	*Removed*
4A	26	1, 8, 9, 26	1, 8, 9, 26
4B	27	1, 2, 23	1, 2, 23, 27
8A	none	1, 2, 8	1, 2, 8
8B	14	14	14
10A	31, 32	5, 23, 27	5, 23, 27, 31, 32
10B	15	1, 2, 8, 15	1, 2, 8, 15

ITC = Item-Test Correlation; a = discriminant coefficient

Exploratory Factor Analysis #2

After misfit items were removed, EFA was re-performed on all six data sets testing for both three- and four-factor solutions to consider confirmation of alignment to either the three shifts or four tested strands of ELA CCSS. Removal of the misfit items did not result in any significant changes in eigenvalues or scree plots for any of the six data sets. See Table 7 for eigenvalues and percent of variance explained after the second EFA was performed.

However, for all six data sets, the number of items that loaded onto factors when testing for both three- and four-factor solutions increased with slightly more items loading when testing the three-factor solution, indicating a better alignment of the data to three factors. Tables 8-13 show the factor loadings for both three- and four-factor solutions per item within each data set. On Test 4A, of the 26 items that remained in the analysis after removal of misfit items, 20 loaded when testing for a three-factor solution and only 19 loaded when testing for a four-factor solution. On Test 4B, of the 26 items that remained in the analysis after removal of misfit items, 23 loaded when testing for a three-factor solution and only 22 loaded when testing for a four-factor solution. On Test 8A, of the 31 items that remained in the analysis after removal of misfit items, 24 loaded when testing for a three-factor solution and only 21 loaded when testing for a four-factor solution. On Test 8B, of the 33 items that remained in the analysis after removal of misfit items, 25 loaded when testing for a three-factor solution and only 24 loaded when testing for a four-factor solution. On Test 10A, of the 28 items that remained in the analysis after removal of misfit items, 22 loaded when testing for a three-factor solution

and only 21 loaded when testing for a four-factor solution. On Test 10B, of the 28 items that remained in the analysis after removal of misfit items, 20 loaded when testing for a three-factor solution and only 18 loaded when testing for a four-factor solution.

Table 7 Eigenvalues and Cumulative Percent of Variance Explained for First Four Components by Test after Removal of Misfit Items

Test	*Component*	*Eigenvalue*	*Cumulative %*
4A	1	4.43	17.0
	2	1.24	21.8
	3	1.18	26.3
	4	1.12	30.7
4B	1	5.57	21.4
	2	1.32	26.5
	3	1.11	30.8
	4	1.10	35.0
8A	1	5.52	17.8
	2	1.48	22.6
	3	1.19	26.4
	4	1.18	30.2
8B	1	7.07	21.4
	2	1.37	25.6
	3	1.34	29.6
	4	1.16	33.1
10A	1	5.02	18.6
	2	1.31	23.5
	3	1.16	27.8
	4	1.08	31.6
10B	1	4.44	15.9
	2	1.31	20.5
	3	1.10	24.5
	4	1.04	28.2

Table 8 Factor Loadings for Three- and Four-Factor Solutions on 4th Grade Test A

Factor Loadings for Three-Factor Solution on 4th Grade Test A				Factor Loadings for Four-Factor Solution on 4th Grade Test A					
Item	1	2	3	Item	1	2	3	4	CCSS Strand
2	**0.394**	0.201	-0.025	2	**0.375**	0.209	-0.043	0.05	RI
3	0.336	0.055	0.270	3	0.306	0.034	0.233	0.297	RL
4	**0.593**	-0.161	0.086	4	**0.606**	-0.147	0.020	0.028	RL
5	**0.402**	-0.136	0.209	5	**0.442**	-0.113	0.158	-0.059	RL
6	**0.672**	-0.134	0.031	6	**0.706**	-0.101	-0.036	-0.119	RL
7	0.195	-0.207	**0.448**	7	0.158	-0.256	0.295	**0.492**	RL
10	0.275	0.334	-0.135	10	0.205	0.315	-0.132	0.221	L
11	**0.518**	0.109	-0.226	11	**0.448**	0.090	-0.257	0.228	RL
12	0.115	**0.452**	-0.019	12	0.092	**0.461**	0.006	0.011	RL
13	**0.470**	0.047	0.088	13	**0.510**	0.086	0.050	-0.169	L
14	0.311	**0.375**	-0.064	14	0.254	0.260	-0.070	0.205	W
15	**0.507**	-0.03	0.199	15	**0.528**	-0.011	0.146	0.014	W
16	-0.100	**0.548**	-0.019	16	-0.169	**0.526**	0.030	0.220	RI
17	0.180	0.077	**0.474**	17	0.229	0.098	**0.448**	-0.013	RI
18	-0.081	**0.544**	-0.019	18	-0.133	**0.533**	0.028	0.133	RI
19	0.047	**0.471**	-0.015	19	0.068	**0.509**	0.020	-0.227	L
20	0.119	0.288	-0.044	20	0.079	0.28	-0.032	0.121	RI
21	0.042	0.217	0.282	21	0.041	0.215	0.284	0.111	RI
22	0.195	0.327	0.117	22	0.172	0.327	0.120	0.113	L
23	0.233	0.136	0.191	23	0.319	0.199	0.180	-0.377	RI
24	-0.098	-0.023	**0.718**	24	-0.023	-0.007	**0.700**	0.007	RI
25	-0.129	**0.404**	0.274	25	-0.118	0.311	**0.402**	0.054	W
27	-0.007	**0.375**	0.050	27	-0.006	**0.391**	0.078	-0.065	W
28	**0.377**	0.340	-0.139	28	0.263	**0.363**	-0.140	-0.074	L
29	-0.187	**0.476**	0.245	29	-0.177	**0.489**	0.289	-0.033	W
30	-0.019	0.339	-0.033	30	-0.155	0.263	-0.012	**0.621**	W

L = Language; RI = Reading Informational Text; RL = Reading Literature; W = Writing

Table 9 Factor Loadings for Three- and Four-Factor Solutions on 4th Grade Test B

Factor Loadings for Three-Factor Solution on 4th Grade Test B				Factor Loadings for Four-Factor Solution on 4th Grade Test B					
Item	1	2	3	Item	1	2	3	4	CCSS Strand
3	**0.684**	-0.116	-0.106	3	**0.673**	-0.111	-0.030	-0.095	RI
4	**0.593**	-0.119	-0.039	4	**0.567**	-0.132	0.040	-0.046	L
5	**0.636**	-0.099	-0.068	5	**0.731**	0.018	-0.315	0.011	RI
6	**0.613**	0.022	0.086	6	**0.620**	0.036	0.017	0.083	RI
7	**0.355**	0.112	0.049	7	**0.433**	0.196	-0.19	0.096	W
8	**0.362**	0.329	-0.083	8	**0.365**	0.323	0.000	-0.076	W
9	-0.190	0.316	**0.38**	9	-0.046	**0.447**	-0.209	0.323	RI
10	0.331	-0.147	**0.425**	10	0.335	-0.084	0.003	**0.414**	RI
11	**0.522**	-0.037	0.003	11	**0.507**	-0.044	0.037	-0.004	L
12	0.281	-0.125	**0.496**	12	0.286	-0.118	0.177	**0.441**	RL
13	0.053	0.052	**0.395**	13	-0.106	-0.127	**0.626**	0.236	RL
14	0.128	0.166	**0.357**	14	0.091	0.116	0.275	0.285	L
15	**0.410**	0.316	0.023	15	0.345	0.238	0.236	-0.029	RI
16	-0.112	**0.558**	0.176	16	-0.124	**0.515**	0.186	0.133	RI
17	**0.495**	-0.046	0.067	17	**0.485**	-0.048	0.047	0.056	L
18	**0.501**	0.138	0.074	18	**0.389**	0.017	0.269	-0.013	RL
19	0.284	0.330	0.004	19	0.228	0.259	0.206	-0.041	RL
20	-0.019	**0.421**	0.058	20	-0.197	0.21	**0.589**	-0.080	RL
21	**0.502**	0.301	-0.18	21	**0.390**	0.176	0.286	-0.238	W
22	**0.399**	-0.015	0.093	22	0.280	-0.138	**0.375**	0.002	W
24	0.138	0.257	0.177	24	0.141	0.248	0.098	0.152	W
25	0.038	0.146	0.281	25	-0.085	0.003	**0.489**	0.159	L
26	**0.412**	-0.016	0.138	26	**0.460**	0.041	-0.085	0.156	W
28	-0.097	**0.763**	-0.138	28	-0.082	**0.746**	0.001	-0.127	RL
29	-0.087	**0.657**	0.044	29	-0.026	**0.691**	-0.066	0.066	RL
30	-0.197	-0.018	**0.746**	30	-0.156	0.012	0.194	**0.681**	W

L = Language; RI = Reading Informational Text; RL = Reading Literature; W = Writing

Table 10 Factor Loadings for Three- and Four-Factor Solutions on 8th Grade Test A

Factor Loadings for Three-Factor Solution on 8th Grade Test A				Factor Loadings for Four-Factor Solution on 8th Grade Test A					
Item	1	2	3	Item	1	2	3	4	CCSS Strand
3	**0.482**	0.168	-0.147	3	**0.512**	0.275	-0.006	-0.333	W
4	**0.640**	0.076	-0.072	4	**0.627**	0.135	-0.093	-0.031	RI
5	**0.651**	-0.053	-0.050	5	**0.649**	0.019	-0.066	-0.078	L
6	**0.373**	0.234	-0.026	6	**0.359**	0.253	-0.031	0.020	RI
7	**0.630**	-0.037	0.019	7	**0.627**	0.018	-0.019	-0.020	W
9	0.014	**0.465**	-0.083	9	0.014	**0.473**	0.017	-0.093	W
10	0.333	-0.013	0.243	10	0.339	-0.020	0.182	0.093	W
11	**0.384**	-0.021	0.318	11	**0.409**	-0.015	0.286	0.027	RI
12	0.277	0.030	0.316	12	0.272	-0.011	0.211	0.212	RI
13	0.236	-0.212	0.347	13	0.291	-0.187	**0.364**	-0.107	RL
14	**0.483**	-0.047	0.260	14	**0.477**	-0.056	0.151	0.166	RL
15	0.218	-0.061	**0.461**	15	0.242	-0.100	**0.384**	0.144	RL
16	0.078	0.070	**0.689**	16	-0.010	-0.078	-0.052	0.280	RL
17	-0.281	**0.470**	0.175	17	-0.303	**0.373**	0.149	0.237	RI
18	-0.083	-0.025	**0.586**	18	-0.016	-0.067	**0.610**	-0.009	RI
19	**0.354**	0.289	0.108	19	0.340	0.280	0.077	0.104	RL
20	-0.006	0.268	0.270	20	0.006	0.224	0.275	0.082	RL
21	0.002	**0.471**	0.057	21	-0.074	0.201	-0.122	0.340	RL
22	0.071	0.278	0.082	22	0.075	0.268	0.115	0.001	W
23	-0.075	-0.018	**0.423**	23	-0.078	-0.110	0.296	0.303	L
24	0.175	0.333	0.163	24	0.168	0.302	0.145	0.114	W
25	0.123	**0.476**	-0.093	25	0.101	**0.470**	-0.054	0.026	W
26	0.188	**0.583**	-0.116	26	0.151	**0.567**	-0.098	0.089	W
27	-0.291	0.304	**0.370**	27	-0.204	0.299	**0.569**	-0.276	RL
28	0.005	-0.009	**0.530**	28	0.047	-0.062	**0.498**	0.098	RL
29	0.126	**0.586**	-0.227	29	0.095	**0.592**	-0.159	-0.013	RI
30	0.108	0.159	**0.512**	30	0.045	0.047	0.015	0.299	RI
31	-0.199	0.337	0.322	31	-0.145	0.312	**0.445**	-0.121	RI
32	0.259	**0.409**	0.010	32	0.226	**0.384**	-0.027	0.157	L
33	0.138	-0.068	**0.357**	33	0.184	-0.069	**0.371**	-0.048	RI
34	-0.024	**0.473**	0.117	34	-0.079	0.342	0.006	**0.378**	W

L = Language; RI = Reading Informational Text; RL = Reading Literature; W = Writing

Table 11 Factor Loadings for Three- and Four-Factor Solutions on 8th Grade Test B

Factor Loadings for Three-Factor Solution on 8th Grade Test B				Factor Loadings for Four-Factor Solution on 8th Grade Test B					
Item	1	2	3	Item	1	2	3	4	CCSS Strand
1	-0.119	0.137	**0.513**	1	0.175	0.192	**0.358**	-0.226	RI
2	0.160	0.255	0.224	2	0.172	0.243	0.166	0.093	RI
3	-0.167	-0.087	**0.696**	3	0.264	0.000	**0.489**	-0.373	RI
4	0.054	0.106	**0.512**	4	-0.036	-0.037	**0.665**	0.140	RI
5	0.162	-0.016	**0.447**	5	0.339	0.017	0.306	-0.066	RL
6	-0.035	0.205	**0.404**	6	-0.064	0.109	**0.499**	0.068	L
7	**0.447**	-0.137	0.298	7	**0.608**	-0.051	0.054	0.013	RL
8	**0.368**	-0.014	0.127	8	0.102	-0.146	0.272	0.343	W
9	**0.689**	-0.257	0.174	9	**0.719**	-0.194	-0.043	0.173	W
10	**0.652**	-0.157	0.021	10	**0.579**	-0.125	-0.126	0.248	L
11	0.309	0.151	0.222	11	**0.455**	0.233	-0.007	0.016	W
12	0.190	0.111	**0.359**	12	0.342	0.154	0.200	-0.024	RI
13	0.322	-0.091	**0.395**	13	**0.381**	-0.097	0.307	0.068	L
15	0.247	0.021	0.228	15	**0.549**	0.180	-0.110	-0.149	W
16	-0.094	0.145	**0.476**	16	-0.184	-0.008	**0.671**	0.095	L
17	-0.076	**0.363**	0.229	17	-0.017	0.350	0.200	-0.007	RL
18	0.042	**0.480**	-0.081	18	-0.132	**0.412**	-0.003	0.218	RL
19	0.286	0.029	0.252	19	0.155	-0.059	0.317	0.220	RL
20	0.298	0.146	0.196	20	**0.379**	0.194	0.030	0.064	RL
21	-0.053	**0.634**	0.066	21	-0.006	**0.655**	-0.023	0.039	RL
22	-0.087	**0.535**	0.158	22	0.145	**0.637**	-0.066	-0.131	RL
23	-0.122	**0.677**	0.084	23	-0.041	**0.706**	-0.014	-0.004	RL
24	0.316	0.254	0.001	24	0.025	0.133	0.133	**0.376**	RI
25	**0.477**	0.262	-0.120	25	-0.059	0.044	0.163	**0.621**	RI
26	0.129	0.272	0.108	26	0.012	0.206	0.160	0.182	W
27	**0.547**	0.138	-0.267	27	0.165	0.044	-0.171	**0.490**	W
28	**0.431**	0.288	-0.215	28	-0.040	0.121	0.000	**0.550**	RI
29	**0.619**	0.066	-0.110	29	0.265	-0.036	-0.027	**0.491**	RI
30	0.104	**0.499**	0.022	30	-0.079	**0.414**	0.104	0.256	L
31	**0.426**	0.288	-0.173	31	0.303	0.311	-0.280	0.267	W
32	-0.040	**0.499**	0.069	32	-0.098	**0.458**	0.093	0.110	W
33	**0.416**	-0.099	0.193	33	**0.391**	-0.102	0.113	0.153	L
34	0.190	0.249	0.173	34	0.043	0.160	0.247	0.228	W

L = Language; RI = Reading Informational Text; RL = Reading Literature; W = Writing

Table 12 Factor Loadings for Three- and Four-Factor Solutions on 10th Grade Test A

Factor Loadings for Three-Factor Solution on 10th Grade Test A				Factor Loadings for Four-Factor Solution on 10th Grade Test A					
Item	1	2	3	Item	1	2	3	4	CCSS Strand
1	**0.398**	-0.159	0.272	1	**0.366**	-0.178	0.172	0.249	RL
2	**0.636**	-0.062	-0.133	2	**0.644**	-0.078	-0.107	-0.042	RL
3	**0.643**	-0.142	0.064	3	**0.657**	-0.129	0.157	-0.172	RL
4	**0.593**	-0.180	0.082	4	**0.599**	-0.173	0.140	-0.098	L
5	**0.351**	-0.022	0.325	5	0.342	-0.011	0.326	0.031	L
6	0.263	-0.081	0.325	6	0.270	-0.046	**0.409**	-0.150	W
7	**0.534**	0.011	-0.028	7	**0.527**	-0.015	-0.069	0.107	W
9	**0.428**	0.175	-0.109	9	**0.427**	0.145	-0.151	0.101	W
10	**0.444**	-0.012	0.205	10	**0.418**	-0.036	0.110	0.235	L
11	**0.399**	0.197	0.012	11	**0.402**	0.183	0.006	0.031	W
12	0.163	0.129	0.174	12	0.109	0.066	-0.081	**0.569**	RL
13	0.022	-0.035	**0.415**	13	-0.063	-0.103	0.060	**0.792**	RL
14	**0.355**	0.230	0.032	14	0.348	0.206	-0.020	0.131	RL
15	0.315	0.268	-0.098	15	0.322	0.249	-0.102	0.018	W
16	0.044	-0.001	**0.609**	16	0.031	0.037	**0.622**	0.012	RL
17	-0.110	-0.018	**0.716**	17	-0.130	0.025	**0.712**	0.051	L
18	-0.044	0.307	0.259	18	-0.047	0.316	0.245	0.048	RI
19	0.227	**0.384**	0.193	19	0.223	0.280	0.170	0.074	RI
20	-0.139	**0.390**	0.256	20	-0.145	**0.402**	0.330	0.079	RI
21	0.157	**0.436**	-0.145	21	0.190	**0.443**	-0.047	-0.209	RI
22	0.294	**0.560**	-0.187	22	0.315	**0.539**	-0.167	-0.037	RI
24	0.073	**0.364**	0.228	24	0.065	**0.357**	0.179	0.129	L
25	-0.052	0.260	0.177	25	-0.043	0.278	0.216	-0.071	RI
26	0.033	**0.441**	-0.161	26	0.038	**0.413**	-0.204	0.088	W
28	0.075	**0.463**	0.173	28	0.093	**0.480**	0.232	-0.109	L
29	-0.174	**0.565**	0.072	29	-0.156	**0.574**	0.105	-0.066	RI
30	-0.247	**0.507**	0.037	30	-0.239	**0.506**	0.026	0.024	L

L = Language; RI = Reading Informational Text; RL = Reading Literature; W = Writing

Table 13 Factor Loadings for Three- and Four-Factor Solutions on 10th Grade Test B

Factor Loadings for Three-Factor Solution on 10th Grade Test B				Factor Loadings for Four-Factor Solution on 10th Grade Test B					
Item	1	2	3	Item	1	2	3	4	CCSS Strand
3	**0.534**	-0.036	-0.034	3	**0.487**	-0.220	0.238	0.058	L
4	**0.393**	0.181	-0.182	4	**0.659**	-0.149	0.013	-0.032	RL
5	**0.615**	-0.338	0.274	5	-0.131	-0.053	**0.665**	0.134	RL
6	**0.471**	-0.009	0.003	6	0.246	-0.003	0.325	-0.019	L
7	**0.463**	-0.014	0.114	7	0.178	0.012	**0.358**	0.082	RL
9	0.194	0.349	-0.119	9	**0.605**	-0.024	-0.135	0.040	RI
10	**0.374**	0.214	-0.12	10	**0.519**	-0.008	0.084	-0.037	RI
11	**0.533**	-0.004	-0.209	11	**0.461**	-0.095	0.255	-0.177	RI
12	**0.370**	0.087	-0.035	12	0.057	0.251	0.349	-0.161	RI
13	0.251	**0.362**	-0.123	13	**0.500**	0.121	-0.011	-0.047	W
14	0.140	0.215	0.055	14	0.186	0.128	0.053	0.068	W
16	**0.360**	0.058	0.096	16	0.306	-0.072	0.184	0.151	W
17	**0.472**	-0.031	0.112	17	0.01	0.158	**0.469**	-0.014	RI
18	0.127	0.309	0.109	18	0.019	**0.373**	0.155	0.018	RI
19	0.136	**0.474**	-0.024	19	0.341	0.311	-0.016	-0.007	L
20	0.097	0.308	0.066	20	-0.032	**0.419**	0.155	-0.051	W
21	**0.384**	0.007	0.117	21	-0.043	0.206	**0.417**	-0.016	W
22	0.129	0.267	0.186	22	0.040	0.276	0.138	0.135	W
23	0.338	0.183	0.048	23	0.191	0.191	0.247	0.001	RL
24	-0.034	0.319	0.144	24	0.265	0.055	-0.158	0.253	RL
25	0.132	**0.546**	0.018	25	0.339	**0.380**	-0.010	0.026	RI
26	-0.073	**0.379**	0.350	26	-0.037	0.320	-0.005	0.332	RI
27	0.091	0.077	**0.629**	27	0.016	-0.098	0.097	**0.721**	RI
28	-0.035	0.198	**0.637**	28	-0.059	0.047	0.028	**0.702**	RL
29	-0.117	**0.366**	0.152	29	-0.292	**0.593**	0.104	-0.034	RL
30	-0.016	**0.617**	0.008	30	0.293	**0.434**	-0.124	0.021	L
31	-0.138	**0.603**	0.048	31	0.161	**0.467**	-0.167	0.042	L
32	-0.035	0.345	0.102	32	-0.123	**0.469**	0.081	-0.023	W

L = Language; RI = Reading Informational Text; RL = Reading Literature; W = Writing

CHAPTER V: DISCUSSION

This study was an endeavor to use exploratory factor analysis (EFA), classical test theory (CTT), and item response theory (IRT) to investigate if factors aligned to the three instructional shifts of CCSS can be confirmed within benchmark assessments designed to measure student progress across three grade levels: 4th, 8th, and 10th. This chapter delivers an overview of the results, interpretation of the findings, limitations, and recommendations for future research. The study addressed three research questions:

1. Do the assessments have sound psychometric properties, such as reliability and validity?

2. Do the assessments show three- or four-factor solutions through factor analysis?

3. Are the benchmark assessments used by TN schools/students aligned to the shifts or strands of ELA CCSS?

The researcher hypothesized that the assessments would have sound psychometric properties, show a three-factor solution, and align to the shifts of CCSS.

To address the first question, "Do the assessments have sound psychometric properties, such as reliability and validity?," Cronbach's alpha was computed for each assessment confirm reliability of the benchmark tests as an internal consistency reliability index. Nunnaly (1978) has indicated 0.7 to be an acceptable reliability coefficient and all six tests exceeded that limit: 4A (a = 0.803), 4B (a = 0.851), 8A (a = 0.839), 8B (a =

0.881), 10A (a = 0.825), and 10B (a = 0.804). We can conclude that all six tests are reliable measurements.

Additionally, to confirm concurrent validity across each grade level series, Pearson's correlation coefficient (r) was computed using total raw test scores for students who took both A and B tests for each grade level. Because each grade level's tests A and B had an r between .60-.79, we can conclude that student performance on each test series is strongly correlated. Therefore, we can conclude that all sets of tests are valid. Thus, the first hypothesis was confirmed.

To address the second research question, "Do the assessments show three- or four-factor solutions through factor analysis?," EFA was conducted to test for whether a three- or four-factor solution fit the data better. We hypothesized that if the data fit a three-solution factor, it could indicate alignment with the three instructional shifts of ELA CCSS but that if the data fit a four-solution factor, it could indicate alignment to the four tested strands of ELA CCSS. The purpose of the study was to investigate if factors aligned to the three instructional shifts of CCSS can be confirmed within the studied benchmark assessments, so we anticipated alignment with the three shifts.

Initial EFA analyses were inconclusive. For example, when examining the eigenvalues, all data sets exhibited at least seven components with the eigenvalues greater than 1 (William, Brown, & Onsman, 2012). However, the scree plots demonstrated a tendency towards one-factor solutions because the slope of the curves reached a steady rate from the second factor. Additionally, factor loadings for each data

set revealed that more items loaded on a factor when testing for a three-factor solution than when testing for a four-factor solution.

Because the results were unclear, but demonstrated a promising tendency towards a three-factor solution, CTT and IRT were utilized to remove misfit items before conducting a second round of EFA. In general, item discriminations and item difficulties in CTT demonstrated similar characteristics to the estimates in IRT. Items were identified for removal based on two index statistics: item-test correlations (ITC) and discriminant coefficients (a-values), which both indicate an item's ability to discriminate between test taker ability.

The removal of misfit items from the second EFA did not result in any significant changes in eigenvalues or cumulative percentages. However, for all six data sets, the number of items that loaded onto factors when testing for both three- and four-factor solutions increased with slightly more items loading when testing the three-factor solution, indicating a better alignment of the data to three factors. However, overall the differences are minimal and the data do not clearly show either a three- or four-factor solution. Thus, the second hypothesis was neither confirmed nor rejected.

The final research question, "Are the benchmark assessments used by TN schools/students aligned to the shifts or strands of ELA CCSS?," is more difficult to answer. Because the answer to the second research question is inconclusive, we cannot ascertain whether or not the assessments are truly aligned to either the shifts or the strands. There is evidence that the assessments are more aligned to a three-factor solution and, therefore, more closely aligned to the shifts of ELA CCSS than the strands

of ELA CCSS. However, without access to the specific item questions, it is not possible to visibly check for item alignment to the shifts.

On the other hand, the researchers did have access to each item's strand alignment, which do not reveal any consistency between factor loadings and strands (i.e. Language, Reading Informational Text, Reading Literature, and Writing) when testing the four-factor solution on all six data sets. Tables 8-13 show the ELA CCSS strand for each item compared to their loading factors. Even if the data fit a four-factor solution, the items do not align with the strands of ELA CCSS. Therefore, the third hypothesis was neither confirmed nor rejected.

It must also be acknowledged that the lack of evidence to support a strong fit to either a three- or four-factor solution may indicate that the assessments are actually more aligned to a one-factor solution, as the scree plots from the first round of EFA indicated. If we accept that the assessments align to a one-factor solution, we can anticipate the assessments are mostly likely aligned to a student achievement or reading comprehension construct. The unidimensional data suggests that students perform roughly the same across items, indicating a single latent factor. However, for benchmarks designed to assess specific strands of standards founded on three specific shifts, that conclusion would is disappointing. The purpose of the assessments is to assist teachers to identify areas of improvement for their students. If the items do not align with either the strands or the shifts, targeting skills for differentiated instruction is not possible. These assessments would only be useful for assessing overall reading comprehension, but not for improving instruction.

Limitations and Recommendations for Future Research

Every study has its limitations and this one is no exception. The most salient limitation is the researcher's lack of access to the actual test items and reading passages. Without being able to fully analyze the content of the tests, it is not possible to know which of the three instructional shifts of ELA CCSS to which each item may possibly align. Even if we consider the stronger alignment to a three-factor solution to be indicative of a finding, we cannot know for sure if that tendency means anything. We will continue exploring other data sets with available test items so that more information about the test items is available for better interpretation.

Additionally, while the data set was robust in number, there was inconsistency in administration dates of each assessment across the state. For example, some students began the test series in August, 2014 while others began in October, 2014. Though there are clear differences in time between the students who took both tests, there could be a confluence in the data based on how much knowledge students gained between August and October. For future studies, we will explore the comparison of data sets in which students took the tests during more consistent time frames.

A final limitation worth mentioning is that there may exist confluence in student performance based on the nested nature of the data. Each reading benchmark assessment included questions related to specific passages. It is possible that the data may not reflect

underlying factors, such as the three key shifts of ELA CCSS, because student performance is more related to passage comprehension. Future studies could account this possibility by considering testlet effects (Wainer & Kiely, 1987).

Benchmark assessments will continue to be utilized in schools as a way to measure student proficiency as the school year progresses. Therefore, it is important that we continue to study their effectiveness and ensure that they actually assess what they are designed to assess. Further research is needed to investigate the alignment of these assessments not just to the standards but to the theoretical underpinnings of the standards. Additionally, because benchmark assessments are only valuable as a tool if used properly, research is needed to evaluate how teachers are utilizing these assessments to improve student outcomes. As the larger test consortia like PARCC begin to see results on the benchmark assessments aligned to their end-of-course tests, it will be valuable to see how predictive the data truly is and how teachers are accessing the data to guide instruction.

While it is disappointing that our study has not been successful in confirming factors aligned to the three instructional shifts of CCSS within these benchmark assessments, we have uncovered some insight into the possibility that they measure only one factor, which may represent reading ability. If the benchmark assessments only measure reading ability, then further research is needed to determine if students perform similarly on state summative assessments to confirm alignment between benchmark assessments and other CCSS-aligned tests. If they are aligned, perhaps a one-factor solution for reading ability is all that is needed to assess student progress.

REFERENCES

Achieve, Inc., The Education Trust, & The Thomas B. Fordham Foundation. (2004). *Ready or Not: Creating a High School Diploma that Counts.* Washington, DC: American Diploma Project.

Achieve, Inc. (2007). *Closing the expectations gap 2007: An annual 50-state progress report on the alignment of high school policies with the demands of college and work.* Washington, DC: Author. Retrieved from http://www.achieve.org/files/50-state-07-Final.pdf

ACT, Inc. (2006). *Reading between the lines: What the ACT reveals about college readiness in reading.* Washington, DC: Author.

ACT, Inc. (2009). *ACT National Curriculum Survey 2009*. Iowa City, IA: Author.

Applebee, A., & Langer, J. (2006). *The state of writing instruction in America's schools: What existing data tell us.* Center on English Learning and Achievement. Retrieved from http://www.albany.edu/aire/news/State%20of%20Writing%20Instruction.pdf.

Baker, L., Dreher, M. J., Shiplet, A. K., Beall, L.C., Voelker, A.N., Garrett, A. J., Schugar, H. R., & Finger-Elam, M. (2011). Children's comprehension of informational text: reading, engaging, and learning. *International Electronic Journal of Elementary Education*, *4*(1), 197-227.

Ballentine, S. (2014, April 28). Indiana approves Common Core replacement standards. *Huffington Post*. Retrieved from http://www.huffingtonpost.com/2014/04/28/indiana-common-core-replacement_n_5228212.html.

Baumann, J. F., & Kameenui, E. J. (1991). Research on vocabulary instruction: Ode to Voltaire. In J. Flood, J. M. Jensen, D. Lapp, & J. R. Squire (Eds.), *Handbook of research on teaching the English language arts* (pp. 604–632). New York, NY: Macmillan.

Beavers, A.S., Lounsbury, J.W., Richards, J.K., Huck, S.W., Skolits, G.J., & Esquivel, S.L. (2013). Partial considerations for using exploratory factor analysis in educational research. *Practical Assessment, Research & Evaluation, 18*(6). Retrieved from http://pareonline.net/getvn.asp?v=18&n=6.

Beck, I. L., McKeown, M. G., & Kucan, L. (2002). *Bringing words to life: Robust vocabulary instruction.* New York, NY: Guilford.

Beck, I. L., McKeown, M. G., & Kucan, L. (2008). *Creating robust vocabulary: Frequently asked questions and extended examples*. New York, NY: Guilford.

Becker, W. C. (1977). Teaching reading and language to the disadvantaged—what we have learned from field research. *Harvard Educational Review, 47,* 518–543.

Bergan, J.R., Bergan, J.R., & Burnham, C.G. (2009). *Benchmark assessment in standards-based education*. (White paper). Retrieved from http://www.ati-online.com/pdfs/researchK12/BenchmarkAssessment.pdf

Binks-Cantrell, E., Joshi, R.M., & Washburn, E. K. (2012). Validation of an instrument for assessing teacher knowledge of basic language constructs of literacy. *Annals of Dyslexia, 62,* 153-171.

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

Bos, C., Mather, N., Dickson, S., Podhajski, B., & Chard, D. (2001). Perceptions and knowledge of preservice and inservice educators about early reading instruction. *Annals of Dyslexia, 51*, 97–120.

Burke, K. (2010). *Balanced assessment: From formative to summative*. Bloomington, IN: Solution Tree Press.

Cohen, D. K., & Bhatt, M.P. (2012). The importance of infrastructure development to high-quality literacy instruction. *The Future of Children, 22(*2), 117-138.

Common Core State Standards Institute. (2014). *Frequently asked questions*. Retrieved from http://www.corestandards.org/wp-content/uploads/FAQs.pdf.

Conley, D.T., & Gaston, P.L. (2013). A path to alignment: connecting K-12 and higher education via the Common Core and the Degree Qualifications Profile. Indianapolis, IN: Lumina Foundation. Retrieved from http://www.luminafoundation.org/publications/DQP/A_path_to_alignment.pdf.

Conway, J.M, & Huffcutt, A.I. (2003). A review and evaluation of exploratory factor analysis practice in organizational research. *Organizational Research Methods, 6* (2), 147-168.

Correia, M.P. (2011). Fiction vs. informational texts: which will kindergarteners choose? *Young Children, 66*(6), 100-104.

Cristol, K., & Ramsey, B. (2014). *Common Core in the districts: an early look at early implementers.* Washington, DC.: Education First and Thomas B. Fordham Institute. Retrieved from http://www.edexcellence.net/sites/default/files/publication/pdfs/Common-Core-In-The-Districts-Full-Report_0.pdf.

Cronbach, L. J., & Meehl, P. E. (1955). Construct validity in psychological tests. *Psychological Bulletin, 52*(4), 281-302.

Darling-Hammond, L. (2007). Standards, accountability, and school reform. In Christine Sleeter (ed.), *Facing Accountability in Education: Democracy and Equity at Risk*, pp. 78-111. NY: Teachers College Press.

De Ayala, R. J. (2009). *Theory and practice of item response theory*. New York, NY: The Guilford Press.

Dreher, M. J (2003). Motivating struggling readers by tapping the potential of information books. *Reading & Writing Quarterly, 19*, 25-38.

Dunn, K., & Mulvenon, S.W. (2009). A critical review of research on formative assessment: the limited scientific evidence of the impact of formative assessment in education. *Practical Assessment, Research & Evaluation, 14*(7), 1-11.

Ebel, R.L., & Frisbie, D.A. (1986). *Essentials of educational measurement.* Englewood Cliffs, NJ: Prentice-Hall.

EPE Research Center. (2013). *Findings from a national survey of teacher perspectives on the Common Core.* Retrieved from http://www.edweek.org/media/epe_survey_teacher_perspectives_common_core_2013.pdf.

Evans, J.D. (1996). *Straightforward statistics for the behavioral sciences*. Pacific Grove, CA: Brooks/Cole Publishing.

Fan, X. (1998). Item response theory and classical test theory: an empirical comparison of their item/person statistics. *Educational and Psychological Measurement, 58*(3), 357-382.

Fang, Z. & Pace, B.G. (2013). Teaching with challenging texts in the disciplines: text complexity and close reading. *Journal of Adolescent and Adult Literacy, 57*(2), 104-108.

Faria, A., Heppen, J. Yibing, L., Stachel, S., Jones, W., Sawyer, K., Thomsen, K., Kutner, M., Miser, D., Lewis, S., Casserly, M., Simon, C., Uzzell, R., Corcoran, A., & Palacios, M. (2012). *Charting success: data use and student achievement in urban schools.* Washington, DC: Council of the Great City Schools. Retrieved from http://www.cgcs.org/cms/lib/DC00001581/Centricity/Domain/87/Charting_Success.pdf.

Fisher, D., & Frey, N. (2012). Text-dependent questions. *Principal Leadership, 13*(1), 70-73.

Graham, S., & Hebert, M. A. (2010). *Writing to read: Evidence for how writing can improve reading. A Carnegie Corporation Time to Act Report*. Washington, DC: Alliance for Excellent Education.

Greenstein, L. (2010). *What teachers really need to know about formative assessment*. Alexandria, VA: Association for Supervision and Curriculum Development.

Hambleton, R. K., & van der Linden, W. L. (1982). Advances in item response theory and applications: An introduction. *Applied Psychological Measurement, 6*, 373-378.

Hambleton, R. K., Swaminathan, H., & Rogers, H. J. (1991). *Fundamentals of Item Response Theory (Measurement Methods for Social Sciences).* California: Sage Publication.

Hambleton, R. K., & Jones, R. W. (1993). Comparison of classical test theory and item response theory and their applications to test development. *Educational Measurement: Issues and Practice, 12*(3), 37-47.

Hayes, D., & Ward, M. (1992, December). *Learning from texts: effects of similar and dissimilar features of analogies in study guides.* Paper presented at the 42nd Annual Meeting of the National Reading Conference. San Antonio: Education Trust.

Heller, R., & Greenleaf, C. L. (2007). *Literacy instruction in the content areas: Getting to the core of middle & high school improvement*. Washington, DC: Alliance for Excellent Education. Retrieved from www.all4ed.org.

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.

Herman, J.L., Osmundson, E., & Dietel, R. (2010). Benchmark assessment for improved learning (AACC Report). Los Angeles, CA: University of California.

Jochim, A. & Lavery, L. (2015). The evolving politics of the Common Core. *Issues in Governance Studies, 69,* 1-13. Retrieved from http://www.brookings.edu/~/media/research/files/papers/2015/05/07-common-core-politics-jochim-lavery/common_core.pdf.

Kerlinger, F. N., & Lee, H. B. (2000). *Foundations of behavioral research*. Belmont, CA: Cengage Learning.

Kim, J. K., & Nicewander, W. A. (1993). Ability estimation for conventional tests. *Psychometrika, 58,* 587-599.

Kline, T. (2005). *Psychological testing: a practical approach to design and evaluation.* Thousand Oaks, CA: Sage Publications, Inc.

Liebtag, E. (2013). Moving forward with Common Core State Standards implementation: possibilities and potential problems. *Journal of Curriculum and Instruction, 7*(2), 56-70.

Leung, C., & Mohan, B. (2004). Teacher formative assessment to support student learning in classroom contexts: Assessment as discourse and assessment of discourse. *Language Testing, 21*(3), 335-359.

Maloch, B., & Bomer, R. (2013). Teaching about and with informational texts: what does research teach us? *Language Arts, 90*(6), 441-450.

Masters, G.N. (1982). A Rasch model for partial credit scoring. *Psychometrika*, *47*, 149-174.

Mathis, W. (2012). *Research-based options for education policymaking: Common Core State Standards.* Boulder, Co: University of Colorado Boulder.

Milewski, G. B., Johnson, D., Glazer, N., & Kubota, M. (2005). *A survey to evaluate the alignment of the new SAT Writing and Critical Reading sections to curricula and instructional practices* (College Board Research Report No. 2005-1 /ETS RR-05-07). New York, NY: College Entrance Examination Board.

Moats, L. C. (1994). The missing foundation in teacher education: Knowledge of the structure of spoken and written language. *Annals of Dyslexia, 44*, 81–102.

Moss, B., & Newton, E. (2002). An examination of the informational text genre in basal readers. *Reading Psychology, 23*(1), 1–13.

Muraki, E. (1993). Information functions of the generalized partial credit model. *Applied Psychological Measurement, 17*, 351-363.

National Assessment of Educational Progress. (2013). Tennessee state profile. Retrieved from http://nces.ed.gov/nationsreportcard/states/.

National Center for Education Statistics. (2015). Nation's report card: Tennessee state profile. Retrieved from http://nces.ed.gov/nationsreportcard/states/Default.aspx?st=TN.

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

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

No Child Left Behind (NCLB) Act of 2001, 20 U.S.C.A. § 6301 et seq.

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

Olson, L. (2005). Benchmark assessments offer regular checkups on student achievement. *Education Week, 25*(13), 13-14.

Pepper, M., Burns, S., Kelly, T., & Warach, K. (2013). Tennessee teachers' perceptions of Common Core State Standards (research brief). Retrieved from http://news.vanderbilt.edu/files/RESULTS-Tennessee_Teachers_Perceptions_of_Common_Core_State_Standards.pdf.

Popham, W.J. (2008). *Transformative assessment*. Alexandria, VA: Association for Supervision and Curriculum Development.

Popham, W.J. (2010). *Everything school leaders need to know about assessment*. Thousand Oaks, CA: Corwin Press.

Preacher, K.J., Zhang, G., Kim, C., & Mels, G. (2013). Choosing the optimal factors in exploratory factor analysis: a model selection perspective. *Multivariate Behavioral Research, 48*, 28-56.

Pritchard, M. E., Wilson, G. S., & Yamnitz, B. (2007). What predicts adjustment among college students? *Journal of American College Health, 56*, 15-21.

Scholastic. (2014). *Primary sources: America's teachers on teaching in an era of change* (3rd ed.). Retrieved from http://www.scholastic.com/primarysources/PrimarySources3rdEdition.pdf.

Scriven, M. (1967). The methodology of evaluation. In R.W. Tyler, R.M. Gagne, and M. Scriven (Eds.), *Perspectives of curriculum evaluation, Volume I* (pp. 39-83). Chicago: Rand McNally.

Shanahan, T., & Shanahan, C. (2008). Teaching disciplinary literacy to adolescents: Rethinking content-area literacy. *Harvard Educational Review, 78*(1), 40–59.

Shanahan, T., & Duffett, A. (2013). *Common Core in the schools: a first look at reading assignments*. Washington, DC: Thomas B. Fordham Institute. Retrieved from http://www.edexcellence.net/sites/default/files/publication/pdfs/20131023-Common-Core-in-the-Schools-a-First-Look-at-Reading-Assignments.pdf.

Spear-Swerling, L., & Brucker, P. O. (2003). Teachers' acquisition of knowledge about English word structure. *Annals of Dyslexia, 53*, 72–103.

Spear-Swerling, L., Brucker, P. O., & Alfano, M. P. (2005). Teachers' literacy-related knowledge and self-perceptions in relation to preparation and experience. *Annals of Dyslexia, 55*(2), 266-296.

Stanovich, K. E. (1986). Matthew effects in reading: Some consequences of individual differences in the acquisition of literacy. *Reading Research Quarterly, 21*, 360–407.

Stien, D., & Beed, P. (2004). Bridging the gap between fiction and nonfiction in the literature circle setting. *Reading Teacher, 57*, 510-518.

Stenner, A. J., Koons, H. H., & Swartz, C.W. (2009). *Text complexity, the text complexity continuum, and developing expertise in reading*. Durham, NC: MetaMetrics.

Tabachnick, B. G & Fidell, L. S. (2007). *Using multivariate statistics*. Boston: Pearson Education Inc.

Tang, K. L. (1996). *Polytomous item response theory models and their applications in large-scale testing programs: A review of literature.* (Report No. RM-96-8). Princeton, New Jersey: Educational Testing Service.

Tavaskol, M. & Dennick, R. (2011). Making sense of Cronbach's alpha. *International Journal of Medical Education, 2,* 53-55.

Tennessee Department of Education. (2013). *The Common Core State Standards: history and fact sheet*. Retrieved from http://tncore.org/sites/www/Uploads/Family/Common_Core_Facts_History.pdf.

Tennessee Department of Education. (2015). *The Tennessee educator survey report.* Retrieved from http://tn.gov/assets/entities/education/attachments/data_survey_report_2015.pdf.

Tennessee Department of Education. (2009a). *The Tennessee English Language Arts Standards: Kindergarten*. Retrieved from http://www.tn.gov/education/standards/english/ENG_Grade_K.pdf.

Tennessee Department of Education. (2009b). *The Tennessee English Language Arts Standards: Grade 1*. Retrieved from http://www.tn.gov/education/standards/english/ENG_Grade_1.pdf.

Tennessee Department of Education. (2009c). *The Tennessee English Language Arts Standards: Grade 2*. Retrieved from http://www.tn.gov/education/standards/english/ENG_Grade_2.pdf.

Tennessee Department of Education. (2009c). *The Tennessee English Language Arts Standards: Grade 3*. Retrieved from http://www.tn.gov/education/standards/english/ENG_Grade_3.pdf.

Tennessee Department of Education. (2009d). *The Tennessee English Language Arts Standards: Grade 4*. Retrieved from http://www.tn.gov/education/standards/english/ENG_Grade_4.pdf.

Tennessee Department of Education. (2009e). *The Tennessee English Language Arts Standards: Grade 5*. Retrieved from http://www.tn.gov/education/standards/english/ENG_Grade_5.pdf.

Tennessee Department of Education. (2009f). *The Tennessee English Language Arts Standards: Grade 6*. Retrieved from http://www.tn.gov/education/standards/english/ENG_Grade_6.pdf.

Tennessee Department of Education. (2009g). *The Tennessee English Language Arts Standards: Grade 7*. Retrieved from http://www.tn.gov/education/standards/english/ENG_Grade_7.pdf.

Tennessee Department of Education. (2009h). *The Tennessee English Language Arts Standards: Grade 8*. Retrieved from http://www.tn.gov/education/standards/english/ENG_Grade_8.pdf.

Tennessee Department of Education. (2009i). *The Tennessee English Language Arts Standards: English I*. Retrieved from http://www.tn.gov/education/standards/english/ENG_Grade_3001.pdf.

Tennessee Department of Education. (2009j). *The Tennessee English Language Arts Standards: English II*. Retrieved from http://www.tn.gov/education/standards/english/ENG_Grade_3002.pdf.

Tennessee Department of Education. (2009k). *The Tennessee English Language Arts Standards: English III*. Retrieved from http://www.tn.gov/education/standards/english/ENG_Grade_3003.pdf.

Tennessee Department of Education. (2009l). *The Tennessee English Language Arts Standards: English IV*. Retrieved from http://www.tn.gov/education/standards/english/ENG_Grade_3004.pdf.

Tennessee General Assembly. Joint Conference Committee. (2014). *Conference Committee Report on House Bill No. 1549 / Senate Bill No. 1835*. Retrieved from http://www.capitol.tn.gov/Bills/108/CCRReports/CC0009.pdf.

Tennessee General Assembly. (2015). *House Bill No. 1035 / Senate Bill No. 1163*. Retrieved from http://wapp.capitol.tn.gov/apps/BillInfo/Default.aspx?BillNumber=HB1035.

Tennessee Government, Office of the Press Secretary. (2014). Haslam lays out next steps from education summit. [Press release]. Retrieved from http://news.tn.gov/node/13106.

Tennessee State Board of Education. (2016). *Math and English Language Arts standards review*. Retrieved from https://www.tn.gov/sbe/article/math-and-english-language-arts.

TNCore. (n.d.). *Frequently asked questions.* Retrieved from http://www.tncore.org/faqs.aspx.

Troia, G.A. & Olinghouse, N. G. (2013). The Common Core State Standards and evidence-based writing practices: the case for writing. *School Psychology Review, 42*(3), 343-357.

U.S. Census Bureau. (2010). *2010 Census Urban and Rural Classification and Urban Area Criteria.* Retrieved May 5, 2014, from http://www.census.gov/geo/reference/ua/urban-rural-2010.html.

U.S. Department of Education. (2010). *Race to the top application for initial funding*. Retrieved from http://www2.ed.gov/programs/racetothetop/phase1-applications/tennessee.pdf.

Utah State Office of Education. (2010). *A comparison of the existing Utah language arts curricula to the Common Core State Standards for English language arts.* (White paper). Retrieved from http://www.schools.utah.gov/CURR/directors/Home/Meeting-Archive/2010-September3-6Comparison.aspx.

Wagner, R. (2014). *PARCC may still be on the table for Tennessee schools.* Times News. Retrieved from http://www.timesnews.net/article/9078740/parcc-may-still-be-on-the-table-for-tennessee-schools.

Wainer, H. & Kiely, G. (1987). Item clusters and computerized adaptive testing: A case for testlets. *Journal of Educational Measurement, 24*, 185-202.

Williams, B., Brown, T., & Onsman, A. (2012). Exploratory factor analysis: a five-step guide for novices. *Australasian Journal of Paramedicine, 8*(3). Retrieved from http://ro.ecu.edu.au/jephc/vol8/iss3/1.

Williamson, G. L. (2006). Aligning the journey with the destination: a model for K-6 reading standards (White Paper). Retrieved from http://cdn.lexile.com/m/uploads/whitepapers/AligningtheJourneywithaDestination.pdf.

Wixson, K.K. & Valencia, S.W. (2014). Suggestions and cautions for addressing text complexity. *The Reading Teacher, 67*(6), 430-434.

Wright, B.D. (1998). Rating scale model (RSM) or partial credit model (PCM)? *Rasch Measurement Transactions, 12*(3), 641-2.

Yopp, H. K., & Yopp, R. H. (2006). Primary students and informational texts. *Science and Children, 44*(3), 22–25.

Young, T. A., & Ward, B. A. (2012). Common Core and informational texts: learn how the Common Core State Standards emphasize and define informational text. *Booklist, 109*(1), 31-37.

APPENDICES

APPENDIX A: IRB APPROVAL

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

MIDDLE TENNESSEE STATE UNIVERSITY

EXEMPT APPROVAL NOTICE

9/25/2015

Investigator(s): Melissa Stugart
Department: Literacy Studies
Investigator(s) Email: mts3i@mtmail.mtsu.edu
Protocol Title: "Common Core State Standards Benchmark Assessments: Item Alignment To The Shifts"

Protocol ID: 16-1060

Dear Investigator(s),

The MTSU Institutional Review Board, or a representative of the IRB, has reviewed the research proposal identified above and this study has been designated to be EXEMPT.. The exemption is pursuant to 45 CFR 46.101(b) **(4) Collection or Study of Existing Data**

The following changes to this protocol must be reported prior to implementation:

- Addition of new subject population or exclusion of currently approved demographics
- Addition/removal of investigators
- Addition of new procedures
- Other changes that may make this study to be no longer be considered exempt

The following changes do not have to be reported:

- Editorial/administrative revisions to the consent of other study documents
- Changes to the number of subjects from the original proposal

All research materials must be retained by the PI or the faculty advisor (if the PI is a student) 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

NOTE: All necessary forms can be obtained from www.mtsu.edu/irb.