DO MULTIPLE SUBTEST EXAMS IMPACT STUDENT PERFORMANCE?

A Dissertation

Presented to

The Faculty of the Education Department

Carson-Newman University

In Partial Fulfillment

Of the

Requirements for the Degree

Doctor of Education

By

Aaron H. Wood

3/26/2018

CARSON-NEWMAN
UNIVERSITY

Dissertation Approval

Student Name/ CNU ID: Aaron Wood

Dissertation Title: *Do Multiple Subtest Exams Impact Student Performance?*

This dissertation has been approved and accepted by the faculty of the Education Department, Carson-Newman University, in partial fulfillment of the requirements for the degree, Doctor of Education.

Dissertation Committee:

Signatures: (Print and Sign)

Dr. Patricia Murphree

Dissertation Chair: Dr. Patricia Murphree

P Mark Taylor

Methodologist Member : Dr. P. Mark Taylor

Christy Walker

Content Member: Dr. Christy Walker

Approved by the Dissertation Committee Date: 3/26/2018

Abstract

The purpose of this study was to determine whether student performance is impacted by dividing an exam into multiple subtests and multiple administrations. This study was designed to determine if there is a saturation point of testing where student performance is negatively impacted. The quantitative study focused on separating exams over multiple administrations and calendar days impacted student performance on an exam. A retired ACT exam was administered to students over three or four calendar days and testing sessions. Student performance was compared using a t-test. Students who completed the exam over four administrations outperformed students who completed the exam over three administration sessions. Students with four administration sessions scored, on average, two composite points higher on the ACT exam. The study suggested that exams with multiple subparts should be administered to students over multiple calendar days when possible.

Key words: subtests, multiple administrations, ACT, saturation point, exam scheduling

Acknowledgements and Dedication

There are so many people who helped and supported me throughout this process. My friends and family have done so much to encourage, guide and reaffirm me along this entire journey. I couldn't have done this without Morgan. She was my biggest cheerleader and biggest supporter along the way. The late Dr. Hayes was a tremendous help throughout the dissertation journey and I am extremely sad to know that others may not be encouraged by her as I was. Dr. Taylor and his blunt, positive and supportive comments and suggestions brought joy into this adventure. Dr. Walker continued to reaffirm that I was on the right track and to not stress out. The entire Carson Newman staff has and will continue to influence my career throughout the courses that I have completed throughout this program. I couldn't have done any of this without a supportive and encouraging school district. I am truly blessed to call Johnson City Schools my district.

I would like to dedicate this to Mrs. Budy. She was the first teacher that encouraged me that being an overachiever is alright. To Dr. Schmalzried for unlocking my true potential as a teacher and educator. Without my family pushing me along the way I'm not sure if I ever would have completed this. Thank you Mom, Dad, Skyler, and Jordan for making me the person I am today. Finally, to Morgan for being the ultimate teammate, supporter and encourager throughout this journey.

Table of Contents

List of Tables

CHAPTER ONE

Introduction

The landscape of teaching has dramatically changed since the days of a single classroom school. The introduction of standardized testing altered the landscape of education in dramatic and irreversible ways. Over the last ten years of standardized testing in Tennessee the administration of the exams changed almost yearly for high school students. The original state exam was known as the Gateway exam and required a passing score in order for a student to receive credit for the tested course. Students were required to pass these exams in "three specific high school courses- Algebra I, English II, and Biology I" (Morgan, 2004). In 2009, the Tennessee Consolidated Assessment Program, or TCAP, phased out the Gateway exams and began administering the End of Course Exams (Zunith, 2012). According to the Tennessee Transition Policy (2010), the required exams were expanded to include English II, and US History. The exams were administered in a single test session without a time limitation. The administration of these exams was then altered to have specific time limits, but there was only a single testing session.

Background of the Study

The testing process changed when the TN Ready exams were implemented for the first time during the 2015-16 school year. These exams were composed of two subtests administered at different points of the course. For the first time, Tennessee divided the tests into multiple subparts for their students. The change to the administration of exams impacted how teachers prepared for the exams, and also changed how the students approached the exams. The research into whether the changes to the administrations of the exams impacted student performance has not been studied which is the basis of this research project.

Statement of the Problem

No research has been completed that investigates whether student performance is impacted when an exam is broken down into multiple subtests. The lack of information about proper administration structure can negatively impact students. Student performance could potentially reach a saturation point where performance on multiple subtests is impacted. If this is the case, then the TCAP End of Course Exams would have negatively impacted student performance by breaking exams down into multiple subtests. A current high school junior has the potential to take up to eleven subtests with a standard course load of English III, Algebra II, U.S. History, and Chemistry.

Purpose of the Study

The purpose of this study was to determine if student performance is impacted by dividing an exam down into multiple subtests and multiple administrations. This study was designed to determine if there is a saturation point of testing where student performance is negatively impacted. Multiple studies have focused on the impact of the addition of extended time to students, but there have been no major studies that have focused on multiple administration settings for all students. A single study focused (Waltz, et al, 2000) on the administration of tests to students with disabilities over two days of testing, but multiple sessions for all students have not been explored.

Theoretical Framework

Assessment has been used throughout education to determine teacher and student performance over a specific span of time. Individuals who administer these assessments, and also individuals who interpret the results must be certain that the assessment is valid and authentic. Authentic assessment should “require students to be effective performers with acquired

knowledge" (Wiggins, 1990, pp. 1). A student must have a proper and valid administration to allow the assessment to be an authentic representation of the student's learning. Bringing the structure of not only the exam, but the administration of that exam, into the conversation of determining the authenticity of an assessment. High stakes exams often lead teachers and students to adjust their learning strategies and goals in order to maximize test scores (Corbett & Wilson, 1988). The National Commission on Testing and Public Policy (1990), states that, "Test scores are at best an estimate of someone's knowledge or ability, and can be affected by numerous outside factors. Inevitably, some who could perform successfully will "fail" tests and thus risk being misclassified and erroneously denied opportunity" (p. 2). There are many outside factors that can impact student performance on the exam.

Research Questions

This study was conducted to research the impact that the administration window has on the authenticity of the assessment.

The study focused on two research questions:

1. Does dividing an exam into smaller subtests impact overall student performance on exams?
2. Does the length of the testing window impact student performance on exams?

Participants of the study were a group of secondary students that included students with disabilities and also general education students. The size of the study due, to resources and funding was, very limited to a small population in the high school. Also, the population of the students was representative of the school itself, but this school is not a representation of all students across the country. These two factors limited the scope of the study, but these areas also be expanded on in future studies if merited.

Limitations and Delimitations

Through the design of the study, there were assumptions that were made. Primarily there was an assumption of normalcy between the groups. Due to the fact that the results were not standardized, scores would not fall onto a standard bell curve of scores. The scores of the exam fell onto a distribution of scores, but those scores would not be normalized. The study focused on the impact of multiple subtests on scores over time. Another limitation of the study was the homogeneity of the variances within the group. The group size has similar variances, but the variances were relative to the school population. The groups were similar in design, but all variances could not be controlled in the school environment. The study intended to look into whether there was a saturation point for a student's performance on exams. The amount of sleep a student got the night before, environment of the school, time of the day when a test is administered, and even the temperature of the testing room. These factors can directly impact the authenticity of the assessment.

Definition of Terms

Saturation point. When the student has exerted as much effort as possible over an extended amount of time and after the saturation point motivation and performance decrease.

Subtest. When a single exam is broken down into multiple administrations, these are also known as subtests or subparts.

Student performance. The raw score increase of a student throughout subtests and also their total raw score after all subtests measures student performance. Student motivation was measured by a Likert scale survey given on each subtest of the exam.

Organization of the Document

Chapter One of this document focuses on the purpose and the organization of the study. The review of related literature is outlined in chapter two. Chapter Three of the study describes the methodology of the research study. After the study, the data collected is summarized in chapter four. Chapter Five focuses on the conclusions, implications and recommendations formed after the study and the analysis of data.

CHAPTER TWO

Review of Related Material

Tennessee Teacher Evaluations

In the state of Tennessee the landscape of teacher evaluations dramatically changed in 1997 with the introduction of the Framework for Evaluation and Professional Growth (FEPG). The new framework was introduced and approved by the Tennessee State Board of Education in 1997 (Wright, 2012). The FEPG was piloted across Tennessee from 1997 to 1999 and was fully implemented in July of 2000. The FEPG contained 44 criteria correlated to six domains:

1. Planning
2. Teaching Strategies
3. Assessment and Evaluation
4. Learning Environment
5. Professional Growth
6. Communication (p.1)

Direct guidance was provided by the state as to identify the data that were allowed to be used in evaluations and the procedures for the observations that mirrored Goldhammer's (1969) model. The eventual final evaluation was completed by the school principal and that decision could determine hiring and firing decisions, course assignments, pay scale adjustments and amount of evaluations the coming year. This model was accepted and used by administrators, but the State Board of Education wanted to increase the rigor of evaluations for all teachers. Teachers were required to be evaluated at least once a year, and up to three times a year based on years of experience teaching. The increased rigor model was in effect until 2010 when the *Tennessee First to the Top Act* (FTTT) altered the way that educators across the state were

evaluated. The new component that was to be involved in school, administrator, and teacher evaluations was student achievement on standardized exams. In January 2011, the resolution was passed that incorporated a measure of student achievement that counted for 15 percent of the teacher's evaluation. This is known as the student achievement portion or growth score (Wilson, 2012). For an individual teacher who taught an EOC tested course, student performance on that exam was incorporated into the total evaluation score. Student performance was transferred into an analysis system known as TVAAS (Goldstein, Behaniak, 2005).

In the late 1980s, two professors from the University of Tennessee used longitudinal data to measure the impact different teachers had on student outcomes (SCORE, 2012). In 2003 William Sanders, a statistician, stated that "beyond the use of a longitudinally merged database to provide input into a value-added analytical process to produce measures for-accountability purposes, the availability of this type of database presents several opportunities for the extraction of positive diagnostic information to be available to practitioners that heretofore could not be provided" (p. 3). This longitudinal growth he suggested was the basis of a teacher's TVAAS score based on their students' past testing history. In the TEAM Model, a teacher's TVAAS score could be selected as a measure for the 15 percent student achievement measure incorporated into a teacher's evaluation. According to SCORE, "TVAAS score for a teacher is determined by looking at the amount of growth above, below, or just at expectations that each of the teacher's students make in a given school year" (p. 4). The expectations that were set for an individual student were based on historical testing data on TCAP exams that the student had previously completed. Once a baseline was set, a student's expectations for achievement within a course were set and their performance on the summative Tennessee exam would determine if the teacher provide that student the opportunity to master the expected amount of content in relation

to their peers, students in other teachers' courses, and in schools across the state. A teacher's TVAAS score was used to help teachers support their students, improve their instruction methods to the best practices, and help teachers collaborate to improve the collective performance of teachers.

This method of assessing students and analyzing teachers on the basis of student performance was brought into question when the assessments that students were given were changed by Tennessee. The validity and reliability of the TCAP exams were brought into question due to the struggles of the transition to new assessments (TDOE, 2015). The transition to the TN Ready exams, and the also to the EOC exams impacted teachers' evaluation scores. Focused on TVAAS scores, some school districts created a policy where a teacher could receive a different salary, bonus or even pay level based on their TVAAS level from the previous assessment year. When the validity of the assessments came to the forefront of discussion, teachers began to question the validity and reliability of using those scores in the evaluation process as well. This only increased as the structure and rigor of the TCAP assessments changed.

History of Standardized Testing

Standardized testing developed into a normalized part of the educational world over a span of time in the United States. One would think that the history of standardized testing lasted as long as the educational system in the United States. According to Carole Gallagher (2003), Horace Mann suggested switching to written tests to assess the knowledge levels of students. Prior to written tests, students were assessed through oral exams. Mann's example aimed to create a method that identified the best practices for education. The best practices that were identified would be used and taught across the country in teacher development. These goals and aspirations were presented to the Boston Public Schools in 1845. "Mann persuaded the Boston

Public School Committee to allow him to administer written exams to the city's children in place of the traditional oral exams" (Gallagher, 2003, p.4). Prior to this time, "Public examinations were generally held once a year and were more in the nature of public displays or exhibitions to show off brilliant pupils or to glorify teachers" (Kandel, 1936, p. 24). This type of exam did not compare teachers to one another. This lack of comparison is what drove Mann to create the written exam.

After administering his exam, the results indicated that there were wide gaps of knowledge among the schoolchildren in Boston. The gaps that were identified allowed Mann to create a series of exams that would attempt to identify when a child was ready to be promoted to the next academic level. The exams were instrumental in improving the educational practices in Boston City Schools. This improvement led to other cities, school districts and educational centers to develop their own version of written exams. The most notable one that was developed was the New York Regents Exam that was adapted and changed, but was still being used as the name of the exams. Urban Schools throughout the United States including California, Kansas, Michigan, New York, New Jersey, Pennsylvania, and Massachusetts began using these new versions of measurement tools to assess their students (Chapman, 1988).

These exams were influenced by the progression that was also being made in mental evaluations. Numerous psychologists were creating exams to determine the mental abilities, disorders, and intelligence levels of an individual. According to Wash and Betz (1995), a French psychologist named Alfred Binet was working on developing a test that could be administered in order to identify, "slow children who would not profit significantly from schooling" (p. 2). Binet's (1916) exam created a scale of intelligence that would identify the mental age of a child. The identified mental age would then allow the administrators of the exam to directly compare

students, and also individual students to their past results. Binet had the goal of removing students with mental disabilities from the school in order to focus on the students who were mentally capable of improving based on the results of his assessment. Binet's assessment became the basis of the test that H.H. Goddard used in New Jersey. Goddard published his version of Binet's scale in 1908. Goddard's (1913) system was implemented as a method of incorporating intelligence testing data into the conversation during decision making about a student. This method was promoted to public schools in 1911. This was the primary event where student testing data was involved in school placement decision making for the student. In 1914 William Stern published a book that provided the current formula for intelligence testing known as an individual's IQ. The formula was to divide an individual's mental age with their chronological age and then multiply by 100. This simple formula had a lasting impact on the landscape of education across the globe. The scientific world now had a set of specific guidelines to measure any student against a norm, and could create guidelines based on student performance. This data was then integrated into those decision-making conversations that Goddard strived to create when determining the placement, or advancement of a student to the next academic level. Numerous psychologists worked on creating and developing a consistent and comparable method of determining the mental age of an individual. At this point, as frequently repeated, necessity created a time of opportunity for invention.

When the United States became involved in World War I, recruiting issues were prevalent. "The United States needed a way to evaluate the ... hundreds and thousands of recruits and would-be officers in rapid fire fashion" (Kaufman, 2009, p. 25). The first exam used was the Stanford-Binet exam which was created in 1916. This exam was a set of items used to determine the mental age of children across the United States. The creator of the exam, Lewis

Terman, was a professor at Stanford University. One of his students, Arthur Otis, was a lead developer of a multiple choice Binet style exam that could be administered to large groups of individuals. A variation of this exam, the Army Alpha, was used by the United States army to assess the IQ of thousands of soldiers and hundreds of potential officers. The goal of administration was to determine whether there were candidates for officer nomination within the large pool of recruits that had applied. Paper and pencil exams were envisioned to be the most efficient and effective way to assess large groups of individuals. This style of administration became the model for all subsequent standardized exams (Rothman, 1995). The Army Alpha was administered to an estimated two million recruits. The efficiency and data produced by the Army Alpha led school systems to developing models that mirrored this style of assessment.

Schools hoped to use assessments to determine not only intelligence, but also the mental capabilities of their students. Standardized testing became the model to determine the classification of students based off their results. The varying results presented the school data to place students in certain classifications and plans of study. This tracking, as it was commonly called, limited the available options for students based on their performance. The locally developed exams were a start, but the exams were localized and specific. There was a need for an exam for elementary level students that combined several content areas into one exam, and also had results that could determine intellectual ability. This necessity was addressed by the Stanford Achievement Tests in 1923 (Gallagher, 2003). These exams were used to place the child into one of two groups: the students who had learned, and the students who had not yet learned. The distinction in the data allowed for teachers to analyze if the teaching strategies had been effective, and if there were any notable gaps in the student's learning. This same type of analysis occurred throughout the year with modern standardized testing models as well. Around

this same time, other states were creating their own versions of standardized exams. The University of Iowa, for example, developed a set of assessments that could be utilized statewide on a voluntary basis. There were two main exams: the Iowa Tests of Basic Skills and the Iowa Tests of Educational Development. The Iowa Test of Basic Skills (ITBS) was developed for grades 3-8. According to Tomblin and Nippold (2014), all of the schools that were in Iowa were required to administer the exams. Therefore, the ITBS were not required to be completed by the schools. There were multiple subtests incorporated into the exam including, "Reading comprehension, Language, Mathematics, and Social Studies" (Tomblin & Nippold, 2014, p. 169). The Iowa Tests of Educational Development (ITED) were designed to provide a comprehensive development of a student during their time at a secondary school (Hendricks, 1967). The original exam included subtests covering 9 different topics including: Understanding of Basic Social Concepts, Background in the Natural Sciences, Ability to do Quantitative Thinking, Ability to Interpret Reading Materials in the Natural Sciences, Ability to Interpret Literary Materials, General Vocabulary and Use of Sources of Information (Hendricks, 1967). These 9 subtests were then summarized into one composite score. The tests aimed to identify how well students were able to apply the information acquired throughout their schooling. Any areas with a decreased performance indicated that the student was not able to think critically with the provided assessment materials. This could then be used by teachers to adapt and alter teaching strategies accordingly. Other states adopted these achievement tests and for over 50 years they were considered the most frequently used commercially available achievement tests on the market (Peterson, 1983). These achievement tests not only changed education on a primary and secondary level, but also dramatically impacted the college level as well.

College admissions staff examined the Army Alpha to determine whether there could be something similar administered that would allow them to determine if a student was prepared enough for their school or not. Therefore, a group of college admissions personnel worked together to create a list of standardized admission requirements for entry into their top-level schools. This group of individuals created a panel in 1923 called the College Entrance Examination Board (CEEB) (Hays, 2017). The CEEB developed a panel with the goal of translating a standardized set of admission standards into an examination that would be used as an admission criteria for the panel of colleges represented. The panel would also oversee the administration of their exam. The exam was developed to test not only student intelligence, but it also aimed to assess the achievement the student had made during prior schooling. The exam was commonly known as the CEEB test.

Desiring new types of information, a professor at Princeton, Carl Brigham, aimed to refine the CEEB test into a more updated version. This new version would help define what a student should have learned through the nature and content of the college preparatory instruction the student had received (Walsh and Betz, 1995). This test was called the Scholastic Aptitude Test (SAT). The SAT was an attempt to provide an equal educational opportunity to all students because each student across the country would be taking the same exam and would be graded on the same scale. The first SAT was taken in 1926 by 8,040 individuals applying for college, however for multiple years, the CEEB exam was the preferred exam for colleges (Crouse and Trusheim, 1988). Schools who were involved in this were known as the College Board. The College Board branched off from the CEEB and focused on the development of the SAT. The SAT was originally designed to be a screener for high achieving students to receive a full tuition scholarship to Harvard University (Zwick, 2004). As the SAT developed it was adopted as the

admissions test for all College Board school applications regardless of scholarship qualification. The success of this screener was proven Henry Chauncer to be able to test large quantities of students across the country at the same time and on the same day and still be secure and scored accurately (Zwick, 2004). After Chauncer proved this could be done, the SAT was recognized as being able to screen every child across the United States for college admissions. The SAT was comprised of sections including verbal skills, mathematical skills, and reading skills. The individual subtest scores would be combined to create a composite score for a student.

As a result of World War II, some Ivy League schools such as Princeton, Yale and Harvard changed their academic calendar to a year round calendar. This would mean that an incoming freshman student would be admitted in May of their senior year, but would start courses in June or early July. The decision created a logistical hurdle. The CEEB exam had a one day testing window with multiple exam administrations on one-day, and would receive their scores, then use those scores to determine eligibility for scholarships and admission into schools. However, with elite schools moving up the start of classes, this stress was influenced on the CEEB test to return results sooner to students. The exam consisted of essays that had to be hand graded, before scores would be released. To shorten the turn-around time, the CEEB was pushed to administer the multiple choice SAT exam and also various achievement exams. In 1947, the Educational Testing Service, a non-profit and non-stock corporation, took control of designing and administering the CEEB exams (Hanson, 1993). This emphasis on a quicker turn-around for scores and results also changed how the exams were administered to students.

The vision of ranking and sorting students into different tiers led assessment designers to create an institutionalized version of administering exams in order to determine a student's aptitude and achievement. The protocol expected of testing administrations included trial exams

and statistics was used to determine the analysis of exam results. An administration guide was created to ensure that rigid instructions were followed for an isolated group of individuals completing the exam. This guide would include administrative verbal directions, and the creation of an exam on the basis of a nationally endorsed and approved set of multiple-choice questions. These questions could cover a multitude of skills and knowledge a student was expected to have learned by the time that they applied for colleges. The exams would also need to be scored externally to remove teacher bias from the grading process of the previous CEEB essay exams. Goodenough (1949) stated it best, “The early tests were far from perfect; nevertheless the called attention to the necessity of using a standard situation, of providing a common basis of reference, if individuals are to be classified in a uniform and meaningful way” (p. 87). Many educators would argue that this would still act as a strong definition of the term standardized.

The SAT was the primary standardized exam that was administered under the new guidelines. The development of a new technology revolutionized testing forever. Reynold B. Johnson created a test scoring machine called the Markograph (Lemann, 1999). This new machine could identify whether a student selected the correct answer by determining the location of the pencil mark on a sheet of paper. Technology dramatically reduced the amount of scoring time needed, and also increased the amount of exams that could be graded in a short time. This was exemplified by high-stakes testing that used a Scantron form that was shipped to a grading facility, and then reported scores back to the student and school. A scantron form was a piece of paper that was printed with a specified number of 4 to 5 bubble sets. The student would then work off of a test packet, select an answer choice, and then bubble in the correlating bubble of the respective answer set. A scantron form could be used for any test as long as the answer sets matched the choices available in the test packet, and the most common choices were A, B, C or

D. Standardized testing would again undertake a transformation during another war, the Cold War.

The Cold War led politicians to focus on the national issues of security, safety and education. This concern was increased when the Soviet Union launched Sputnik in 1957. The satellite launch led politicians to focus closely on the education that students were receiving and finding ways to assess that learning. In order to assess learning, the role of standardized exams was expanded tremendously. In 1965, the Elementary and Secondary Education Act (ESEA) was passed and increased the use of standardized testing in schools. The increase in the role of standardized testing was influenced primarily by the decrease in the average SAT score across the nation. In order to address the declining scores, the College Board who produced the SAT and the Educational Testing Service (ETS) who produced the ACT, were tasked with determining the cause for decline and possible interventions that could be initiated to increase the SAT scores again. According to Wirtz (1977), the study concluded that weak educational standards, increased television watching and the slow deterioration of the family were what contributed to the decrease of SAT scores over the fourteen year span. The one area the government determined that could be impacted was the weak educational standards. Standardized testing also received more focus and attention as a result of the minimum-competency movement. The focus of schooling prior to World War I was vastly different than the focus after World War II. The industry and career workforce received almost half of the students who were graduating from high school. Schooling was looked at as a way to prepare a child to enter the work force by learning the skills necessary for a trade, or as an avenue to attend college. However this view shifted after World War II when more holistic subjects such as health, civics, and physical education were included into the schooling of students. Education

was now focused on the entire person and not just the core academic subjects. Students who were graduated were meant to be healthy parts of a democratic society. To instill this into the students the student was enrolled in courses focused on improving the whole person. The holistic approach also impacted the methodology behind grading, and made it far less likely that a student would fail an individual course (Huddleston and Rockwell, 2015). This perception of lowering the standards so that fewer students failed courses was coupled with the decrease in SAT scores to create a minimum-competency from their graduating students. For many states this took shape with the goal that a high school graduate should be able to complete basic mathematics problems, read and write. To attain these basic skills, numerous states passed legislation that used standardized exams to determine whether a student had reached the minimum-competency measurement (Rothman, 1995). The legislation had an impact; but the release of *A Nation at Risk* (National Commission on Excellence in Education, 1983) added gasoline to the fire of incorporating standardized testing to determine student performance by including the following:

> Our nation is at risk. Our once unchallenged preeminence in commerce, industry, science and technological innovation is being overtaken by competitors from throughout the world... the educational foundations of our society are presently being eroded by a rising tide of mediocrity that threatens our very future as a Nation and a people. (p. 5)

By 1989, in response to the report, 47 states had adopted policies that increased statewide standardized testing programs within the schools. Tennessee was one of the 47 states that adopted policies for their state-wide testing program.

Tennessee began administering the adopted version of a graduation exam in the early 1980s (Morgan, 2004). The name and nature of theses exams changed over the years. This was a result of changes to standard levels, format and also the context for the exams. Graduation exams were rising in popularity among states because the schools could prove that their students were ready to move on to college or the career workforce. Tennessee named the state mandated version of the exam the Competency Test. The Competency Test focused on skills in mathematics and language arts. Students would originally take this exam in ninth grade. However, students not receiving a passing score would not be allowed to graduate. Therefore, many students had to retake the Competency Test multiple times in order to graduate. The requirement to pass the exam prior to graduation dramatically impacted the graduation rate in Tennessee. In 2001 the graduation rate of Tennessee, because of the Competency Test requirement, was only 60 percent (Morgan, 2004, p. 2). This low number forced Tennessee legislators to look into developing a new style of graduation exams. These new exams would be known as the Gateway Exams. The Gateway Exams would no longer be comprehensive assessments, but would be subject specific assessments. The original three subjects were Biology, Algebra I and English II. Students were required to pass the subject-specific exam in order to earn the credit for the course and meet graduation requirements. Students not passing the exam they were required to retake the entire course again. Re-taking the exam was not an option for the Gateway Exams. The Gateway Exams implemented in the 2001-2002 school year in Algebra I, Biology and English II. The three exams were required for a student to graduate in Tennessee; but the state was also developing exams in English I, English 3, US History, Algebra II, Geometry, Chemistry, in response to the No Child Left Behind Act in 2002. These exams provided a single exam for each subject that was being tested. In January 2008, a new high

school policy was passes that eliminated Gateway Exams and replaced them with End-of-Course (EOC) exams.

The EOC exams were introduced in conjunction with new standards that were also being introduced in Tennessee. According to Justin Wilson (2009, p. 2) Tennessee adopted new standards because the state received a grade of "F" for "Truth in Advertising About Student Proficiency" from the National Chamber of Commerce. The report found that on the Tennessee eighth grade exam in 2005, 87% of students were scored at the *proficient* level. However, the National Assessment of Educational Progress (NAEP) found that only 26% of student were proficient in reading and 21% proficient in math. Tennessee had set the proficiency standards far too low for their exams, and therefore created new standards and exams to attempt to correct this issue. The initial step was to administer diagnostic exams yearly in eighth, tenth, and eleventh grade. The suite of assessments used were developed by ACT, Inc. The three exams were the EXPLORE, PLAN and the ACT. The exams were used to determine student college readiness at different points of the academic career. The college readiness standards were determined nationally so they would identify whether Tennessee students were progressing at a positive rate compared to students across the nation. New standards meant that a new exam needed to be developed in order to determine student growth. Thus, EOCs were developed and implemented.

Students in the class of 2013 were the first class subject to taking the full suite of EOC exams. Students prior to the class of 2013 were still held to the Gateway passing requirement. Students who needed to complete the EOCs were required to take the exams in ten different subjects as funds were available to develop the exams: Algebra I, Algebra II, English I, English II, English III, US History, Biology, Chemistry and Physics. The Physics exam was never developed or implemented. Instead of requiring a student to pass the EOC to graduate, as the

Gateway was, these exams would be incorporated into the individual course grade. This allowed the possibility of a student taking the exam, failing the exam, but still passing the course dependent on their course grade. These exams were composed of one multiple choice test that had no time limit. These exams were in place until the 2015-16 when TN Ready exams were introduced.

Tennessee once again changed the academic standards to ones similar to Common Core Standards. As with each time the standards change, a new assessment was required. These exams were a direct replacement of the EOC exams. Besides the different standards, there were major changes with the TN Ready exams compared to EOC. First of all, the TN Ready exams were now timed. In the past, EOCs were completely untimed and a student had as long as they felt they needed to complete the exam. The TN ready exams were split into two different administrations. For English and US History, the first administration was the composition of an essay. This was the first standardized written exam that Tennessee had introduced that was incorporated into student grades. The administrations were also during two different parts of the course. One administration, or Part I, was administered when the students were around 75-80% complete with the course. Part II was administered near the end of the course. This two-part administration was new, but continued to change and evolve. However, there was an issue with the new vendor for the exams, and the first administration of TN Ready exams was a debacle. The new TN Ready exams were also intended to be delivered online. The online platform failed, and paper copies could not be produced soon enough. This caused the cancellation of grade 3-8 testing across the state. The fallout of cancelling exams caused more changes to TN Ready exams. Due to the tarnished name, the exams were reverted back to the original name of EOCs. These EOCs took on a different style of administration as the original EOCs. Instead of one

administration, the exams were broken down into three or four different administrations. The goal was to not impact class instructional time as much as one long test historically had. However, the issue brought to the forefront was whether breaking the exam down into multiple subparts changed the authenticity of the exam.

Authentic Assessment

With the dramatic changes that happened in regard to standardized testing, it was debated whether or not the standardized exams were still authentic assessments. Parents and teachers often felt that the test was no longer an accurate representation of the student's learning throughout the course. According to Janet Powell (1993), an authentic assessment, "means making your assessment strategies match your instructional practices" (p. 3). There are multiple definitions for authentic assessment, but based off this one, many teachers have argued that standardized testing is not an authentic assessment for students.

Based on Powell's definition, a student in Tennessee should be given a multiple choice exam for all of the student's assessment. Assessment ideally reflected the style of teaching that was occurring in the classroom regularly and consistently. This idea changed how many instructional groups looked at assessment in the classroom. The National Commission on Testing and Public Policy (1990) recommended that assessment should be an ongoing process that helps emphasize what students know instead of a group of tests that help identify what the student has not learned yet. This was a tough challenge for teachers due to textbook development companies. Most of the instructional materials that teachers could receive had assessment materials that were accompanied with them. The assessment materials were typically focused on chapter and unit assessments that determine whether students understood the basic vocabulary and facts stated in the textbook readings. Teachers were posed with the decision between making their own

assessments and using the textbook provided assessments. EOC exams that are administered to students are focused on the standards that the students should master by the completion of the course. This brought about questions as to what an authentic assessment was and whether or not Tennessee exams are authentic.

Savery and Duffy (1995) described the authenticity of an assessment as the similarity between both the criterion situation the assessment is based on and also the cognitive demands of the assessment. Gulikers, Bastiaens, and Kirschner (2004) defined authentic assessment as "an assessment requiring students to sue the same competencies, or combinations of knowledge, skills, and attitudes, that they need to apply in the criterion situations in professional life" (p. 69). The criterion situation would determine the resemblance level that an assessment would have and therefore the authenticity of that assessment. One of the hurdles faced was the aspect that authenticity is subjective (Huang, 2002) and was also dependent on one's perceptions. This incorporates that what a student thought was authentic may not be perceived as such by the teacher. Perceived authenticity, in turn, was true between the teachers and the test developers as well. Although teachers and test developers spend a tremendous amount of time developing their own authentic assessments, the time and effort was all for nothing if the students did not perceive them as authentic as well. Huang (2002) suggested that pre-authentication can be interpreted either as that the design of one was impossible, or that it was very important to examine the experience of the users of the assessment. If the users of the assessment did not value the importance of the assessment, one could argue that the assessment was not authentic.

Validity

One value of authentic assessment that is relative to the authenticity was the construct validity of the assessment. Embretson (2007) suggested that the construct validity of an exam

was based on the meaning of test scores, and the implication of those scores. One way to improve the construct validity was to understand how item-type-specific features and testing affect knowledge, skills, and abilities. Construct validity can be summarized as: does the exam that was administered measure the goal of administration? Brown (2000), suggested that a construct is, “an attribute, proficiency, ability, or skill that happens in the human brain and is defined by established theories” (p. 9). The most difficult aspect of construct validity was the aspect that it must be studied from and individual perspective. Therefore, the amount of information that is available about the validity of the test, the stronger the construct validity of the exam would be. Teachers, parents and students in Tennessee argued that the amount of information available was inadequate to support a valid exam. Therefore, the construct validity of the exam was relative to public perceptions as well. Some of the largest threats to the validity of an exam are the reliability and consistency of the exam (Brown, 2000).

The construct validity of an exam could be impacted by problems in multiple categories. These problems, stated by Brown (1996), created five different categories: “environment of the test administration, administration procedures, examinees, scoring procedures, and test construction” (pp. 188-92). These problems are typically addressed by the psychometrics that are involved in the test development process. The recent history of Tennessee testing did not support the psychometric measurements used during test development. When exams were changed from an online testing platform to a paper on the day that testing was supposed to begin influences the environment of the test administration. Also, the cancellation of exams in grades three through eight led to parent belief that the exams were no longer high stakes exams. The cancellation of exams for grades three through eight also impacted the secondary level students. Many students, parents, and teachers felt that it was unfair to cancel testing in these grades. The environment for

testing was one where students felt as though they were being punished as the only grades that were still required to take the TN Ready exams. This perception could have dramatically impacted the construct validity of the exam. These perceptions also should have impacted the interpretation and use of the test scores. The use of the test scores was impacted when the Tennessee Department of Education decided to not include the test results into teacher evaluations. However, the scores and test results were included in both teacher and school effectiveness. The construct validity of the exams was impacted by the perceptions of the stakeholders, and therefore was a problem for the validity of the exam. Stakeholders still have negative perceptions of Tennessee testing and those perceptions might have impacted the construct validity of the EOC exams. This could have been an example of consequential validity, because that type of validity incorporated the social consequences of the exam as well. The College Board (2017) refers to the consequential validity as the social consequences surrounding using the exam for a particular purpose. The intention of tests was to provide a benefit to society. These social consequences could be felt from local schools. Test scores in Tennessee indicated a lower amount of achievement made by students during the chaotic year of testing. This decrease in scores was attributed to the new standards upon which the exams were based. However some teachers, schools, and parents questioned the authenticity of the exam based on the consequential validity as well. The authenticity of the assessment is also relative to the criterion-related validity of the exam.

The criterion-related validity focused on the relationship between a specific outcome on an exam and the test score necessary to receive that outcome. An example is the college and career readiness benchmark of the ACT. In order to be considered college and career ready, a student needed to receive a composite score of 21 on the ACT. The criterion for success that a

score of 21 suggested was student scheduling and grade point average for their freshman year of college. Students scoring below a 21 as a composite, often had to enroll in remediated courses their first year in college due to their status of not being college and career ready. In order to determine this validity needs to be correlated to the criterion in order to accurately represent the criterion behavior (College Board, 2017). Other studies were not only completed on later behavior, but they were also conducted on the predictive aspect of correlation as well. Predictive correlation included academic success in a specific course, admission into a particular program, and acceptance into specific schools. Tennessee used the assessment program, TVAAS, to predict future student test scores based on past scores. According to the Tennessee Department of Education (2015), a minimum of three past test scores could accurately predict future performance on Tennessee Consolidated Assessment Program (TCAP) exams. The criterion validity of that process was questioned as the testing program was impacted by cancellations, changes and modifications. The criticism was focused on the change in standards and administration methods. Changing the exams to an online platform, the amount of administration sessions, and the change of standards was perceived by the public to have impacted the criterion-referenced validity. Students in Tennessee were placed on one of four tiers of results based on performance on a TN Ready exam: the lowest performing students were placed in the Below Basic tier, the next highest performing students were placed in the Basic tier, the next highest fourth of students were placed on the proficient tier, and the highest performing students were placed on the Advanced tier (TDOE, 2015). When Tennessee switched to EOC tests in 2016-17, the tiers were respectively re-named: Basic, On-Track, Approaching and Mastered. The changes to the administration and the exam standards also influenced the reliability of the exam.

Reliability

Reliability is the measurement of the correlation that occurs between two separate events (Center for Public Education, 2006). Reliability in regard to standardized testing meant that if a student completed the same exam repeatedly they would receive approximately the same score. Tennessee used psychometric data to analyze the reliability of exams due to the fact that exams could not be administered to students a second time. However, some teachers and students argued that the difference between taking TN Ready exam online instead of a paper and pencil version affected the reliability of the exam. Some researchers agreed that the reliability between the two tests was impacted due to the platform of the exam. Dillon (1994) suggested that reading was around 20 to 30% slower on a computer screen as compared to on paper. Belmore (1985) suggested that if individuals were given sufficient practice on the task comprehension should be similar across the two media. Wastlund *et al.* (2005) noted that tasks on a computer were more tiring and increased stress more than paper-based task counterparts. The increase let to a greater investment of cognitive functions to complete the same task on a computer instead of a paper task. Noyes *et al.* (2004) determined that there was no significant difference between a task that was presented on a computer or on paper in terms of cognitive workload. However, the striking finding was that significantly more workload was reported by subjects on the computer-based task, as compared to a paper task. This meant, according to this study, that students needed to work harder on a test on an online platform compared to students who took the exam on paper. This scenario could have been applied in 2016-17 when TCAP schools had the option of testing high school students online or on paper. However, in 2017-18 all secondary EOC exams were delivered on the online platform. The reliability of the exam was not impacted by the platform change according to psychometricians at the TDOE (2017). Further research into the difference

between online and paper testing could determine the reliability and validity concerns of standardized testing. In regard to standardized testing another factor in the reliability and validity of the exam was the test administration itself.

The creation of a standardized exam is one that establishes that the reliability and validity of an exam are ensured. Each question of a standardized exam must pass a battery of evaluation to remove bias. Removing bias ensures the fairness and sensitivity of a test item. Standardization of an exam could be a lengthy process. Hawkins (1995) suggested a SAT exam could take up to 18 months to develop. This 18 month period includes preparation, pre-testing, and publishing of the questions. Typically, TCAP assessments were developed over a two-year period. EOC exams were administered with field test items that were not included in a student's exam grade. For example, an exam that was composed of fifty operational items and twenty field test items would be scored based on the fifty operational items. The field test items would be analyzed for bias and performance to determine the reliability and validity of an individual item. Tennessee administered complete field tests when new exams were being developed for EOCs in subjects such as Chemistry, Geometry, US History and others (TDOE, 2010). These field tests were not considered operational exams. Therefore, the students who participated in the field tests did not receive a performance score and baselines for future students were established. TN Ready exams were among the tests that incorporated field test items. For the essay portion of the English and US History exams students were required to complete multiple writing prompts. At least one of the prompts were field test items that were being analyzed for reliability and validity. This process continued with the multiple choice items on the EOCs that were developed to replace TN Ready exams.

Test Administration Practices

According to the U.S. Department of Education (2013), “no library of best practices exists that could help state educational agencies and local educational agencies with test administration” (p. 1). A test could be reliable and valid, but through test administration and ancillary factors that test could be impacted dramatically. One such influence is the time of day that a student is assessed. Pope (2016, p. 1) suggested that having math in the first two periods of the school day instead of the last two increases the math GPA of students. Also, scheduling math early in the morning also increased math standardized test scores. Goldstein (2006) studied when the test was given throughout the day, and the student’s testing time preference correlation impacted test scores. The study suggested that IQ can rise or fall up to 6 points dependent on when the test is administered for an individual student.

Sievertsen, Gino and Piovesan (2016) studied the implication of what time of day the assessment was given and the impact that had on student performance. The study was based on testing in a school being relative to student schedules and the availability of computer labs class could use. This dependence was similar to the conundrum that many schools faced with the scheduling of online TN Ready exams that relied on specific computer requirements. The study concluded that for each hour later in the day that a test was administered test performance decreased by 0.9% of a standard deviation. It was also concluded that if a 20-30 minute break was taken the test performance improved by 1.7% of a standard deviation. The study concluded that this was a result of cognitive fatigue felt by students as the school day wore on. Students who had just returned from a scheduled break negated the impact of the later hour of testing. Sieverten *et al*. (2016) incorporated information about circadian rhythms and the time of day preferences of students. Randler and Frech (2009) found that students performed better on school

tasks when they were completed at their preferred time of day at school. For elementary students the preferred time of day was determined to be in the morning. However, for students who are experiencing puberty, the time of day preference shifts to the afternoon instead. This study concluded that students who have a preference for morning tasks performed better in the school environment (Randler and Frech, 2009).

Cognitive fatigue can occur "when individuals must succeed on high-stakes tests or perform critical cognitively designed tasks that take long periods of time to complete" (Ackerman and Kanfer, 2009, p. 163). When it came to the SAT Ackerman and Kanfer suggest that concerns over test time focus on one or more of lines of reasoning including: (1) additional time testing on the SAT negatively affected student performance; (2) testing fatigue increases as the total amount of time testing also increases; and (3) testing fatigue became an influential factor on SAT performance. Aria, as early as 1912, stated that as time on task increased the solution time to mental multiplication tests also increased. Davis (1946) investigated whether there was a change in performance over a 70 minute pilot simulation task. A major finding from this study was that during the early phase of time-on-task, individuals might respond to increases in perceptions of cognitive fatigue by creating more effort and so increase performance, and as time-on-task increased exhaustion of effort resources might reduce effort and performance below initial baseline level (Davis, 1946). An issue with cognitive fatigue was that most studies had the subjects report their own levels of cognitive fatigue. The reporting of this characteristic seemed to be reflective on the amount of effort that the subject perceived was needed to complete the task. If the individual task required more effort, the subject tended to report that he or she was working harder (Ackerman and Kaufman, 2009). Overall, subjects in the study increased their composite score in longer testing sessions on the SAT by an average of 13 points. The conclusion

of the Ackerman and Kaufman (2009) study suggested that the hypothesis focusing on longer testing sessions negatively impacted test scores was not supported. Findings did not support the decision by TDOE to break the newer EOC standardized tests down into multiple administrations instead. This change was to avoid cognitive fatigue for students. These results also found that even though the subjects reported a higher amount of cognitive fatigue their scores were actually improved with the longer test session.

Walz *et al.* (2000) whether or not a student receiving an accommodation to take a statewide assessment over multiple days. Prior to EOCs in Tennessee, multiple day administrations were not common statewide. This type of administration was only available as an accommodation for students who received prior approval on an ACT or SAT exam. Students who receive extended time have been found to improve their performance when given time-and-a-half on each subtest of the SAT (Mandinach, E et al., 2005). Fuchs et al. (2000) found that a student's score was not impacted by receiving extended time on a standardized assessment. Although offering extended time may benefit some students, the extended time may also fatigue students if the session becomes too long with the added time (Elliot, Marquart, 2003). The accommodation of multiple day testing would allow the benefit of allowing extra time with the potential of also minimizing the impact of fatigue by having more than one day to complete exams. The study administered the exam over two administration sessions to both students who receive accommodations and general education students. In regard to the special education students, the reading rate displayed on the assessment over multiple days and in a single day were very similar. However, general education students demonstrated lower levels of performance when taking the assessment over multiple days than when taking exams in one day. A limiting factor of

Walz's study was the control group was the use of a reading fluency test and not a standardized assessment.

A study by Perucca (2013), required students to take a standardized exam and administered it to groups of students in two different methods. The control group was given the test in one four hour session while the experimental group had three eighty-minute sessions to complete it. There was no a statistical difference in the test scores between the two groups. The selected population was elementary-aged students. A study focusing on changing a three day administration exam to more or less days could not be found. Studies focused on breaking an exam down from one administration session to either one or three exam administrations. However, there are no studies on the impact of the amount of subtests an exam is broken down into.

Testing over multiple days is an accommodation that was found as on option for ACT exams. Students could receive the accommodation of testing with varying amounts of extended time over multiple days. Students could receive time and one-half for each session over multiple days, double time for each session over multiple days, and triple time for each session over multiple days. Students were placed in these timing codes based on individual disabilities in order to allow access to the exam for all students.

The state of Tennessee provided a set list of accommodations that could be offered to students based on individual disabilities stated in a 504 plan or an IEP. These accommodations have changed and adapted to the platform that testing occurred on for TCAP testing. The paper testing era had different accommodations as compared to the online testing platform. Over both platforms students were offered extended time on TCAP assessments when they began to be timed assessments. Extended time is defined as, "students who qualify for extended time can

receive up to double time on an individual subtest" (TDOE, 2017, p. 48). Students can also receive accommodations such as adult transcription, assistive technology, rest/breaks, scribes, and large print test booklets. The option of multiple administrations was not an accommodations that was offered to students. However, in the 2016-17 TCAP testing season, the exams were broken down into one to four subtests based on the specific subject for all students.

Standardized testing intended to show growth and progress of students over a set amount of time. These increased standards were also applied to students with disabilities. However, due to carrying disabilities the access to the exams varied based on individual situations. Testing accommodations were provided to students to facilitate participation in standardized assessments. Elliott, Kratochwill, and Schulte (1999) suggested that "testing accommodations are changes in the way a test is administered or responded to by a student. Testing accommodations are intended to offset distortions in test scores caused by a disability without invalidating or changing what the test measures" (p. 2). Available accommodations varied between states and even exams. In Tennessee, some subject exams had different available accommodations than others. For example, a student was given the opportunity to have the test read aloud to them in a mathematics course, but was not allowed to in an English course. This difference was based of ensuring the validity of the exam. If a student was assessed on reading competency having the test read aloud to them eliminated the reading component of an exam. For EOC exams, students were not allowed to use a calculator on a specific subpart of the exam in order to assess mathematical skills. No students, including those with disabilities, were allowed to use a calculator to protect the validity of the exam.

Common accommodations for assessments included a special test setting. These settings included environments such as small groups, special education classrooms and even the use of

study carrels to decrease the amount of distractions. The timing of the test could be altered to allow the student an equal opportunity to access the exam. Individuals with reading or processing deficiencies could have extended time to ensure they were able to complete the entire test within a similar timeframe as a student without accommodations. For the ACT exam, students were able to receive time-and-one-half, double-time, and even triple-time on each subtest. The individual student applied for the accommodations and the ACT determined if the student qualified to receive the accommodation. How the exam was presented to the student could also be modified. These modification could be having the test read aloud to the student, having a braille copy of the exam available or a large print test book for students with vision impairments. The transition to online testing has allowed for a more seamless delivery of these modifications. The addition of an automatic zoom and delivery of text to speech through headphones was developed with the new EOC exams administered in Tennessee (TDOE, 2017). Students could also have the ways that they respond their answers to the exam modified. Students could receive accommodations such as adult transcription where they verbally responded to a question and the adult bubbled the corresponding answer for the student if physically bubbling was not possible. Elliott, Kratochwill and Schulte (1999) identified eight different domains of accommodations that included motivation, assistance prior to the administration of the test, test scheduling, setting, assessment directions, assistance during the assessment, use of equipment or adaptive technology, and changes in the test format.

High Stakes Testing And Teacher Evaluations

High stakes testing was not only used to assess student learning, but scores from high stakes exams were also incorporated into teacher evaluation scores. Efficiently assessing teacher

performance was a continual issue of debate and discussion. This efficient assessment had varied in approach over the history of teaching in the United States.

In the beginning of the educational system, the role of a teacher was considered the role of a public servant. The individual that the teacher reported to as a supervisor had total and unlimited power to determine the evaluation criteria for their teachers. The definition of an effective teacher was created and built by that supervisor and was used to hire and fire teachers (Burke & Krey, 2005). Teacher evaluations followed this pattern for a majority of the 1700s. For a teacher to be hired or fired, the supervisory committee held that decision in their hands. As the educational field continued to expand, the evaluation process also evolved with the profession. As industry and business started to create larger urban areas, that growth also created larger and more complicated school systems. The increased complication of the school system not only impacted the assessment of the students, but the complexity also influenced the manner in which teachers were also evaluated. The 1800s also led to teachers adopting specific disciplines that they focused and taught. Specific disciplines were then assigned to administrators as well, and was the creation point of the *principal* teacher that had developed into the building principal (Marzano, R, Frontier, T, Livingston, D, 2011). Even though the trend of specialization was found primarily in urban areas, the trend continued to spread to smaller cities and other rural areas. This specialization continued to be found in schools as teachers were designated as science, math, a specific grade level, and a designated course intensity. This complexity created a trend of increasing the complexity of the profession of teaching, and therefore the evaluations of teachers became more complex as well. Blumberg (1985) noted that the instruction that a teacher provided began to be the focus of supervisory evaluations. In 1845 the Superintendent of the Common Schools of the State of New York stated the following:

> Too much reliance ought not to be placed upon visitation to the schools, to give method to the teacher and efficacy to his instructions. Instruction is the primary object of visitation, and … more instruction can be given to teachers of a town when assembled together in one day. (p.131)

The focus on instruction continued to be a part of evaluations of teachers throughout the 21st century. However, another movement was developing for teacher evaluations in the late 19th century.

John Dewey was known as one of the forefathers of the American Education System. He saw the backbone of the educational system as democracy. Schools were the grounds where students were able to develop citizenship and also to develop the principles of democracy as well (Dewey, 1938). This mindset created progressive ideas such as student-centered education which is where the individual interests and needs of the student were the focus of that student's education. There was also a large push to integrate differentiated content areas into the instruction for individual students. Dewey's philosophical goal was to allow students the opportunity to move from learners in a passive educational environment to citizens as active members in society. This view was contradictory to Frederick Taylor who had a much more scientific approach to management.

Taylor's (1911) philosophy was not focused on education, but on the management of the efficiency of a factory. According to Taylor, over the numerous ways to complete a task there were some methods that would be more efficient, and therefore there was also a method that would be considered the best. Taylor applied his methods to rudimentary tasks such as shoveling coal, dividing labor and even training programs for workers. This philosophy of the best method to complete a task moved into the realm of education. Teacher education programs incorporate

courses on the current best practices for education, classroom management and even data analysis. Cubberly (1929) stated that:

> Our schools are, in a sense, factories in which the raw products (children) are to be shaped and fashioned into products to meet the various demands of life. The specifications for manufacturing come from the demands of twentieth century civilization and is the business of the school to build its pupils according to the specifications laid down. (p. 338)

The specifications were based in the based practices that were founded out of the management philosophy propagated by Taylor. The result of this style of thinking was used by William Wetzel in 1929 to define the effectiveness of a school or teacher based off student learning. His proposal was based on the combination of three components: the use of aptitude tests to determine a child's ability level, the establishment of clear objectives for each course, and the use of reliable measures of student learning. These three components were combined to determine whether a teacher or school was effective in the given time frame between assessments. Mirroring the shift in reliance on standardized testing after World War II, the assessment of schools and teachers were also shifted to focus on individual teacher's effectiveness rather than the school as a whole. Contrary to the development of a teacher as an individual, the supervisor who evaluated teachers had their role defined in specific terms. These terms covered the daily running of the school and all of the aspects that led to that. These aspects included curriculum, classroom environment, resources, personnel, attendance, public relations and staff training as well. The principal was looked at more as a manager of a factory than an instructional leader. The evaluations that teachers received from administration also mirrored a factory style of leadership where a checklist or categories were used to assess a teacher's

performance during classroom visits. This style led to the adoption of clinical supervision models that were extremely popular in the late 1960s and early 1970s. In 1980, Bruce and Hoehn found that about 90 % of school supervisors used some type of clinical supervision model. The most popular model that was developed in response to this movement was developed by Robert Goldhammer (1969). His publication, *Clinical Supervision: Special Methods for the Supervision of Teachers* was based on numerous classroom visits and supervisor conferences. The process was focused on five steps and included teachers in the reflection after a supervision. The first phase was the pre-observation conference of the teacher and supervisor that discussed the overview and objectives of the upcoming observation. The second phase was the actual classroom observation that focused on what was discussed during the pre-observation conference. Then, the supervisor organized the data they collected during the analysis phase. After the analysis, a supervision conference was performed where the teacher and supervisor discussed the data collected and the teacher reflected back upon their lesson. Finally, the supervisor's analysis was discussed with the teacher to ensure the fairness of the analysis. This model continued to be adapted and changed over the coming years until a new piece of data was being collected, and that was student high stakes test scores.

Cognitive Fatigue

A potential source of bias that has been identified for standardized testing is cognitive fatigue. Cognitive fatigue for students can take on multiple variations. Common variations for students included length of the school day, time during the school day the test is administered and the length of the test administration. Researchers suggested cognitive fatigue various ways. Preliminary research focused on the prolonged work on the performance of tasks that involved

memory functions (Ebbinghaus, 1896-1897). Studies centered on time-on-task and the impact of cognitive fatigue on performance.

Martyn (1913) investigated individual responses to conditions that cause fatigue. The study had three participants who completed mental multiplications over a one-hour work period. The participants displayed three different reactions to the task. One participant showed improved performance over the one-hour work period. A second participant showed no significant impact on their performance on tasks throughout the hour. The third participant performed at a lower level, had distracting thoughts, and length of task also had a detrimental effect on the individual's physical well-being.

This study was similar to the study completed by Aria (1912) on the solution time for a mental multiplication test. The study concluded that the amount of time that was used to complete the math tasks nearly doubled for individuals who were scheduled for twelve hours of multiplication. Individuals in the study took, on average, 24% longer for each item during a two-hour section of administration. These studies had other investigations that disagreed with their findings. Carmichael & Dearborn's 1947 study concluded that an individual's reading performance was not impacted even with a six hour administration period. This finding was supported by Davis (1946), who found that 75% of the individuals who were a part of a study that tasked individuals with a strenuous seventy-minute pilot simulation task. The study found that individuals performed in ways similar to the study completed by Martyn (1912). The vast majority of participants showed that participants seemed to not be impacted by the length of the simulation. Subject attention level and experience of fatigue stayed constant throughout the simulation. A second response that was found was the increased performance of individuals regardless of the reported experience of fatigue felt by the subjects. Also similar to Martyn's

study (1912), the remaining participants tended to scale back their effort in a manner to prevent cognitive fatigue. This withdrawal from the task by means of reduction of effort was a symptom of cognitive fatigue. Davis (1946) proposed another inherent possibility that he believed he witnessed in some of his participants. The proposal suggested that individuals might have initially increased their performance due to the environment, but as cognitive fatigue set in the effort averaged to the initial level of performance. This average placed the individual participant in the steady level of performance reporting category, and that category accounted for 75% of participants.

Ackerman and Kanfer (2009), investigated whether the length of an exam had an impact on test taker performance. The study administered the SAT exam to groups of students in three different administration lengths. The control group completed the SAT in the standard amount of time, four and one half hours. One experimental group completed the exam with an additional hour than standard, making the administration around five and one half hours. An alternate experimental group completed the exam in an hour less that standard, making their administration time around three and one half hours. The study also included self-reporting questionnaires that participants completed during the same administration of the SAT battery of tests. Ackerman and Kanfer (2009) found that students who were placed in the longer test group reported a higher sense of cognitive fatigue. However, the students in that group also performed better than the standard time administration group on the SAT exam. Students who reported high levels of cognitive fatigue did not show an immediate reduction in fatigue at the end of the testing session. Ackerman and Kanfer's (2009) conclusion that to "explicitly train new affect-action bonds- for individuals to respond to feelings of cognitive fatigue with higher, rather than lower levels of cognitive effort- may help to preclude premature withdrawal of task effort and

associated lower levels of performance" (p. 177). This conclusion suggested that an individual's subjective fatigue impacted performance on tasks over an amount of time.

Bills (1943) suggested that there were three main aspects to fatigue. One aspect was the physiological fatigue that is shown through a reduction in an individual's physical capacity. This fatigue was typically onset by strenuous activity of exercise that left the body worn down and fatigued. A second aspect of fatigue was objective fatigue in work performance. This type of fatigue was displayed by a decrease in the individual's performance. This decrease was regarded as getting less accomplished in the same amount of time, or producing an inferior product in the same amount of time. The thirst aspect of fatigue was subjective fatigue. This type of fatigue was self-reported by the individual in response to a task or environment. Åhsberg (1998) proposed that subjective fatigue was a response to concepts including stress, pain, alienation, boredom, burnout and anxiety. In regard to fatigue that was related to work-related perceptions, five dimensions were proposed. Those dimensions were "lack of energy, physical exertion, physical discomfort, lack of motivation and sleepiness" (p. 21). Perucca (2013) suggested that a student's sustained attention and academic performance were impacted by numerous stress factors rooted in being a low socioeconomic status student. The study suggested that a student's low socioeconomic status (SES) impacted the amount of attention that the student could focus on an individual task with and therefore would impact their performance. Burnett and Fanshawe (1997) concluded that the environmental stress of a low SES increased an individual's anxiety. That increased anxiety was associated with school exams, parental pressure, comparisons to other students, poor performance and the amount of time of exams. This anxiety was a large factor in standardized test performance as well.

Test Anxiety

Text anxiety is a factor that significantly decreases a student's performance on an academic achievement assessment (Zeidner, 1998). Psychological, behavioral, and physiological reactions commonly resulted from a poor performance, or failure, on an evaluation (Zeidner, 1998). Performance on standardized exams was a common cause factor in the presence or absence of anxiety based reactions. After the implementation of NCLB the amount of anxiety a student felt increased. The reliance on standardized exams to determine final course grades in subjects created a high stakes exam situation. Students who experienced a lower level of test anxiety often outperformed classmates who experienced a higher amount of anxiety. Newspapers and other investigations have looked into the impacts of high stakes testing on school culture, parents and teachers. Abrams, Pedulla and Madaus (2003) suggested that the increase of state testing programs have resulted in increased student anxiety, increased stress, lower motivation and a teacher's increasing focus on test preparation. Increased stress and anxiety could lead to a poor performance on an assessment in a course. The poor performance could lead to a decreased performance on a standardized exam as well. Text anxiety was often self-reported and contained some bias based on student perception of performance. Segool (2009) investigated the experience and perception of high stakes standardized assessments within a young student population. The study found that students perceived high stakes tests with more stress and anxiety than normal classroom assessments. Increased anxiety on standardized exams was magnified as the amount of standardized tests that students completed dramatically increased after NCLB was passed. Family support and parental involvement were factors that could help mitigate the amount of stress a student experienced on a daily basis. Students in a low SES setting were not in an environment where those mitigating factors were present (Burnett &

Fanshawe, 1997). Unlike the students in Ackerman and Kanfer's 2009 study, younger students may not have received the same training and education of coping skills and behaviors to mitigate the impact of test anxiety.

Test anxiety could also lead to higher instances of cognitive interference. Cognitive interference impedes the direction of an individual's attention during a test-taking situation (Sarason & Stoops, 1978). In a high stakes testing environment, thoughts of poor performance and criticism of one's own performance were characterized as cognitive interference. This led to more cognitive faculties being used to focus on the interference rather than the exam itself. Distracting thoughts also negatively impacted student performance on the assessment. Previous negative performance on an assessment could also lead to higher amounts of perceived anxiety on subsequent assessment. Modest amounts of perceived anxiety were sometimes associated with an increase in student performance for some students (Alpert & Haber, 1960). Modest amounts of anxiety were often found among students. However, the negative impact of anxiety was most commonly found when there was also an excessive amount of anxiety perceived. Excessive amounts of anxiety were commonly found related to high-stakes standardized exams. The increased amount of standardized exams could be extrapolated to conclude that the amount of anxiety felt among students in regard to standardized exams increased as the amount and reliance on standardized exams directly increased.

Saranson and Palola (1960) administered simple memory tasks to individuals under varying amounts of pressure. The students who received a high amount of pressure experienced high levels of test anxiety. Students in that environment also performed at a lower level on the memory tasks than individuals in a lower pressure environment. This study was replicated numerous times with varying exams, assessments and tasks and the performance of the students

supported the original findings. The collective research suggested that test anxiety, cognitive interference and reduced academic outcomes were correlated (Cassady & Johnson, 2002).

In Tennessee the reliance on standardized tests to determine graduation, course averages, and teacher evaluations increased the anxiety felt by students and teachers. The Gateway Exams were required in order to graduate. Students with poor past test performance often felt an increased perception of anxiety on the exam. The perception of higher anxiety often led to lower test performance regardless of content mastery (Cassady & Johnson, 2002). No known studies had investigated the correlation between multiple administrations and perceived test anxiety. Teacher anxiety increased with the standardization of state assessments in Tennessee. Tests were most recently included in course averages, teacher effectiveness scores, and pay scale advances. The stakes of the TCAP assessments increased dramatically which might have impacted student and teacher performance due to cognitive interference and anxiety. Inclusion of student performance on teacher evaluations increased the amount of anxiety felt by teachers who taught a TCAP tested course.

Summary

Throughout history the importance of standardized testing in schools has dramatically increased. This was the result of not only teacher interest, but political pressure and recruiting tactics. Throughout the entirety of the history of standardized testing, the attempt was to determine what level of performance an individual placed in based on their performance on a standardized test. The implications ranged from the awarding of a leadership position in the army to even a lucrative scholarship offer from a college. High-stakes exams in Tennessee are intended to determine the growth that a student has made after an academic year of learning within a

course. Student performance on the exam was used to determine student growth, teacher effectiveness scores, and school-wide growth for district and state analysis.

The conversation among teachers and schools centered on whether the newly developed high stakes exams that are being administered through TCAP are authentic assessments. Stakeholders were concerned that the reliability and validity of the TCAP were impacted by the testing administration issues in Tennessee. Perception of the exams influenced the validity and reliability of the exams among stakeholders. Each new development of the TCAP included an altered assessment schedule. Gateway exams were untimed assessments that were completed in a single session. A student needed to receive a passing score on the Gateway as a graduation requirement. The assessment that replaced the Gateway was the EOC. EOCs were also untimed single session exams that were incorporated into the student course grade. These exams no longer required a passing score for completion of the course and graduation. The student instead only needed a passing average, after the incorporation of the EOC grade, in the course average. Replacing EOCs were TN Ready exams. TN Ready exams were instituted due to the change in Tennessee standards. TN Ready exams were composed of a Part I and a Part II that were both timed. The exams were also composed of two subtests that were administered during different parts of the course. The student score was incorporated into their course average. Due to testing issues, the TN Ready exams were changed back to the name of EOCs. The new version of EOCs were broken down into either one, three, or four subtests based on the tested subject. Each of the subtests were timed subtests with a varying amount of time for each section..

Studies have investigated the impact that multiple day administrations have for students who require accommodations. There have also been studies that have examined the impact of breaking an exam down into one or three administrations. However, there has not been a study

that has examined the impact of breaking an exam down into two, three, or four administration sessions for both general students and also students who require accommodations to access the exam.

CHAPTER 3

Research Methodology

The quantitative study focused on the administration of a standardized exam to groups of students in differing number of subtests. The student participants were enrolled in two Advanced Placement Chemistry classes and two ACT preparatory classes. The four groups had a range of predicted scores on the ACT exam, and also had differing amount of experience on standardized exams. The population of the groups included students from Northeast Tennessee who attended an urban school district. The school selected was a public high school with a historical trend of high performing students on state and national exams.

Population and Sample

The school selected for the study was selected as a matter of convenience for the researcher. The local school district approved the research study prior to the administration of the subtests. The school was a large urban high school in Northeast Tennessee. According to the 2015-16 Tennessee Report Card, the school had a reported enrollment of 2227 students. Of these students, 24% qualified as economically disadvantaged and 13% fall within the students with disabilities subcategory. Four classes from the school were selected for the study. Students and parents completed consent forms and had access to the results of the study upon request (Appendices A & B).

For the purposes of the study, the classes were broken into two categories. The control group consisted of Ecology, English III, Anatomy and Physiology, English II, Physics, ad Bridge Math classes. The control group would take the test over three testing administrations. The remaining Anatomy and Physiology, English III, English II, Chemistry II, and Bridge Math

classes were designated as the experimental group. The test was administered to the experimental group in four parts.

Description of Instruments

A retired ACT exam was administered. The exam was used to predict student performance on a future ACT exam. There are four subsections of the ACT exam. The subsections are English, mathematics, reading and science. All subtests of the exam were administered under standard procedures and environment meeting ACT requirements. A standardized environment and access to materials ensured the validity and reliability of the exam was minimally impacted. Students in the control group completed the English and mathematics sections in one administration, and they completed the reading and science sections in the third administration. The experimental group completed each subsection in four separate administrations.

All student names and identifying details were removed from the data for confidentiality. Also, all results were available to students and parents for review after the completion of the study. The school used the data to assess student predicted scores for future ACT exams. Teachers will have access to all materials after the data have been collected and analyzed in order to address any student needs identified through the assessment.

Analysis of Data

This study was conducted to research the impact that the administration window has on the authenticity of the assessment.

The study focused on two research questions:

1. Does breaking an exam into smaller subtests impact overall student performance on exams?

2. Does the length of the testing window impact student performance on exams?

Research question one focused on the impact of breaking exams down into multiple subparts and the impact on student performance. To assess whether there was a statistical difference, the average ACT composite score was compared between subgroups. The analysis was completed using a t-test. A t-test focuses on the difference between the averages of two different groups (Siegle, 2017). The groups were separated by the number of administrations of the retired ACT exam. The null hypothesis stated that changing the number of administrations would not impact a student's ACT composite score. The null hypothesis would be accepted if there was no statistical difference between groups.

Research question two focused on the impact of the length of the testing window. The average scores of the groups were compared based on the number of calendar days used to administer the exam. A t-test was used to compare the average composite ACT score between groups (Siegle, 2017). Groups were similar in composition and demographics. Composite scores within groups that had a statistically significant difference supported the hypothesis. No statistical differences supported the null hypothesis of research question 2.

The study was initiated in the Fall of 2017. During the Fall Semester the research proposal was approved by the local school district. Upon receiving local approval, the research proposal was submitted to the Carson-Newman University Institutional Review Board for approval. This approval allowed the researcher to identify, collect consent forms, and administer the ACT exam to the approved students. The researcher then initiated collection of data relative to the study. The retired ACT exam was administered during the Spring 2018 semester. After administration, the data were collected, analyzed, discussed, and reported in following chapters.

CHAPTER FOUR

Analysis of Data

Introduction

The purpose of this study was to determine whether student performance is impacted by dividing an exam into multiple subtests and multiple administrations. This study was designed to determine if there is a saturation point of testing where student performance is negatively impacted. The study focused on 276 students who were enrolled in courses in which teachers volunteered their classes as participants. Students who returned consent forms were then separated by class into two groups: experimental group and control group. The 132 students were placed in the control group that tested over three administration sessions. The additional 136 students were placed into the experimental group that tested over four administration sessions. The study was designed to determine if there is an impact on student performance by changing the amount of administrations a multiple part exam was given over. All students completed a full-length retired ACT exam over multiple administrations. The exam was administered to students over the school learning management platform Canvas. The study focused on two primary research questions that guided the statistical analysis.

1. Does dividing an exam into smaller subtests impact overall student performance on exams?
2. Does the length of the testing window impact student performance on exams?

The goal of the study was to collect and provide information for administrators to determine the best practices for multiple administration scheduling. The research questions were analyzed using independent t-tests.

Students that participated in the study were enrolled in the Sophomore through Senior year at the time of the study. The control group was composed of 45 Seniors, 61 Juniors, and 26 Sophomores. Gender representation of the groups was 51 males and 81 females. The experimental group was composed of 10 Seniors, 51 Juniors and 75 Sophomores. The demographics are summarized in Table. 4.1 and 4.2.

Table 4.1

Grade Distribution of Groups

Group	Sophomores	Juniors	Seniors	Total
3 Day Administration	26	61	45	132
4 Day Administration	75	51	10	136

Table 4.2

Gender Distribution of Groups

Group	Males	Females	Total
3 Day Administration	51	81	132
4 Day Administration	67	69	136

Results and Findings

The study focused on two research questions that were analyzed with an independent samples t-test. The first research question posed was:

1. Does dividing an exam into smaller subtests impact overall student performance on exams?

The question was analyzed by comparing the mean ACT composite score achieved by both groups of the study. The null hypothesis presented was that dividing a 4 part ACT into 3 administration sessions would have no impact on student performance.

An independent-samples t-test was conducted to compare ACT and multiple administrations. There was a significant difference in the mean ACT composite for students who completed the exam in three administrations (M=18.13, SD=4.65) and four administrations (M=20.08, SD= 4.46); t (266)= -3.51, p=0.0005. These results show that students who completed the ACT exam over 4 administrations scored significantly on average two composite points higher than students who took the exam over three administrations.

Table 4.3

Student Performance over Multiple Administrations

Group	Mean ACT Score	SD	n	*df*	*t*
3 Administrations	18.13	4.65	132	266	-3.51
4 Administrations	20.08	4.46	136		

p=0.00035

The second research question posed was:

2. Does the length of the testing window impact student performance on exams?

The question was analyzed by comparing mean ACT composite scores between students who completed the ACT over three days and students who completed it over four days. The null hypothesis that stated that there would be no significant difference between the scores of students who completed the exam over a differing number of calendar days.

An independent-samples t-test was conducted to compare ACT and multiple calendar days. There was a significant difference in the mean ACT composite for students who completed the exam in three calendar days (M=18.13, SD=4.65) and four calendar days (M=20.08, SD= 4.46); t (266)= -3.51, p=0.0005. These results indicate that students who completed the ACT exam over four calendar days scored significantly on average two composite points higher than students who took the exam over three calendar days.

Table 4.4

Student Performance over Multiple Calendar Days

Group	Mean ACT Score	SD	n	*df*	*t*
3 Calendar Days	18.13	4.65	132	266	-3.51
4 Calendar Days	20.08	4.46	136		

p=0.00035

Summary

The study analyzed student performance with an independent-samples t-test. Student performance was compared over both differing number of calendar days of administration and also administration sessions. The t-test was significant to the 0.00035 level, showing a very strong significance between the number of calendar days, and also number of administration sessions, on the student performance on an ACT. The students who completed the ACT over four administration sessions significantly outperformed students who completed the assessment over three administrations. Students who completed the assessment over four calendar days significantly outperformed students who completed the same assessment over three calendar days. The students in the experimental group outperformed the control group on average by 1.96 composite points.

CHAPTER FIVE

Conclusions, Implications and Recommendations

The quantitative study investigated the impact of multiple administration assessments on student performance. The lack of information on this practice led to the quantitative study using a retired ACT exam with high school students. The purpose of the study was to determine if breaking the ACT down into multiple administrations impacted students' performance on the overall exam. A single study (Waltz, et al, 2000) focused on the administration of tests to students with disabilities, but multiple sessions for all tested students had not been performed. A retired ACT exam was administered to two randomly selected student volunteers to investigate the two research questions. Volunteer classes were divided into a control group and experimental group based off teacher scheduling preference. Due to interruption of lesson plans, teachers were provided the option of either scheduling option. Not all requests were approved due to necessary group composition requirements. The control group completed the ACT in three administration sessions. The experimental group completed the ACT in four administration sessions. Both groups completed the English test in the first administration session, and the Mathematics test in the second administration session. Then, the control group completed both the Reading and Science tests in the third administration session. However, the experimental group completed the Reading and Science tests in two different administration sessions. The control group completed the administrations over a three-day testing window, and the experimental group completed the tests over a four day testing window. Finally, the composite ACT scores were compared using an independent-samples *t*-test between the groups. The four-day/administration group outperformed the three day by, on average, 2 composite points higher than the three day/administration group.

Discussion of Conclusions

The study focused on two research questions:

1. Does dividing an exam into smaller subtests impact overall student performance on exams?
2. Does the length of the testing window impact student performance on exams?

Research Questions

The research questions focused on whether breaking an exam into multiple subtests and administrations over multiple calendar days impacted students' performance on the overall exam. The composite ACT scores from the groups were compared using an independent-samples *t*-test. The analysis revealed that there was a statistically significant difference in the composite scores between the two groups. The statistical difference was based on the *t*-test outcome of P=0.00035. The three-administration sessions group scored, on average, two composite points higher than the four administration group. The three-administration group completed the exam over three calendar days, and the four administration group competed the exam over four calendar days.

Conclusions

The independent-samples *t*-test findings refuted the null hypothesis that there was no impact from multiple administrations on student performance. The researcher concluded that students performed significantly higher on an exam that was separated over multiple administrations. The findings suggest that school districts should separate the multiple subtests of Tennessee End of Course exams over multiple administrations. This decision would significantly improve student performance in comparison to students who completed the exam in fewer administration sessions. The study supported the authenticity of the assessments even when the administration sessions were changed. The student performance was impacted, but

students performed at a higher level when the standardized exam was administered over more sessions and calendar days. Implementing multiple administration sessions could enhance the best practices for test scheduling and administration. This study suggested that students would perform better if those subtests were administered over multiple days and broken into multiple subtests.

Akerman and Kanfer (2009) suggested that the student performance would be impacted by cognitive fatigue. Decreasing the length of time of an exam would decrease the cognitive fatigue that students would perceive. The current study investigated whether testing fatigue would set in over multiple administrations. Student performance suggested that multiple administrations and calendar days positively impacted student performance. Multiple administrations decrease the amount of time that students were in a high stakes testing environment. This decrease of time also decreased the amount of cognitive fatigue that students would perceive. The increase in student performance may have been impacted by lower test anxiety. Shorter testing sessions and less time in a standardized testing environment decreased test anxiety. Breaks between administrations helped decrease the amount of anxiety that students could perceive.

The state of Tennessee provided districts with the flexibility to administer the multiple subtests of TCAP exams in the manner the district saw fit. However, this study suggested that there was a best practice for scheduling the multiple subtests. The performance of students increased when they took the same standardized exam over four days instead of three days. School districts that decided to put multiple subtests of exams together on the same day had the potential to negatively impact their students.

Recommendations for Further Studies

This study was limited by multiple factors, and could be followed up by multiple other research studies. The administration of the subtests was completed using an online testing platform. This platform could have impacted the baseline performance of all students, as it was the first time the students completed an ACT exam on an online platform. All students completed the exam on the same platform. This meant that all students were impacted similarly by the online platform. More research is needed to determine the change to online testing platforms. Testing platforms changed with each learner management system that was used. Additional research is needed to determine whether different platforms impacted student performance on varying exams. Student motivation and standardized exams deserve further study. Due to the limitation of the inability to use the TCAP exams, student motivation may have been impacted in this study. TCAP exam scores have been incorporated into student grades, and therefore could have impacted student motivation. More research is needed to determine whether longer testing windows impacted student motivation and final course grades. Further studies that investigate the impact of multiple administration exams and course grades would determine whether student performance is further impacted by multiple administrations.

References

Abrams, L., Pedulla, J. & Madus, G. (2003). Views from the classroom: Teachers' opinions of statewide testing programs. *Theory in Practice*, *42*(1), 18-29.

Ackerman, P. D., Kanfer, R. (2009). Test length and cognitive fatigue: an empirical examination of effects on performance and test taker reactions. *Journal of Experimental Psychology*: *Applied*, *15*(2), 163-181.

Åhsberg, A. (1998). *Perceived fatigue related to work*. Stockholm, Sweden: Arbetslivsinstitutet.

Alpert, R., Haber, R. N. (1960). Anxiety in academic achievement situations. *Journal of Abnormal and Social Psychology*, *61*(2), 207-215.

Aria, T. (1912). *Mental fatigue*. New York, NY: Columbia University.

Belmore, S. M. (1985), Reading computer-presented text. *Bulletin of the Psychonomic Society*, *23*(1), 12-14.

Bills, A. (1943). *The psychology of efficiency*. New York, NY: Harpes.

Binet, A., Simon, T., Kite, E. (1916). *The intelligence of the feeble-minded*. Baltimore, MD: Wiliams & Wilkins Company.

Blumberg, A. (1985). *The school superintendent: Living with conflict*. New York, NY: College Press.

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

Brown, J. D. (2000). What is construct validity*? Shiken: JALT Testing & Evaluation SIG Newsletter*, *4*(2), 8-12. Retrieved August 14, 2017, from http://hosted.jalt.org/test/bro_8.htm

Bruce, R. & Hoehn, L. (1980). *Supervisory practice in Georgia and Ohio*. Paper presented at the Annual Meeting of the Council of Professors of Instructional Supervision, Hollywood, FL.

Burke, P. & Krey, R. (2005). *Supervision: A guide to instructional leadership*. Springfield, IL: Charles C Thomas.

Burnett, P., Fanshawe, J. (1997). Measuring school-related stressors in adolescents. *Journal of Youth and Adolescence*, *26*(4), 415-428.

Carmichael, L., Dearborn, W. (1947). *Reading and visual fatigue*. Oxford, England: Houghton Mifflin.

Cassady, J., Johnson, R. (2002). Cognitive test anxiety and academic performance. *Contemporary Educational Psychology*, *27*(2), 270-295.

Center for Public Education. (2006). A guide to standardized testing: The nature of assessment. Accessed October 14, 2017, from http://www.centerforpubliceducation.org/Main-Menu/Evaluating-performance/A-guide-to-standardized-testing-The-nature-of-assessment

Chapman, P. (1988). *Schools as sorters*. New York, NY: New York University Press.

College Board. (2017). Validity Evidence. Accessed October 10, 2017, from https://research.collegeboard.org/services/aces/validity/handbook/evidence#consequential validity

Crouse, J., Trusheim, D. (1988). *The case against the SAT*. Chicago, IL, University of Chicago Press.

Cubberly, E. (1922). *Public school administration: a statement of the fundamental principles underlying the organization and administration of public education*. Boston, MA: Houghton Mifflin Co.

Davis, D. (1946). The disorganization of behavior in fatigue. *Journal of Neurology and Psychiatry*, *9*, 23-29.

Davis, D. R. (1946). This disorganization of behaviour in fatigue. *Journal of Neurology, Neurosurgery, and Psychiatry*, *9,* 23–29.

Dewey, J. (1938). *Experience and education*. New York, NY. Kappa Delta Pi.

Dillon, A. (1994). Reading from paper versus screens: a critical review of the empirical literature. *Ergonomics*, *35*(10), 1297-1326.

Ebbinghaus, H. (1896-97). On a new method for testing mental abilities and its use with school children. *Zeitschrift für Psychologie und Psysiologie der Sinnesorgane*, *13*, 401–459.

Elliott, S., Marquart, A. (2003, January). *Extended time as an accommodation on a standardized mathematics test: An investigation of its effects on scores and perceived consequences for students with varying mathematical skills*. Wisconsin Center for Education Research, Paper No. 2003-1.

Embretson, S. E. (2007). Construct validity: a universal validity system or just another test evaluation procedure? *Educational Researcher*. Accessed October 10, 2017, from http://journals.sagepub.com/doi/abs/10.3102/0013189X07311600.

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

Gallagher, C. J. (2003). Reconciling a tradition of testing with a new learning paradigm. *Educational Psychology Review, 15*(1), 83-99.

Goddard, H. (1913). *Standard method for giving the Binet test*. Vineland, NJ, Training School.

Goldhammer, R. (1969). *Clinical supervision: Special methods for the supervision of teachers*. New York, NY: Holt, Rinehart, and Winston.

Goldstein, D., Hahn, C. S., Hasher, L., Wiprzycka, U. J., Zelazo, P. D. (2006). Time of day, intellectual performance, and behavioral problems in morning versus evening type adolescents: is there a synchrony effect? *Personality and Individual Differences, 42*(3), 431-440.

Goldstein, J., Behunaik, P. (2005, August). Growth models in action: Selected case studies. *Practical assessment, research & evaluation, 10*(11).

Goodenough, F. (1949). *Mental testing: Its history, principles and applications*. New York, NY: Rinehart.

Guilikers, J., Bastiaens, T. J., Kirschner, P. A. (2004). A fiver-dimension framework for authentic assessment. *Educational Technology Research and Development, 52*(3), 67-86.

Hanson, F. A. (1993). *Testing testing: Social consequence of the examined life*. London, England: University of California Press.

Hawkins, B. (1995). A multiple choice mushroom: Schools, colleges rely more than ever on standardized tests. *Black Issues in Higher Education, 11*(25).

Hays, D. (2017). *Assessment in counseling: Procedures and practices*. Alexandria, VA: American Counseling Association.

Hendricks, J. (1967). *The Iowa Tests of Educational Development as predictors of academic success at Utah State University*. All Graduate Theses and Dissertations. Paper 2831.

Huang, H.M. (2002). Towards constructivism for adult learners in online learning environments. *British Journal of Educational Technology, 33,* 27-37.

Huddleston, A. P., Rockwell, E. C. (2015). Assessment for the masses: A historical critique of high-stakes testing in reading. *Texas Journal of Literacy Education*, *3*(1), 38-49

Kandel, I. L. (1936). *Examinations and their substitutes in the United States*. New York: Carnegie Foundation for the Advancement of Teaching.

Kaufman, A. S. (2009). IQ testing 101. New York, NY: Springer Publishing Company.

Lemann, N. (1999). *The big test: The secret history of the American meritocracy*. New York City, NY: Farrar, Straus and Giroux.

Mandinach, E., Bridgeman, B., Cahalan-Latusis, C., Trapani, C. (2005). The impact of extended time on SAT test performance. *College Board Research Reports, 8.*

Martyn, G. (1912). A study of mental fatigue. *British Journal of Psychology, 5,* 427-446.

Marzano, R., Frontier, T., Livingston, D. *(2011). Effective supervision: Supporting the art and science of teaching.* Alexandria, VA: ASCD.

Morgan, J. G. (2004). Tennessee's graduation exams: past, present, and future. *Office of Education Accountability*, (November). Retrieved August 14, 2017, from http://www.comptroller.tn.gov/Repository/RE/highstaketest.pdf

National Commission on Excellence in Education (1983). *A nation at risk: The imperative for educational reform*. Washington, DC: National Commission on Excellence in Education.

National Commission on Testing and Public Policy. (1990). *From gatekeeper to gateway: Transforming testing in America*. Chestnut Hill, MA.

New York (State) Superintendent of Common Schools. (1845). *Annual report of the superintendent of common schools*. Albany, NY.

Noyes, J. M., Garland, K. J. (2008). Computer- vs. paper-based tasks: Are they equivalent? *Ergonomics*, *51*(9), 1352-1375.

Perucca, D. (2013). *Divided timed and continuous timed assessment protocols and academic performance.* Walden University, United States of America.

Peterson, J. (1983). *The Iowa testing programs*. Iowa City, IA: University of Iowa Press.

Pope, N. G. (2016, March). How the time of day affect productivity: evidence from school schedules. *The Review of Economics and Statistics*, *98*(1), 1-11.

Powell, J. C. (1993). What does it mean to have authentic assessment? *Middle School Journal*, *25*(2), 36-42.

Randler, C., Frech, D. (2009). Young people's time-of-day preferences affect their school performance. *Journal of Youth Studies, 12*(6), 653-667.

Rothman, R. (1995). *Measuring up: Standards, assessment, and school reform*. San Francisco, CA: Jossey-Bass.

Rothman, R. (1995). *Tests of significance.* San Francisco, CA: Jossey-Bass.

Sanders. W. (2003, April). *Beyond No Child Left Behind*. Paper presented at the annual meeting of the American Educational Research Association. Chicago, IL.

Sarason, I. G., Harmatz, M. G. (1965). Test anxiety and experimental conditions. *Journal of Personality and Social Psychology, 1*(5), 499-505.

Sarason, I., Stoops, R. (1978). Test anxiety and the passage of time. *Journal of Consulting and Clinical Psychology*, *46*(1), 102-109.

Savery, J. R., Duffy, T. M. (1995). Problem based learning: An instructional model and its constructivist framework. *Educational Technology*, *35*(5), 31-38.

Schulte, A., Elliott, S., & Kratochill, T. (1999). Effects of testing accommodations on standardized mathematics test scores: An experimental analysis of the performances of students with and without disabilities. *School Psychology Review, 30*(4), 527-547.

SCORE. (2014). Measuring student growth in Tennessee: Understanding TVAAS. Retrieved Nov 1, 2017 from http://tnscore.org/wp-content/uploads/2014/10/TVAAS-Memo.png

Segool, N. (2009). *Test anxiety associated with high-stakes testing among elementary school children: Prevalence, predictors and relationship to student performance* (Doctoral dissertation). Retrieved from ProQuest.

Siegle, D. (2017). *Educational research basics*. University of Connecticut. Retrieved October 10, 2017 from: https://researchbasics.education.uconn.edu/t-test/

Sievertsen, H. H., Gino, F. & Piovesan, M. Cognitive fatigue influences students' performance on standardized tests. *Proceedings of the National Academy of Sciences in the United States of America, 113*(10), 2621-2624.

Stern, W. (1914). *The psychological methods of testing intelligence*. Baltimore, MD: Warwick & York.

Taylor, F. (1911). *The principles of scientific management*. New York, NY: Harper & Brothers.

Tennessee Department of Education. (2010, April 26). High school transition policy frequently asked questions. Retrieved August 14, 2017, from http://www.scsk12.org/schools/wooddale.hs/site/documents/HSTransitionPolicyFAQ.pdf

Tennessee Department of Education. (2015, September). Tennessee task force on student testing and assessment. Retrieved August 14, 2017, from https://tn.gov/assets/entities/education/attachments/tst_assessment_task_force_report.pdf

Tennessee Department of Education. (2017, April). Test administration manual. Retrieved October 10, 2017, from, https://tn.gov/assets/entities/education/attachments/tst_tcap_tam_spring_2017_eoc_online.pdf

U.S Department of Education. (2013). Testing integrity symposium: issues and recommendations for best practices. *National Center for Education Statistics*. Retrieved October 10, 2017, from https://nces.ed.gov/pubs2013/2013454.pdf

Walsh, W. B., Betz, N. (1995). *Tests and assessment*. Englewood Cliff, NJ: Prentice-Hall.

Walz, L., Albus, D., Thompson, S., & Thurlow, M. (2000). Effect of a multiple day test accommodation on the performance of special education students. *National Center on Educational Outcomes*, (December). Retrieved August 14, 2017, from https://nceo.info/Resources/publications/OnlinePubs/archive/AssessmentSeries/MnReport34.html

Wastlund, E., Reinikka, H., Norlander, T., Archer, T. (2004). Effects of VDT and paper presentation on consumption and production of information: Psychological and physiological factors. *Computers in Human Behavior*, *21*, 377-394.

Wetzel, W. (1929). Scientific supervision and curriculum building. *The School Review*, *37*(2), 112-123.

Wirtz, W. (1977). *On further examination: Report of the advisory panel on the Scholastic Aptitude Test score decline*. New York City, NY: College Entrance Examination Board.

Wright, R. (2012). Recent policy changes in Tennessee: Teacher evaluations. Retrieved October 28, 2017 from Tennessee Comptroller of the Treasury website: http://www.comptroller.tn.gov/Repository/RE/Teacher%20Evaluations.pdf

Zeidner, M. (1998). *Test anxiety: The state of the art*. New York, NY: Plenum.

Zwick, R. (2004). *Rethinking the SAT: The future of standardized testing in university admissions.* New York, NY: RoutledgeFarmer.

Appendix A

Parent Consent Letter

Appendix A

Parent Consent Letter

Information Letter and Consent Form for Parents or Guardians
Permission for Research with Children

Date Pending Approval

Dear **Parent(s) or Guardian(s)**:

I am writing to ask permission for your child to participate in a Carson Newman University research project on multiple subtest exams. This project will be conducted at Science Hill High School over three to four days. We are interested in determining if multiple subtest exams impact student performance. The project is designed to help us understand more about children's performance on TCAP exams

The project in which your child has been invited to participate is expected to require no of time out of class. The assessments will be given during a normal class period. However, the decision about participation is yours. To help you in this decision, a brief description of the project is provided. Students will be administered a retired ACT exam. This ACT exam will be broken down into sections over multiple days. The amount of days varies by class. The student will receive a projected ACT composite score. The projected ACT composite score will be used in the research study data analysis.

All children's performances are considered confidential and individual children's results will not be shared with school staff. However, information based on the results of the group of participants will be provided. Only high school students who have parental permission, and who themselves agree to participate, will be involved in the study. Also, children or parents may withdraw their permission at any time during the study without penalty by indicating this decision in writing to the researcher. There are no known or anticipated risks to participation in this study.

I would like to assure you that this study has been reviewed and approved by the Institutional Review Board at Carson Newman University. In addition, it has the support of the principal at your child's school. However, the final decision about the participation is yours. Should you have any concerns or comments resulting from your child's participation in this study, please contact Dr. Deborah Hayes, dlhayes@cn.edu.

We would appreciate it if you would permit your child to participate in this project, as we believe it will contribute to furthering our knowledge of multiple administration testing. Please complete the attached permission form, whether or not you give permission for your child to participate, and return it to the school by (**date pending approval).**

If you have any questions about the study, or if you would like additional information to assist you in reaching a decision, please feel free to contact me Aaron Wood at WoodA@jcschools.org 423-232-2190 or my faculty supervisor, Dr. Deborah Hayes at, DHayes@cn.edu. Thank you in advance for your interest and support of this project.

Sincerely,

Aaron Wood
Testing, AP, IB Coordinator
Science Hill High School

Appendix B

Student Consent Letter

Appendix B

Student Consent Letter

Consent Form – Child
(Accompanies the information letter about the study)

I have read the information letter concerning the research project entitled Do Multiple Subtest Exams Impact Student Performance conducted by Aaron Wood at Carson Newman University. I have had the opportunity to ask questions and receive any additional details I wanted about the study.

I acknowledge that all information gathered on this project will be used for research purposes only and will be considered confidential. I am aware that permission may be withdrawn at any time without penalty by advising the researchers.

I realize that this project has been reviewed by and approved by the Institutional Review Board at Carson Newman University, and that I may contact this office if I have any comments or concerns about my son or daughter's involvement in the study.
If I have any questions about the study I can feel free to call Aaron Wood at WoodA@jcschools.org 423-232-2190

☐ Yes – I would like my child to participate in this study

☐ No – I would not like my child to participate in this study.

Child's Name **(please print)** _____

Child's Birth Date _____ Gender of Child ____ Male ____ Female

Parent or Guardian Signature _____ Date _____

Student Signature _____ Date _____