Teacher Evaluation and Other Implications of Student Growth and Achievement in Public Schools in Georgia

A Dissertation submitted
to the Graduate School
Valdosta State University

in partial fulfillment of requirements
for the degree of

DOCTOR OF EDUCATION

in Organizational Leadership

in the Department of Leadership, Technology, and Workforce Development
of the James L. and Dorothy H. Dewar College of Education and Human Services

May 2022

Carrie O'Bryant

Ed.S., Brenau University, 2006
M.Ed., Brenau University, 2003
B.B.A., Kennesaw State University, 2000

This dissertation, “Teacher Evaluation and Other Implications of Student Growth and Achievement in Public Schools in Georgia” by Carrie O’Bryant, is approved by:

Dissertation Committee Chair

J L Pate

James L. Pate, Ph.D.
Professor of Curriculum, Leadership, and Workforce Development

Committee Members

Lantry L. Brockmeier

Lantry L. Brockmeier, Ph.D., Researcher
Professor of Curriculum, Leadership, and Workforce Development

Barbara J. Radcliffe, Ph.D.
Professor of Teacher Education

Associate Provost For Graduate Studies and Research

Becky K. da Cruz

Becky K. da Cruz, Ph.D., J.D.
Professor of Criminal Justice

Defense Date April 6, 2022

FAIR USE

This dissertation is protected by the Copyright Laws of the United States (Public Law 94-553, revised in 1976). Consistent with fair use as defined in the Copyright Laws, brief quotations from this material are allowed with proper acknowledgment. Use of the material for financial gain without the author's expressed written permission is not allowed.

DUPLICATION

I authorize the Head of Interlibrary Loan or the Head of Archives at the Odum Library at Valdosta State University to arrange for duplication of this dissertation for educational or scholarly purposes when so requested by a library user. The duplication shall be at the user's expense.

Signature ______________________________

I refuse permission for this dissertation to be duplicated in whole or in part.

Signature ______________________________

ABSTRACT

There is a longstanding debate over how to evaluate teachers. Various models of teacher evaluation tools including observations, teaching artifacts, or frequent walkthroughs, provide evidence of daily classroom performance. Although there are questions on the most efficient means by which to complete evaluations, most agree evaluative tools must focus on teacher growth. The Georgia TKES evaluative tool, enacted in 2014, was created to address the multifaceted subject of teacher evaluation.

The study employed a non-experimental, quantitative design to determine the extent by which TKES scores, academic setting (self-contained or departmentalized), and levels of the percentages of ED students predict student growth and student achievement. Additionally, the reliability and interpretability of the TKES components were examined. Data representing 4,000 fourth and fifth grade ELA and math teachers were provided by the Georgia Department of Education.

Results from the study indicated standards predictive of student growth and achievement by subject and grade. In fourth grade ELA, the following standards were predictive of student growth: 4 (proficient and exemplary) and 8 (exemplary). Standards 9 (proficient level) and 3 (proficient and exemplary levels) had a negative impact. In fourth grade math, the following standards were predictive of student growth: 1 (exemplary) and 3 (exemplary). There were no standards associated with student growth in fifth grade ELA or math. For student achievement, the following standards were predictive of achievement in fourth grade ELA: 3 (exemplary), 7 (exemplary), 8 (exemplary), and 10 (exemplary). Standard 8 (needs development) was associated with a negative impact on student achievement. In fourth grade Math, the following standards at exemplary levels were associated with predicting student achievement: 1, 2, 7, and

8. In fifth grade ELA and math, the following standards at exemplary levels were associated with predicting student achievement: 1, 4, 7, 8, and 9. The study revealed there was a significant difference between levels of economically disadvantaged students and student achievement across fourth and fifth grade ELA and math. Setting was also significant in fourth grade math. In the final research question, TKES Standards were reduced to two components for fourth grade ELA and three components for the remaining sub-questions.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

DEDICATION

This dissertation is dedicated to my family. My husband, Scott, supported me through this process. Sometimes, that wasn't easy. He covered practices, duties, etc., so that I could work on my "paper" and remained steady through all my ups and downs. Emma, my daughter, was one of my biggest supporters. She constantly told me she was proud of me, even though she was probably the most affected by my workload. Jake and Josh, my twins, along with their precious companions, Abbie and Carson, also provided constant support. Finally, my mom and dad never doubted my ability to complete this degree. Overall, my family's role was too significant to claim this degree as my own.

I also thank God for helping me through this trying time. When I applied to begin my Doctorate, I had no idea what challenges were headed my way. I began my administration career at almost the same time I started this degree. During my program, I served as an Assistant Principal and then a Principal. I said so many prayers during that time, and I know God provided direction that allowed me to navigate this stressful time. His "nudges" often kept me going and gave me reassurance and guidance.

I am so thankful for the support of so many others, including colleagues and friends. There are too many to name, but each of them provided support in their own way. However, I must name specifically my friend, Joey Moss. We began this journey together, we kept each other going, and I am so thankful for her constant support.

Everyone in my circle had a role in supporting me, and I am so thankful. I am also so thrilled to slow down, support those who have supported me, and love on my dogs, who sat with me for countless hours as I worked.

ACKNOWLEDGEMENTS

Thank you to my chair, Dr. Pate. He was a constant source of support. His reassurance often made me believe I could finish this, and he talked me off a “cliff” more than once. Also, Dr. Brockmeier pushed me much farther than I wanted, but as he said, “In the end, we will have something for which we could be proud.”

Chapter I

INTRODUCTION

Overview

Throughout the history of formal education, educational leaders stressed the need to consistently evaluate teaching and learning in schools to maximize student learning opportunities. Although the publication of *A Nation at Risk (1983)* brought an increased urgency for education reform in the United States, there are documented attempts to improve education in other areas of the world for more than 600 years (Barrette, Morton, & Tozcu, 1995; The National Commission on Excellence in Education, 1983). Multiple reports including the No Child Left Behind Act of 2001 (2002) and the Every Student Succeeds Act (2017) have echoed similar findings of *A Nation at Risk*, indicating a need to increase the quality of education.

While most educators agree there is a need to improve education, the means to accomplish the improvement is a great source of debate, likely due to the many factors affecting education. Examples of these factors included the educational policy, education programs within the school, school leadership, instructional format, teacher quality, parent's education level, socioeconomic status, overall health, nutrition, level of home support, etc. Because many of these variables were outside the scope of an educator's control, educational leaders must focus on the factors they can control including teacher evaluations and the classroom setting.

Many educational leaders agreed that, among the factors over which they have influence, improving teacher evaluation was the most effective way to improve the overall quality of teaching and learning (Barrette et al., 1995; Carbough, Manzano, & Toth, 2017). Research indicated a direct link between effective teaching and student achievement. While there was a distinct need to monitor education quality through an effective teacher evaluation system, there remains much debate on evaluating teachers effectively (Marzano, Frontier, & Livingston, 2011).

Perhaps due to the complexity of teacher evaluation and the difficulty of applying a one-size, fits all approach, it is difficult for educational leaders and researchers to develop an evaluative system that meets the diverse needs of educational systems. Many models fell short and failed to impact the critical areas surrounding student achievement and growth. Unfortunately, most education evaluation systems highlighted poor practices and other factors affecting education but failed to guide teacher growth (Marzano et al., 2011). As a result, mediocre teaching was widely acceptable (Weisburg, Sexton, Mulhern, & Keeling, 2009). Effective teacher evaluative systems must celebrate instructional successes while highlighting areas for potential teacher growth. (Marzano et al., 2011).

In alignment with Marzano's views, researchers cautioned against forming ineffective teacher evaluation systems. Such systems do not increase the overall quality of instruction (Toch & Rothman, 2008). Effective evaluation methods must focus on increasing student achievement and student growth; the systems must not be used as punitive measures (Tucker & Stronge, 2005). The evaluation process must highlight areas for development and current successes in the classroom. These evaluations must include research-driven strategies for optimal growth and success (Anderson, Butler, Palmiter, & Arcaira, 2016).

One such evaluative tool, the Teacher Keys Effectiveness System (TKES), was established in Georgia to serve as a useful evaluative tool and provide Georgia teachers with avenues for professional growth. The primary goal of TKES was to optimize student learning and growth by improving the quality of instruction. Through the 10 performance standards, TKES incorporated multiple factors affecting teaching and learning to produce an overall, comprehensive score for teachers (Teacher-Keys-Effectiveness-System, n.d.).

Conceptual Framework

The concepts guiding this study included the standards of the Teacher Assessment on Performance Standards (TAPS). They are as follows: Professional Knowledge, Instructional Planning, Instructional Strategies, Differentiated Instruction, Assessment Strategies, Assessment Uses, Positive Learning Environment, Academically Challenging Environment, Professionalism, and Communication (Leader Keys Effectiveness System: Fact sheets, 2012). These standards served as independent variables for all research questions except research question 3. In research question 3, the level of the percent of economically disadvantaged (ED) students and academic setting (self-contained or departmentalized) serve as the independent variables.

John Hattie's theory of high-impact strategies explained the interactions between these variables. Although not identical, Hattie focused on similar themes that closely mirror TAPS Standards. Hattie's most effective strategies are directly connected to the standards highlighted in the TAPS within the Teacher Keys Effectiveness System (TKES) platform (Hattie, 2008). Through his research and meta-analyses, Hattie constructed a list of strategies that affect student achievement, both negatively and positively. His theories are widely accepted in the field of school improvement (Fisher, Frey, & Hattie, 2016) and commonly serve as a reference when considering areas for best practices. Hattie's research was considered a milestone in educational reform, and it is broadly accepted as a guideline for evaluating potential areas for school improvement (Terhart, 2011).

Hattie categorized his strategies based on effectiveness in increasing student achievement. Based on a numerical value, these strategies were rated as the following: having the potential to accelerate learning considerably, having a positive impact, having a small positive impact, or having an adverse impact. The potential for the strategy to accelerate learning increases as values approach 1.0. Likewise, as values fall closer to 0 or even into negative values, the likelihood that strategies have little or negative impact increases (Fisher et al., 2016). For example, the effect size of .40 is considered the hinge point for an acceptable level of growth during a typical instruction year. An effect size of .2 is deemed to be small, .40 is regarded as a medium or average effect, and .60 or larger is classified as a high or large effect exceeding the typical growth within a year of instruction. Effect sizes of less than zero are considered to have a negative impact on learners (Hattie, 2008).

For each of the standards within the TAPS, there are closely related Hattie strategies. Because of the alignment between these standards and Hattie strategies, the Hattie framework provided an appropriate theory on which to base this research. For example, for the standard of Professionalism, Hattie identified the strategy of Teacher Credibility with a high effect size of .9 and the potential to accelerate learning considerably (*250 + influences on student achievement,* 2017). Additionally, many other examples of strategies were closely related to standards within the TAPS. Assessment Strategies and Uses were relevant to Feedback, Evaluation, and Reflection. Differentiated instruction was closely associated with the Hattie strategy of Response to Intervention. Teacher Clarity could be associated with Communication and Classroom Discussion. Conversely, individualized instruction, which could result from Differentiation and Assessment, was rated at .23 (*250 + influences on student achievement,* 2017).

Interestingly, again in alignment with the TKES observation format, Hattie (2008) contended that classroom observation should focus on teacher evaluation. Evaluative systems must concentrate on teacher behaviors that are observed during instruction. Hattie argued that administrators could provide a more pertinent gauge for teacher efficacy due to observations rather than what can be provided by educational assessments (Fisher et al., 2016).

Although not entirely contrary to Hattie's ideas, Marzano's model focused less on classroom observations for specific strategies or standards. Instead, Marzano highlighted the importance of planning and conferencing with the teacher to achieve professional growth. Marzano relied heavily on student achievement as evidence of effective teaching (Carbaugh et al., 2017). The diagnostic model recognized effective instruction through student evidence. The model highlighted 23 essential behaviors related to teacher effectiveness. These behaviors were categorized into the following categories: standards-based planning, standards-based instruction, conditions for learning, and professional responsibilities ("The focused teacher evaluation model: Summary and implementation", 2020).

Cheung and Slavin (2016) did not base their opinions on teacher effectiveness or the potential to impact student achievement through the implementation of numerous strategies. Instead, Cheung and Slavin (2016) believed each educational approach must be studied singularly. Further, educators should narrow their instructional efforts to a limited number of proven strategies and perfect the strategy's delivery for increased student achievement to occur (Cheung & Slavin, 2016). Similarly, the view of limited, well-developed strategies closely resembled Mike Schmoker's opinions as outlined in *Focus: Elevating the Essentials to Radically Improve Student Learning* (2014). Schmoker (2014) believed educators relied consistently on a small number of research-driven strategies. Schmoker believes teachers should become masters of best practices. Educators must resist the temptation to jump from one educational fad to another (Schmoker, 2014). He added, however, that educators must not become bored with the routine implementation of a select number of strategies. The known strategies must be implemented, without fail, so that students receive the maximum benefit. Schmoker also cautioned against relying too heavily on specific assessments to gauge student performance and make sweeping instructional changes based on various instruments.

In addition to understanding these influential theorists with popular ideas regarding effective instruction and school improvement, understanding the standards within TAPS, student growth percentile (SGPs), and student achievement is imperative. For each standard, a rubric was published to aid the administration in selecting the appropriate score for teachers. The rubrics essentially require the evaluator to determine the level at which the teacher executes each function. The scale for teachers for each standard is as follows: (4) exemplary - continually demonstrates the standard, (3) proficient - consistently demonstrates the standard, (2) needs development - inconsistently meets the standard, and (1) ineffective – does not display the behaviors associated with the standard ("Leader Keys Effectiveness System: Fact sheets", 2012).

Depending on their career longevity and proficiency level, teachers are placed on different evaluation plans within the TKES platform. Teachers with three or fewer years of experience are rated on these standards six times each year under the Full Plan. Two of the observations are full formatives. Formatives are 30-minute observations, during which teachers are rated on all ten standards. The remaining four observations are shorter, ten-minute snapshots of teachers' progress toward standards. During these shorter observations, teachers are rated on fewer standards. Experienced teachers with perceived shortcomings may also be evaluated on the Full Plan (Leader Keys Effectiveness System: Fact sheets, 2012).

Teachers with an adequate academic standing, or level 3, with more than three years of teaching experience, are evaluated on the Flexible Plan. The Flexible Plan consists of three observations. One observation is a thirty-minute formative observation. The other two observations are ten-minute observations. If teachers maintain a level 3 rating, they remain on the Flexible Plan (Leader Keys Effectiveness System: Fact sheets, 2012).

Understanding the dependent variables, the SGP and Georgia Milestones Assessment System (GMAS) student mean scale score (MSS), is also imperative to understand the variables' interaction. The student growth model's overarching goal is to provide educators with an authentic picture of student progress. Students are compared with peer groups to determine if their growth pattern is in line with peers, below peers, or higher than peers. Essentially, a student's growth is compared to a student with a similar academic history. Students exceeding the scores of their peer group will have a high student growth percentile or SGP. Likewise, students with less growth than their peer group will have a lower SGP. Teachers are assigned an SGP based on the mean of their class (*A guide to the Georgia student growth model*, 2014). The other dependent variable utilized in this study was the GMAS MSS.

Understanding the overall role of the TAPS within the context of this study was essential to meet the needs of the study so that educational leaders may learn to isolate the standards with the most significant impact on student achievement and, therefore, could increase the emphasis on providing professional development for these standards. The application of standards and how the standards apply to student achievement could add further knowledge regarding how teachers implement the TKES framework strategies and the impact on student growth percentiles. Because Hattie's strategies are closely related to TKES platforms' standards, the theory provides a framework on which to base the study.

Statement of Problem

As educators continue to strive to increase student achievement, teacher evaluation systems must provide direction, so administrators and teachers understand the area in which they need to grow professionally and identify critical factors in addressing student growth and achievement. Further, professional development opportunities must focus on areas with a strong relationship with student growth and achievement. To provide the essential feedback necessary for teacher development, evaluative tools must promote teacher reflection and growth and provide direction regarding professional development. Reflection and development must be directly related to the areas that significantly impact teacher growth and student progress and achievement. Identifying the standards most closely associated with student growth is critical as there is a gap in the literature regarding the relationship between standards and student growth and achievement. Further, understanding the overall reliability and validity of the TAPS instrument is critical if educators are to use it as a tool on which to make crucial decisions.

Purpose of the Study

The purpose of this study was to clarify the relationship between individual TAPS summative standard scores, student growth percentiles, and student achievement scores for students in fourth and fifth grade to determine if specific standards are predictive of student growth percentiles or student achievement. The study identified the predictability between TAPS standard scores and student growth percentiles, specifically in reading and math. Additionally, the study determined whether academic setting, self-contained or departmentalized, and the levels of the percentage of ED students predicted student achievement. Finally, this study's results will allow educational leaders to isolate the standards with the most significant impact on student achievement so they may increase the emphasis on providing professional development for these standards. Additionally, following the results from this study, administrators may make more informed hiring decisions based on an individual teacher’s TAPS scores so that new faculty members have the necessary skills to maximize student growth. Using the evidence from this study, educators may make these decisions with confidence following the further determination of the reliability and validity of the TKES instrument.

Research Questions

Each of the questions below, including the sub-questions, guided this study:

RQ1: Are summative scores on TKES Standards (Professional Knowledge, Instructional Planning, Instructional Strategies, Differentiated Instruction, Assessment Strategies, Assessment Uses, Positive Learning Environment, Academically Challenging Environment, Professionalism, and Communication) significant predictors of the teacher's SGP level on the Georgia Milestones Assessment System (GMAS)?

a. Are summative scores on TKES Standards significant predictors of the teacher's SGP level (levels I, II, III, or IV) on the fourth grade English/Language Arts portion of the Georgia Milestones Assessment System (GMAS)?

b. Are summative scores on TKES Standards significant predictors of the teacher's SGP level (levels I, II, III, or IV) on the fourth grade Math portion of the Georgia Milestones Assessment System (GMAS)?

c. Are summative scores on TKES Standards significant predictors of the teacher's SGP level (levels I, II, III, and IV) on the fifth grade English/Language Arts portion of the Georgia Milestones Assessment System (GMAS)?

d. Are summative scores on TKES Standards significant predictors of the teacher's SGP level (levels I, II, III, and IV) on the fifth grade Math portion of the Georgia Milestones Assessment System (GMAS)?

RQ2: Are summative scores on TKES Standards (Professional Knowledge, Instructional Planning, Instructional Strategies, Differentiated Instruction, Assessment Strategies, Assessment

Uses, Positive Learning Environment, Academically Challenging Environment, Professionalism, and Communication) significant predictors of the teacher's mean scale score on the Georgia Milestones Assessment System (GMAS)?

a. Are summative scores on TKES Standards significant predictors of the teacher's mean scale score on the fourth grade English/Language Arts portion of the Georgia Milestones Assessment System (GMAS)?

b. Are summative scores on TKES Standards significant predictors of the teacher's mean scale score on the fourth grade Math portion of the Georgia Milestones Assessment System (GMAS)?

c. Are summative scores on TKES Standards significant predictors of the teacher's mean scale score on the fifth grade English/Language Arts portion of the Georgia Milestones Assessment System (GMAS)?

d. Are summative scores on TKES Standards significant predictors of the teacher's mean scale score on the fifth grade Math portion of the Georgia Milestones Assessment System (GMAS)?

RQ3: Is there a significant difference in academic setting (departmentalized or self-contained) by level of economically disadvantaged (ED) students on the teacher's mean scale score on the Georgia Milestones Assessment System (GMAS)?

a. Is there a significant difference in academic setting (departmentalized or self-contained) by level of economically disadvantaged (ED) students on the teacher's mean scale score on the fourth grade English/Language Arts portion of the Georgia Milestones Assessment System (GMAS)?

b. Is there a significant difference in academic setting (departmentalized or self-contained) by level of economically disadvantaged (ED) students on the teacher's mean scale score on the fourth grade Math portion of the Georgia Milestones Assessment System (GMAS)?

c. Is there a significant difference in academic setting (departmentalized or self-contained) by level of economically disadvantaged (ED) students on the teacher's mean scale score on the fifth grade English/Language Arts portion of the Georgia Milestones Assessment System (GMAS)?

d. Is there a significant difference in academic setting (departmentalized or self-contained) by level of economically disadvantaged students on the teacher's mean scale score on the fifth grade Math portion of the Georgia Milestones Assessment System (GMAS)?

RQ4: How many reliable and interpretable components are there among the following variables: Professional Knowledge, Instructional Planning, Instructional Strategies, Differentiated Instruction, Assessment Strategies, Assessment Uses, Positive Learning Environment, Academically Challenging Environment, Professionalism, and Communication?

a. How many reliable and interpretable components are there among the following variables: Professional Knowledge, Instructional Planning, Instructional Strategies, Differentiated Instruction, Assessment Strategies, Assessment Uses, Positive Learning Environment, Academically Challenging Environment, Professionalism, and Communication among fourth grade English/Language Arts teachers?

b. How many reliable and interpretable components are there among the following variables: Professional Knowledge, Instructional Planning, Instructional Strategies, Differentiated Instruction, Assessment Strategies, Assessment Uses, Positive Learning

Environment, Academically Challenging Environment, Professionalism, and Communication among fourth grade math teachers?

c. How many reliable and interpretable components are there among the following variables: Professional Knowledge, Instructional Planning, Instructional Strategies, Differentiated Instruction, Assessment Strategies, Assessment Uses, Positive Learning Environment, Academically Challenging Environment, Professionalism, and Communication among fifth grade English/Language Arts teachers?

d. How many reliable and interpretable components are there among the following variables: Professional Knowledge, Instructional Planning, Instructional Strategies, Differentiated Instruction, Assessment Strategies, Assessment Uses, Positive Learning Environment, Academically Challenging Environment, Professionalism, and Communication among fourth grade math teachers?

Methodology

This study utilized a quantitative approach to answer the research questions from data provided by the Georgia Department of Education (GADOE). The random sample included fourth grade and fifth grade ELA and math teachers. The TAPS summative scores for individual fourth and fifth grade teachers, teachers' student growth level on the Georgia Milestones Assessment System (GMAS), teachers' mean scaled score, again, as indicated by the GMAS, the academic setting (self-contained and departmentalized), and percentage of free and reduced lunches served as variables.

The appropriate quantitative analysis was performed to answer each of the research questions. An ordinal logistic regression analysis was used in Question 1. Question 2 utilized a factorial ANOVA. Separate linear regression models were conducted for Question 3, and a factorial analysis was used in Question 4.

Definition of Key Terms

Economically Disadvantaged (ED). The percentage of ED is calculated by dividing the number of students eligible to receive free or reduced-priced meals by the total school enrollment. (Every Student Succeeds Act reporting requirements, n.d.)

Estimated Marginal Mean (EMM). The estimated marginal mean is the mean of Y for each group of the independent variable at one specific value of the covariate.

Mean Scale Score (MSS). The mean is the average of a set of scale scores. The mean scale score is found by adding all the scale scores in a given distribution and dividing that sum by the total number of scale scores. (Georgia's Teacher Keys Effectiveness System, n.d.)

Student Growth Percentiles (SGP's). Established by the Georgia Department of Education, student growth percentiles detail a comparison for students of any score level compared to like students to establish a level of growth for teacher and school accountability purposes (*A guide to the Georgia student growth model*, 2014).

Teacher Assessment on Performance Standards (TAPS). Standards developed by the Georgia Department of Education explicitly state the criteria by which Georgia educators will be scored (Leader Keys Effectiveness System, 2018).

Teacher Effectiveness Measure (TEM). Teachers are assigned a TEM score, which indicates a teacher's overall progress toward meeting or exceeding TKES Standards, or TAPS (Leader Keys Effectiveness System, 2018).

Teacher Keys Effectiveness System (TKES). TKES is the current teacher evaluative system in Georgia by which all Georgia educators are held accountable. (Leader Keys Effectiveness System, 2018).

Value-Added Model (VAM). Value-added models are evaluative systems that use student test scores to measure a teacher's effectiveness. VAMs use statistical measures to account for student characteristics and estimate a teacher's contribution to student progress. ("Value-added modeling 101", 2020).

Summary

The following chapters describe the components of the study. Chapter 2 includes a detailed literature review summarizing the history of teacher evaluations, effective teaching practices, and the components of TKES. Chapter 3 outlines the study's quantitative methodology, including the specific analysis, population, and limitations for each question. Chapter 4 provides an analysis of the collected data and tables to clarify the results. Chapter 5 includes conclusions, interpretations, and recommendations for future research.

Chapter II

REVIEW OF LITERATURE

Need for Teacher Evaluation

Since the release of *A Nation at Risk*, American educators have strived to improve education in public schools and to ensure students receive the level of education provided to students in other countries (The National Commission on Excellence in Education, 1983). Due to its bleak outlook on public education in the United States, the report may be partially responsible for the ongoing search for reform throughout the educational system. As commissioned by the Secretary of Education, *A Nation at Risk* called attention to many public education problems. Among these concerns were a lack of rigor, the need for clearly defined standards, varying requirements for teacher readiness programs, and less than competitive pay for teachers (The National Commission on Excellence in Education, 1983).

Since that time, stakeholders have maintained a call for educational reform. Numerous reports and studies had similar findings as *A Nation at Risk* (The National Commission on Excellence in Education, 1983). For example, in 1985, the Quality Basic Education (QBE) Act called for four significant changes meant to improve education, including teacher certification and evaluation, reading competency to advance to grade two, and equity in funding and resources (Education Quality Basic Education Act). Following the QBE Act in 2001, the No Child Left Behind Act (2002) emphasized accountability and a focus on reading. Additionally, the Every Student Succeeds Act (ESSA) of 2017 continued to demand action to ensure educators meet all students' needs. Essentially, the pressure placed on American school systems to improve education and increase student growth has gained momentum since the publication of *A Nation at Risk* (Patterson, 2019).

This environment of dissatisfaction in the public-school system and the internal drive for teachers to improve their craft continued to fuel the ongoing desire to increase student achievement. How and why teachers are evaluated remains a focal point for educational reform and schools' improvement in the United States and worldwide. Although most agree an evaluative tool is necessary, there have been longstanding debates and questions surrounding the best way to evaluate teachers (Coe, Aloisi, Higgins, & Major, 2014).

Countless studies have attempted to determine the most effective way to improve education (Cosner, Kimball, Barkowski, Carl & Jones, 2015; Derrington, 2014; Derrington & Campbell, 2015; Donaldson & Papay, 2014; Kowalski & Dolph, 2015; Warnock, 2015; Whitehurst, Chingos, & Lindquist, 2015). The studies found that school improvement was primarily linked to teacher development. Therefore, administrators and educators attempted to find ways to increase student achievement by providing valuable feedback to teachers, allowing for reflective teaching, and encouraging growth (Fisher et al., 2016).

Not only must researchers identify the appropriate tools and processes to increase teacher effectiveness, but the leaders in public schools must be able to duplicate the results. An effective teacher evaluation system could provide the guidance needed for educational leaders (Anderson et al., 2016). However, educators must remain committed to delivering the evaluative tools that encourage professional growth rather than concentrating on other, less constructive teacher attributes and characteristics (Anderson et al., 2016).

History of Evaluations

Historically, there have been many opinions regarding the most effective way to evaluate teachers (Coe et al., 2014). As early as the fifteenth century, educational leaders evaluated teachers in some manner for effectiveness (Barrette et al., 1995). Although the history of teacher evaluations is extensive, there was little importance placed on teacher evaluations until the early 1700s because people did not consider education as a professional discipline. As industry became more important in the 1800s, people began to value acquiring knowledge (Marzano et al., 2011).

During the latter part of the nineteenth century and the early twentieth century, thinkers such as John Dewey and Frederick Taylor recognized multiple purposes for public education. They questioned if learning was to promote democratic ideals or for scientific management. No matter the choice, the value of knowledge increased in typical households. These movements led to the need for more critical teacher evaluations. During this time, many teachers were assigned a grade from A to F for specific lessons. The kinds and amount of constructive feedback varied greatly among administrators and regions (Marzano et al., 2011).

Following World War II, the educational pendulum swung away from the scientific method to a more individualistic approach (Marzano et al., 2011). Administrative responsibilities also increased. Administrators acknowledged the role of instructional leaders. However, the administrator's managerial duties often remained the priority and left little time for teacher development (Marzano et al., 2011).

From the late 1960s to the early 1970s, there was a movement toward clinical supervision and evaluation. Several authors published works specifying practices meant to increase each teacher’s pedagogical growth (Marzano, et al., 2011). Goldhammer (1969) detailed supervision methods in his book entitled *Clinical Supervision: Special Methods for the Supervision of Teachers*. He developed a five-phase process of supervision to guide teachers through reflection. Cogan (1973), a professor at Harvard, was among the first to publicly emphasize that continual instructional improvement should serve as the supervisor’s primary focus.

Even with the improvements, there remained much evidence of teacher evaluation processes' inadequacies, particularly in providing growth opportunities. Supervisors gave teachers very little, or no information regarding evaluations and improvement areas, and most teachers being evaluated were not given areas for growth. Evaluative systems did not recognize top performers or identify underperformers. Information from teacher performance was typically used only in extreme cases for remediation and dismissal purposes (Weisburg et al., 2009).

The next major shift in supervision occurred during the 1980s through the work of Madeline Hunter. Her 7-step model for lessons gave clear guidance for administrators (Hunter, 1980). Further, she identified the purposes for supervisory conferences that also facilitated the guiding work for administrators during conferences. Hunter initiated professional development to create a common language for evaluations (Hosford, 1984). Her work significantly impacted many teacher evaluation processes and gave administrators specific guidance (Marzano et al., 2011).

The mid-1980s served as a critical turning point for dialogue about the teacher evaluation process. Glatthorn, McGreal, and Glickman set the stage for modern teacher evaluation (Marzano et al., 2011). Glatthorn (1984) considered the teacher's individual career goals and suggested that teachers have input over their development. McGreal (1983) reiterated the need for professional development differentiation when he detailed teachers' options based on their instructional needs. Glickman (1985) affirmed the most crucial function of the teacher evaluation must be for instructional improvement.

Following the sweeping changes to teacher evaluation in the 1980s, "The Rand Study," a study conducted to determine the types of evaluation practices occurring in the United States, found that the majority of systems in place still did not allow for pedagogical development

(Wise, Darling-Hammond, McLaughlin, & Bernstein, 1984). Further, the evaluations were inconsistent among school systems. Recommendations resulting from the findings of the study included abandoning prescriptive evaluation methods, aligning goals to system priorities, providing administrators with adequate time for evaluations, delivering ongoing training for administrators, aligning the purpose of evaluations with the goals of the system, allocating targeted resources appropriately, using expert teachers for the supervision of peers, involving teacher organizations, and holding teachers accountable for educational decisions (Wise et al., 1984).

In 1996, Charlotte Danielson published *Enhancing Professional Practice: A Framework for Teaching*. Danielson highlighted 76 elements of quality teaching. Teachers were evaluated on the elements using four performance levels (unsatisfactory, basic, proficient, and distinguished) (Danielson, 1996). Like earlier models, the elements of Danielson's model continued to impact current evaluation systems (Marzano et al., 2011).

Tucker and Stronge (2005) championed the importance of including student achievement with classroom observations when evaluating teachers. They stated an undeniable link between student learning and teacher effectiveness. Therefore, student achievement must be an essential source of feedback when considering teacher effectiveness (Tucker & Stronge, 2005). Toch and Rothman (2008) criticized the lack of consistent teacher evaluations that reflected teacher effectiveness.

Immediately following, in 2009, *The Widget Effect* heavily criticized teacher evaluation practices in the United States (Weisburg et al., 2009). The report was based on a mixed-method data analysis from over 1,300 administrators and 15,000 teachers from four states and twelve school districts. Advisory panels met three times between June 2008 and April 2009 to analyze

Arkansas, Colorado, Illinois, and Ohio evaluations. The districts included in the study were committed to reform and reported facing significant student achievement challenges. For example, some schools faced high rates of economically disadvantaged (E.D.) students, ranging from 42 percent to 84 percent (Weisburg et al., 2009). According to the report, school districts assumed teaching effectiveness was the same from teacher to teacher, with administrators noting very little difference in teaching effectiveness among those evaluated (Weisburg et al., 2009). The authors cited a failure to treat teachers as individuals and recognize individual contributions or identify exceptional teachers. Leaders gave 98% of teachers a satisfactory rating, and half of the districts had not dismissed teachers rated poorly. Almost 75% of teachers did not receive specific feedback that promoted professional growth. Many leaders ignored individual strengths and weaknesses (Weisburg et al., 2009).

Existing Evaluative Models

Most current systems employ many or all components of these primary models or combinations. These commonly used evaluative systems include the value-added model, evaluations based on teacher observations, Danielson's (1996; 2014) Framework Model, and the Marzano Focused Teacher Evaluation Model (Carbaugh et al., 2017: Raudys, 2018). The history of teacher evaluation indicates researchers continue to study and debate which model, or combination of models, are the most effective way to increase student achievement (Robinson, n.d.).

Value-added teacher evaluation models (VAM) attempt to predict students' future scores based on previous scores, or the student's growth, through a specific assessment. One key advantage of the VAM include comparing the same students and their development over time (Carbaugh et al., 2017; Raudys, 2018). The study found that administrators' time commitment decreased as the evaluation's primary focus was based on a growth calculation. Disadvantages include the possibility that teachers' ratings are affected by student characteristics rather than a teacher's ability. The VAM growth model also identified the best and worst teachers, but a broad band of teachers was deemed satisfactory. Additionally, if students were high achievers, a considerable growth percentage might not be possible (Raudys, 2018).

Researchers of the value-added model remained cautious about its effectiveness if it was the only model used. VAMs imply that test scores alone are valid indicators of teacher performance. Opponents argued that VAMs did not promote teacher growth. As a result, value-added measures have proven unstable across statistical years, models, and classes (Raudys, 2018).

As noted by Baker and colleagues (2010), teacher effectiveness only predicted 4% to 16% of the variation in a teacher's ratings from one year to the next. Among teachers ranked in the top 20% of teaching effectiveness, fewer than a third remained in the highest performance level during the following year. Additionally, by the third year, many of the teachers ranked in the top 20% initially had moved to the bottom 40% of teaching effectiveness. Further, they contested that because measures' validity might be inconsistent, it would be unfair to base high-stakes personnel decisions on a single data point (Baker et al., 2010).

Milanowski (2011) added that VAMs alone did not provide enough information to improve teacher effectiveness. He added that a combination of systems is needed to assess teaching practice. A successful teaching evaluative system includes a combination of observations, teaching artifacts, and frequent walkthroughs that provide evidence of daily classroom performance (Milanowski, 2011). Weiner and Jacobs (2011) agree that VAMs have valid attributes but should not be used solely to evaluate teaching effectiveness. VAMs could possibly identify the most and least effective teachers based on student test scores. They may also contribute to stronger analyses of programs, influences, and the validity of evaluation methods. Still, VAMs do not help determine why individual teachers excelled or struggled or provide direction for improvement. Weiner and Jacobs (2011) further dispelled the validity of the sole use of VAMs in teacher evaluation by highlighting the inconsistencies of VAMs when analyzed over time.

The American Statistical Association (ASA) took a stance against using VAMs because they typically measure correlation, not causation (Strauss, 2014). The ASA also stated most VAM studies find that teachers account for between 1% to 14% of the variability in test scores, and ranking teachers by VAM scores may have unintended consequences (American Statistical Association, 2014). Overall, researchers cautioned against value-added measures as being accurate. They argued VAMs were not a valid assessment of teacher performance. (Baker et al., 2010; Darling-Hammond, 2009; Darling-Hammond, 2013; Fuhrman, 2010).

Another frequently used evaluative model is based on teacher observation. During observations, administrators observe classroom instruction. The observations might be announced or unannounced and typically last between ten and thirty minutes (Raudys, 2018). Well-designed observation instruments include rubrics that promote consistent, reliable results and multiple raters (Milanowski, 2011; Raudys, 2018; Robinson, n.d.). Using this evaluative model, the administration becomes more familiar with the classroom, teachers, and the school environment. Videos may also be used as an instrument to aid feedback given to teachers (Raudys, 2018).

However, similar to VAMs, evaluations based only on observations had certain disadvantages (Raudys, 2018.) For example, the administrator may see only a small portion of the instruction. Additionally, the observations are very time-consuming. Coupled with the various demands placed on administration, observations may be completed in haste without deep and meaningful feedback (Raudys, 2018).

During the Measure of Effective Teaching (MET) study, organized by the Bill & Melinda Gates Foundation, teachers were asked to build a video library of teaching practices. The teachers were given video equipment and asked to capture videos of teaching practices 25 times. Of the 337 teachers that initially participated, 67 teachers permitted their lessons to be evaluated by 129 raters, comprised of 53 administrators and 76 peers. To increase reliability and gain insight on the benefits of longer versus shorter observations, raters were asked to score half the lessons after the initial 15 minutes and then again after the entire lesson. The remaining half of the lessons were only rated at the end (Ho & Kane, 2013). Ho and Kane (2013) noted the following from the evaluated lessons: Observers rarely used the lowest and highest categories on the rubrics. On any given rated behavior, an average of 5% of scores were at the unsatisfactory level, while only 2% of ratings fell in the advanced categories. The majority of scores fell into the middle two categories, basic and proficient. Additional concerns regarding observations included the likelihood of administrators assigning higher scores to the teachers with whom they were more familiar (Ho & Kane, 2013). Ho and Kane (2013) also found they were able to ensure the reliability of .65 or above through the use of multiple observers with a mixture of full and shorter, 15-minute observations.

Danielson’s Framework Model (1996; 2014) is an evaluation model based on four domains: planning and preparation, classroom environment, instruction, and professional responsibilities. This model encourages more focused collaborative feedback discussions, increased instructional reflection, and evidence-based feedback (Raudys, 2018). Danielson’s model was modified to its current form in 2013, following the Common Core Standards' release. Like other models, the rubric had four levels with examples and attributes. The domains were broken down further into characteristics that support each domain's domains (Danielson, 2014).

The Marzano Focused Teacher Evaluation Model (Carbaugh et al., 2017) is also popular among evaluative models. The model is a systematic step-by-step approach for observations. Observers were expected to score 23 elements from four domains over the year, using a four-point rubric. The domains were as follows: rigorous standards-based system, focus on student results, instructional frameworks with pathways to scaffold instruction, and access to tools and resources with a continuum for professional growth. The questions below were provided to direct the rater: What elements am I seeing when observing a teacher? Does the teacher use the strategy correctly? What technique or techniques does the Teacher use to monitor the desired effect/outcome? What percent of students demonstrate achievement of the desired effect at the appropriate level of the target? After monitoring student evidence and determining the number of students who demonstrate the desired effect, does the teacher make an adaption? The model relies on multiple observers for interrelated reliability (The focused teacher evaluation model, 2020).

Researchers determined that all observed elements were significantly correlated to student gains by analyzing over 250,000 observations on approximately 13,000 teachers based on the Marzano Focused Teacher Methods. The observation scores were a significant predictor of value-added models, and the teacher observation score was the most significant predictor of student growth on state assessments (The focused teacher evaluation model, 2020).

Although there are worthy components in all models, one model alone might fail to provide the information needed for teacher growth. However, a combination might provide the information necessary for optimal professional development (Robinson, n.d.). Milanowski (2011) suggested that student measures and teacher behaviors were two very different behaviors and should be analyzed separately to determine teacher effectiveness. Baker et al. (2010) noted that sound evaluations required balancing the many attributes of effective teaching. A report from the National Council for Teacher Equality (Putman, Ross, and Walsh, 2018) identified these attributes as the following: student surveys, objective measures of student growth, multiple rating categories, annual evaluations and observations, professional development tied to evaluations, and written feedback following observations.

Purpose of Evaluations

As educators attempt to provide a quality education through which students become productive members of society, the question remains, how do educational leaders ensure the quality of education while encouraging the needed improvement in teaching effectiveness (Weisburg et al., 2009)? Through the implementation of and feedback from the evaluative systems, educators have the potential to identify possible areas of personal growth. Identifying essential standards relative to teacher growth would allow educators to participate in professional development, focusing on the areas most closely related to teacher efficacy and student growth (Anderson et al., 2016).

In Georgia, the teacher evaluative measure was referred to as TKES, or the Teacher Keys for Effectiveness System ("Georgia's Teacher Keys Effectiveness System," n.d). TKES echoes the beliefs that observation tools must provide quality feedback for teacher growth. The flexible and Full Plan options (“Georgia’s Teacher Keys Effectiveness System,” n.d.) are driven by individual, specific teacher goals and opportunities for growth.

Approaches to teacher assessment considered moderately valid include several components (Robinson, n.d.). Peers, principals, or outside evaluators conduct observations and provide a variety of viewpoints on teaching effectiveness. The tool must include a means to assess gains in student achievement, and student ratings may be included and may also affect the instrument's validity. There must also be allowances for principal judgment, teacher self-report, and analysis of classroom artifacts and teacher portfolios (Robinson, n.d.).

Many leaders agreed that more resources were needed for quality feedback, regardless of the purpose (Skedsmo & Huber, 2018). Kane, Taylor, Tyler, and Wooten (2011) argued that scoring individual practices allowed for understanding variations in skills among teachers. For example, math achievement scores grew more when teachers scored high in classroom management. Reading skills improved dramatically when teachers scored high in questioning and discussion. Proponents of standardized observations believed classroom practice evaluations were valuable, even if the results were not significant predictors of student achievement. Effective classroom procedures might be more easily identified and replicated (Kane et al., 2011).

Marzano and colleagues (2011) noted that evaluative systems focusing on development have three primary characteristics. The system must be comprehensive and specific. It must include a developmental scale and acknowledge and reward growth. The system must revolve around systematic strategies, content strategies, and strategies enacted at the moment of instruction or unplanned strategies (Marzano et al., 2011).

Effective Teaching Practices

Research-based teaching practices must be incorporated into evaluative systems for teacher growth to occur. If teacher evaluation's pervasive purpose is to increase student achievement, the most impactful strategies must be the assessment's driving force. Evaluation systems cannot overlook proven strategies, and leaders must consider the research when developing evaluative systems or evaluating teachers if teacher growth was the ultimate point of teacher observation (Marzano et al., 2011).

While there are many theories and ideas regarding quality instruction, John Hattie (2008) combined thousands of theories and studies to develop a robust list of effective teaching practices. Based on their meta-analysis of over 800 studies, he and his colleagues determined attributes that typically impact instruction. Hattie (2008) did not discover or discuss new teaching practices, but his premise was that teachers should know their impact on students and teach accordingly. Hattie (2008) explained that learning would be visible with consistent, effective classroom practices. When he later updated his list of strategies in *250 plus influences on student achievement (2017),* Hattie maintained his previous findings since they proved to be highly effective practices (Hattie, Fisher, & Frey, 2017). Only slight changes occurred in the effect sizes for the methods between the 2008 and 2017 findings (Hattie et al., 2017).

From these influences, Hattie and his colleagues (2017) assigned a ranking to show the effect of each of the forces on students. They suggested an effect size of .4 equals one year's expected student growth. Therefore, any effect size greater than .4 was considered beneficial to students. Consequently, some influences were ranked with negative effect sizes, meaning students were negatively impacted or regressed. Examples of negative influences included lack of sleep, low socioeconomic status, and physical and cognitive delays (Hattie et al., 2017). Most negative influences fall outside of the teacher's control, except for influences the teacher might potentially have mitigated, such as motivation, students feeling disliked, retention, boredom, and performance goals (Hattie, 2008).

Over 100 attributes researched by Hattie could result in over one year's growth if consistently implemented. The highest of these characteristics, resulting in an effect size of over 1.0, were typically related to classroom instruction. Examples include collective teacher efficacy, self-reported grades, teacher estimates of achievement, cognitive tasks analysis, response to intervention, Piagetian programs, and the jigsaw method (Hattie et al., 2017). Other influences that have historically ranked high include the following effective practices: teacher credibility, providing formative evaluation, micro-teaching, classroom discussions, comprehensive intervention for learning disabled students, teacher clarity, and feedback. Collective teacher efficacy is the teacher's belief that they can impact students (Hattie, 2008). Similarly, Hattie (2008) determined students with teachers who believed they could have a high impact on learning made more gains than students with teachers who lacked the belief in themselves to impact instruction positively. Also, self-reported grades involved understanding the student's expectations and pushing the learning to exceed the expectations (Hattie, 2008).

Several more strategies with highly effective ratings involved differentiation and assessment (Hattie et al., 2017). Response to intervention, ranked as the third most effective practice, is also a crucial component of the differentiated classroom. Teachers precisely understand what the student needs, provide consistent interventions to meet the learner's needs, and frequently measure progress. In terms of effectiveness, Hattie’s number eight strategy, comprehensive interventions for learning disabled students, had similar results but requires more specialized learning programs. Providing formative evaluation is another strategy for receiving high effect sizes in Hattie's research. It requires the teacher to assess before and during the learning process to guide instruction, which leads to appropriate differentiation. Similarly, another of Hattie's top strategies is using the assessment data. The use of assessment data allows teachers to provide robust, targeted feedback so students can isolate the skills or information on which they need to concentrate (Hattie et al., 2017).

Hattie and his colleagues (2017) also found teachers' attitudes toward themselves and their students significantly impact student growth. The fourth most influential strategy, teacher credibility, is vital to a productive classroom environment. The four factors of credibility include trust, competence, dynamism, and immediacy. They noted that students do not engage with teachers they felt were not credible, creating an adverse climate and potentially poor classroom management. The relationships were ranked with a .72 impact, or almost twice the expected impact level. Students taught by teachers who maintained positive and respectful relationships with students made higher gains than their peers who did not have positive relationships with their teachers (Hattie et al., 2017).

Because Hattie's top influences are primarily related to instruction and teacher behavior, incorporating his strategies is instrumental in improving education (Hattie et al., 2017). Through this study, educators have a comprehensive list of effective teaching practices (Marzano et al., 2011). Although other theorists may disagree with the science behind Hattie's (2008) effect sizes, most typically agree with the researcher regarding the overall positive or negative effect the strategies had on teaching and learning (Marzano et al., 2011). Further, research completed before and after Hattie's initial study generally concurred with his findings (Hattie, 2008). Due to the positive impact on classroom instruction, including these components in an evaluative system (Marzano et al., 2011).

Table 1

Alignment of TAPS and Hattie's Effective Teaching Practices

	Hattie's Strategies
Standard 1: Professional Knowledge	Collective teacher efficacy, Teaching strategies, Teacher subject matter Knowledge, Teacher education
Standard 2: Instructional Planning	Prior achievement, Metacognitive strategies
Standard 3: Instructional Strategies	Jigsaw method, Self-reported grades, Microteaching, Classroom discussions, Feedback, Teacher clarity, Direct instruction, Teaching strategies
Standard 4: Differentiated Instruction	Response to intervention, Comprehensive interventions
Standard 5: Assessment Strategies	Formative evaluation, Evaluation
Standard 6: Assessment Uses	Targeted feedback, reflection
Standard 7: Positive Learning Environment	Teacher-student relationships, Classroom behavior, Peer tutoring, Classroom management, Classroom cohesion
Standard 8: Academically Challenging Environment	Teacher expectations, Goals, Student expectations
Standard 9: Professionalism	Teacher credibility, teacher not labeling students, Professional development
Standard 10: Communication	Questioning, Teacher verbal ability

Note. TKES Standards and related Hattie Strategies.

Cautions for Educators Seeking Effective Evaluation Systems

Educational theorists have weighed in on the problems and precautions regarding evaluations. Marzano and colleagues (2011) noted the evaluation tools with different purposes look very different. Further, because approximately 76 percent of administrators want to focus on teacher growth, teacher development is critical and should be the priority of evaluation systems. They also advised of the inconsistencies between the information provided by current evaluative models and noted the types of feedback teachers need to improve their effectiveness (Marzano et al., 2011).

Anderson et al. (2016) warned that educators must identify the purpose of teacher evaluations before making substantial changes to evaluative systems. They stated that effective teacher evaluation systems must catalyze teacher improvement and student growth. If utilized properly, teacher evaluations may lead to substantial academic growth for students. States must attempt to capture the benefit of effective assessment by including successful instruction elements and embedding the core concepts associated with student growth into the teacher evaluation systems (Anderson et al., 2016). Anderson and colleagues (2016) also stated that teachers deserve feedback that increases teacher efficacy and student achievement. The sole purpose of teacher evaluation must center around teacher advancement and professional growth (Anderson et al., 2016).

The question remains why there is a lack of pervasive, efficient, evaluative systems for teachers. Opponents of the overarching educational evaluation reform argue there is very little difference between teachers, and the use of the vast resources necessary to develop such a system would be a waste of resources (Berliner, 2018). The American Statistical Association (ASA) argued that most student achievement variance might be attributed to teacher characteristics and instruction (Berliner, 2018). Haertel (2013) concluded, based on his review of teacher evaluation studies (Goldhader, Brewer & Anderson, 1999; Nye, Konstantopoulos, & Hedges, 2004), there is only an approximate 10 percent variance in student growth and achievement between teachers. Opponents argue the cost of an evaluation is too high for the low return. The rate of poor teachers is meager, and the system to identify the few failing teachers is too costly, insensitive, and potentially insulting to those who are masters in the craft of teaching (Berliner, 2018).

Robinson (n.d.) noted challenges in developing effective evaluative systems. She warns that evaluators may witness as little as 0.1% of instruction before assigning scores. Teachers may construct a lesson far above the quality of typical education if they know the observation time.

The author also noted a lack of trust, time, and equity as common pitfalls of evaluative systems (Robinson, n.d.).

Those opposing widespread teacher evaluation systems also argue that the educational systems do not have the funds to support poor teachers' development and should not identify them (Berliner, 2018). Further, Berliner reported industries outside of the educational field typically pay for performance and train for upcoming changes. However, the educational system does not. Because educators are not trained on the inadequacies, it is not beneficial to identify them. An example would be the Common Core. There were few professional development opportunities and a lack of funds to prepare teachers for the vast changes required by the Common Core initiative (Berliner, 2018).

Opponents also agree that there are serious flaws following unsatisfactory observations. For example, observations are rarely used as the primary reason for rewarding, punishing, or firing. There is often no follow-up or negative repercussions to insufficient observations. Further, using the standardized test results alone for teacher evaluation creates a validity problem, while the use of observations may create reliability concerns (Berliner, 2018).

Berliner (2018) suggested another essential point: the severe implications of teacher observations and their corresponding actions. Opponents argue that the correlation between student achievement and teacher effectiveness may be relatively low and difficult to capture in teacher evaluations. Given that lives and careers may be threatened by poor observations, which may or may not be valid, some educators warn that educational leaders must proceed with caution when determining the means of future teacher evaluations (Berliner, 2018).

Another reason educators have difficulty determining the most effective means of teacher evaluation is to first settle on the evaluative tool's fundamental purpose (Berliner, 2018). Leaders must determine the primary purpose of evaluations before developing a mechanism to evaluate teachers. The overall purpose must be resolved before sweeping education evaluation reform can occur (Berliner, 2018).

The New Teacher Project was founded based on the belief that educators must have the avenue to develop teachers (Skedsmo & Huber, 2018). The organization has a mission of ensuring poor and minority students have equal access to effective teachers while maintaining the following beliefs regarding teacher evaluations: all the teachers are good to great, there are no poor teachers, excellence goes unrecognized, while professional development is inadequate, no special attention is given to novices, and poor performance goes unaddressed (Skedsmo & Huber, 2018). The New Teacher Project and the Race to the Top Initiative concur with Marzano and his colleagues' (2011) opinion that evaluative tools must focus on growth. These value-added models attempt to measure teachers' effect on school attributes, including achievement scores and levels of learning (Skedsmo & Huber, 2018). While reviewing the results of 35 educational research studies using an experimental or quasi-experimental control group or statistical modeling with hypothesis testing, Darling-Hammond, Hyler, and Gardner (2017) maintained sustained professional growth occurs when the focus of the evaluation is on student growth and learning, when the feedback is clear, with specific goals, when teachers are encouraged to grow continually, and when a mentor in a trusted environment provides input. The leaders promote and support continued learning (Darling-Hammond et al., 2017).

The concern with teacher evaluation is the one-size-fits-all approach that fails to recognize the varying needs of schools and students (Primer: Education Issues – Variables Affecting Student Achievement, 2014). Since schools have variables that affect their population, designing an evaluative tool that considers these variations is essential. Further, the evaluative tool must allow for these differences. School variables that may affect achievement include exterior influences, i.e., poverty or Title One status, the number of English Language Learners, and the overall makeup of the community that the school serves. Internal variables include school setting, whether students are served in one classroom or by content-specific teachers, tracking, Title status, and class size (Primer: Education Issues – Variables Affecting Student Achievement, 2014).

Factors Affecting the Learning Environment

Many variables affect the learning environment. Teaching and learning differ greatly based on those variables. The failure to consider those factors in teacher effectiveness could be irresponsible. Although some factors hinge on school-based decisions and settings, many are beyond the school systems' control. Regardless of whether the factors are internal or external, they can affect teaching and learning and must be considered when evaluating teachers (Najimi, Sharifirad, Amini, & Meftagh, 2013).

There are many internal factors affecting teaching and learning. Examples include curricular decisions, program decisions, physical building environment, building leadership, staff collaboration, and staff development (Najimi et al., 2013). Building administration may not have the ability to control these characteristics. However, some of the factors are within the scope of school administrators, such as classroom setting.

Administrators make decisions regarding the classroom setting in elementary schools. Many schools remain self-contained throughout the primary years. According to Montero (2020), proponents of the self-contained models contend that teachers build stronger relationships with students. Further, instructional time is maximized because students do not transition to other classes (Montero, 2020).   However, some leaders may choose to design a departmentalized model. In a departmentalized model, the intention is for teachers to become content specialists. A teacher in a departmentalized grade level may teach their subject several times each day to multiple classes. This model may lead to more peer collaboration (Montero, 2020).

Like most educational factors, much research surrounds best practices and building delivery models. In a 2015 study involving seven schools and 19 classrooms, Woods (2017) examined the impact of departmentalization and self-contained instruction on fourth grade and fifth grade reading and math achievement. The study's questions included: 1) Is departmentalization in fourth grade or fifth grade more academically impactful than self-contained classes?; 2) Does departmentalization result in increased teacher efficacy or morale as opposed to self-contained teachers?; 3) Do teachers prefer one model to another? 4) Are the reasons data-related?; and 5) Have schools been successful in using the strategy of their choice to meet federal AMO goals? Through a t-test, end-of-grade test scores were analyzed to compare the means of students being served in self-contained and departmentalized settings. Likert scale surveys were used in the chi-square analysis for questions 2 and 3. For question 4, Woods (2017) reviewed the history of goal acquisition. The researcher concluded a significant difference in fourth grade ELA or Math achievement on the end-of-grade assessments when students were served in a departmentalized model (Woods, 2017).

In contrast, there was no significant difference in achievement in fifth grade based on delivery models. However, 86% of all teachers felt they were more knowledgeable in the content in the departmentalized model. Teachers also thought they had more time to prepare for lessons, as 91% reported a more manageable workload when teaching in departmentalized settings (Woods, 2017). Woods (2017) suggested future research examining the role of other factors, i.e., teacher effectiveness, professional development, and school culture and climate, on student achievement.

In a similar study in 2014 in Virginia, Nelson (2014) documented a significant difference in the mean scores on a state test in Virginia for students based on the classroom setting through a causal-comparative design. The research questions that lead the study included: 1) Was there a statically significant difference between departmentalized fifth grade students' mathematics achievement based upon classroom organizational structure, gender, racial minority/racial non-minority status, as measured by students' scores earned on the Virginia 2011 Grade 5 Mathematics SOL Test, while controlling for students' Virginia 2010 Grade 4 Mathematics SOL Test scores?; 2) Was there a statistically significant difference between departmentalized and non-departmentalized fifth grade students' mathematics achievement based upon classroom organizational structure and gender, as measured by students' scores earned on the Virginia 2011 Grade 5 Mathematics SOL Test, while controlling for students' Virginia 2010 Grade 5 Mathematics SOL Test scores?; 3) Was there a statistically significant difference between departmentalized and non-departmentalized fifth grade students' mathematics achievement based upon classroom organizational structure and racial minority/racial non-minority status, as measured by students' scores earned on the Virginia 2011 Grade 5 Mathematics SOL Test, while controlling for students' Virginia 2010 Grade 4 Mathematics SOL Test scores?; 4) Was there a statistically significant difference between departmentalized and non-departmentalized fifth grade students' mathematics achievement based upon gender and racial minority/racial non-minority status, as measured by students' scores earned on the Virginia 2011 Grade 5 Mathematics SOL Test, while controlling for students' Virginia 2010 Grade 4 Mathematics SOL Test scores?; 5) Was there a statistically significant difference between departmentalized and non-departmentalized fifth grade students' mathematics achievement based upon classroom organizational structure, as measured by students' scores earned on the Virginia 2011 Grade 5 Mathematics SOL Test, while controlling for students' Virginia 2010 Grade 4 Mathematics SOL Test scores?; 6) Was there a statistically significant difference between departmentalized and non-departmentalized fifth grade students' mathematics achievement based upon gender, as measured by students' scores earned on the Virginia 2011 Grade 5 Mathematics SOL Test, while controlling for students' Virginia 2010 Grade 4 Mathematics SOL Test scores?; and 7) Was there a statistically significant difference between departmentalized and non-departmentalized fifth grade students' mathematics achievement based upon racial minority/racial non-minority status, as measured by students' scores earned on the Virginia 2011 Grade 5 Mathematics SOL Test, while controlling for students' Virginia 2010 Grade 4 Mathematics SOL Test scores? The study's participants included 239 students, with 90 served in departmentalized settings and 149 served in non-departmentalized or self-contained settings. The researcher concluded that students in departmentalized homerooms scored approximately thirty points higher than students in a self-contained setting. There was no difference among subgroups (Nelson, 2014). Nelson recommended further research in other content areas and school districts and settings.

Dymond (2017), alternatively, had differing results. He also studied the impact of setting on student achievement. His study analyzed data from 696 fourth grade students taking the Measures of Academic Progress (MAP) test in South Carolina. Qualitative data were also collected from volunteer participants. The research questions included the following: 1) Are there significant differences in fourth grade achievement scores (MAP) in math between students in self-contained and departmentalized classrooms?; 2) What components of organizational structures do teachers relate as having an impact on the quality of the math program?; and 3)

What are the identifiable culture classroom differences in environments in comparing self-contained and departmentalized settings? Dymond (2017) found that self-contained students outperform their departmentalized counterparts by 9.8%. In the qualitative portion of the study, teachers attributed the higher score to the social and emotional connection made by teachers who served their students for the entire school day. Dymond recommended further research specific to performance in each of the four Northwest Evaluation Association (NWEA) goal areas: Algebraic Thinking & Operations, Number Sense & Operations, Measurement & Data Analysis, and Geometry. Other future research implications included data analysis at different grade levels and including parent responses and more teachers within the study (Dymond, 2017).

In a 2014 correlational design study, Jack (2014) posed the following research questions: 1) Is there a significant difference in fifth grade mathematics CRCT scores between urban students who receive mathematics instruction in a departmentalized classroom setting and urban students who receive instruction in a self-contained classroom setting?; 2) Is there a significant difference in fifth grade mathematics CRCT scores between urban students who receive mathematics instruction in a departmentalized classroom setting and urban students who receive instruction in a self-contained classroom setting when controlling for percentage of students eligible for the free and reduced lunch program (FRL), race/ethnicity, and school size?; and 3) Is there a statistically significant interaction between percentage of students eligible for the free and reduced lunch program and organizational structure in terms of fifth grade CRCT mathematics performance? The study was based on 46 schools in two districts in Georgia on the 2010-11 Criterion-Referenced Competency Test (CRCT). Jack (2014) found no significant difference among students based on an academic setting. Other factors affected growth and performance.

For example, the percentage of students being served free and reduced lunches was a significant predictor of student achievement on the CRCT (Jack, 2014).

Mitchell's (2013) non-experimental correlational study of 3,456 students in Southern California also concluded that classroom setting had no significant impact on ELA or Math achievement as measured by the California Standards Test (CST). The researcher acknowledged other factors might have skewed the results. For example, the teachers from the study had no intentional content-specific training. Therefore, the teachers were not considered content specialists (Mitchell, 2013). The questions that guided the study included 1) Controlling for gender, language, and prior achievement, is there a statistically significant difference in Grade 6 English Language Arts scale scores on the 2010-2012 California Standards Test between students who received instruction in departmentalized classrooms and students who received instruction in self-contained classrooms?; 2) Controlling for gender, language, and prior achievement, is there a statistically significant difference in Grade 6 Mathematics scale scores on the 2010-2012 California Standards Test between students who received instruction in departmentalized classrooms and students who received instruction in self-contained classrooms?; and 3) How do teachers in departmentalized and self-contained settings describe their experience within each setting? Mitchell (2013) recommended further research, including a qualitative analysis of teachers' perceptions regarding classroom setting and student achievement and broadening the study's scope to include other content areas.

Although educators can make decisions that affect the delivery model, many other factors affecting student achievement are outside the school building. According to "The Effect of Poverty on Student Achievement" (2009), one of the most damaging influences on education is poverty. The family income level is one of the most powerful indicators of student achievement because of the vast array of disadvantages that accompany children of poverty. Children living in homes with a low socioeconomic level are more likely to have prenatal disadvantages. In these homes, there is an increased risk of injury or illness without adequate medical care. Other factors associated with poverty and the negative impacts on education include nutritional problems, exposure to pollutants, unsafe neighborhoods, lack of adult supervision and attention, residential instability, family violence, and lack of access to educational activities and materials (The Effect of Poverty on Student Achievement, 2009).

Multiple studies found a correlation between socioeconomic status and school achievement. Children of poverty, for example, are at significant risk for academic failure. The students enter school with lower reading skills and have more significant risks for adverse reading outcomes. The students frequently have attendance problems and drop out more often, and their attention and concentration may be impaired due to the myriad of problematic environmental factors associated with poverty (Bellani & Bia, 2018; Boatwright & Metcalf, 2020; The effect of poverty on student achievement, 2009; Jensen, 2010; Vallaster, 2019). Furthermore, Bellani and Bia (2018) stated the effects of poverty are long-term. Children from E.D. homes are 23% more likely to stop attending school after primary grades. Between 90 and 95% of low-income students have parents who did not complete high school (Bellani & Bia, 2018). Vallaster (2019) indicated that educators of children of poverty must focus on meeting the students' immediate physical and social-emotional needs to succeed and close the gap. To do this, teachers must build relationships. Based on these findings, there is an increased need for collaborative support to counteract poverty's negative impacts (Vallaster, 2019).

Teacher Keys Effectiveness System

To standardize and strengthen the teacher effectiveness system in Georgia and bridge the gap between effective teaching practices and student achievement, lawmakers passed House Bill 244 in 2013. The law established the Teacher Keys Evaluation System (TKES). All state educational systems began evaluating teachers using the TKES criteria during the school year 2014-2015 (H.B. 244, 2013-2014). The new evaluative system was required to be implemented by the 2014 school year. Under the system, teachers, assistant principals, and principals receive an overall effectiveness rating annually. The evaluation system is based on multiple measures prioritizing student growth and achievement (H.B. 244, 2013-2014).

There are three primary components of TKES, the Teacher Assessment on Performance Standards (TAPS), Student Growth, and Professional Growth. A Teacher Effectiveness Measure (TEM) is calculated using scores in each component. TAPS provides evaluating administrators with rubrics on which to base teacher ratings. Although all standards may not be evaluated on walkthroughs, formative assessments require ratings on all ten standards. While there is administrative discretion when assigning the summative scores, scores on the observations generally inform the summative assessment ratings. The TAPS portion of the TKES evaluation is 50% of the overall teacher effectiveness rating.

All TKES evaluated educators have three required conferences, including a pre-evaluation conference, mid-year conference, and summative evaluation conference. Following the pre-evaluation conference, administrators may begin walkthrough and formative observations for the evaluation's TAPS portion. During the ten-minute walkthrough observations, teachers are rated on select standards. During the formative, or thirty-minute observations, teachers are assessed on all Performances Standards ("Georgia's Teacher Keys Effectiveness System," n.d).

There are two plans under the TKES platform. The full plan, which is used for induction teachers, teachers with less than three years of experience, or teachers in need of remediation, requires six observations yearly. A teacher on the Full Plan will have four walkthrough observations and two formative observations yearly. Teachers who have completed the initial phase of teacher induction and have satisfactory ratings have a reduced number of observations consisting of two walkthrough observations and one formative observation under the flex plan. Teachers who do not maintain satisfactory ratings may be returned to the Full Plan at the discretion of administration ("Georgia's Teacher Keys Effectiveness System," n.d).

All teachers are rated on four levels for each standard. The levels are exemplary (level IV), proficient (level III), needs development (level II), or ineffective (level I). The standards by which teachers are evaluated for observations include Standard 1: Professional Knowledge, Standard 2: Instructional Planning, Standard 3: Instructional Strategies, Standard 4: Differentiated Instruction, Standard 5: Assessment Strategies, Standard 6: Assessment Uses, Standard 7: Positive Learning Environment, Standard 8: Academically Challenging Environment, Standard 9: Professionalism, and Standard 10: Communication ("Georgia's Teacher Keys Effectiveness System," n.d). Each standard has the performance standard, performance indicators, and an appraisal rubric to help administrators assign scores from the rubrics to each standard. The purpose of these aids is to allow for more consistent scoring across administrators and systems. Additionally, evidence of implementation, or easily identified characteristics of effective teaching practices, are included in the documentation to ease scoring (Georgia's Teacher Keys Effectiveness System Meaningful Feedback Professional Growth Flexibility to Innovate, 2018).

Standard 1, Professional Knowledge, centers around the teacher's understanding of the subject matter or concepts. Other key factors include pedagogical thinking, decision making, learner knowledge, and cultural or community knowledge. Teachers with sufficient professional knowledge ask open-ended questions, facilitate interdisciplinary approaches and understanding, and use various sources to encourage meaningful connections. When evaluating professional knowledge, administrators may look for higher-level questioning techniques, alternative explanations, inquiry-based learning, student-directed activities, and engaging scenarios (Georgia's Teacher Keys Effectiveness System Meaningful feedback professional growth flexibility to innovate, 2018). Teachers scoring in the proficient or exemplary level in Standard 3 possess confidence in delivering content. They understand the essential knowledge and adapt teaching methods to the learner's needs. Professional teachers are culturally competent and strive to understand the communities and cultures they teach ("Georgia's Teacher Keys Effectiveness System," n.d).

Mastering Standard 2, Instructional Planning, requires effective planning through standards, strategies, resources, and data to meet learners' diverse needs. Teachers, who plan effectively, ask the following about curriculum when planning: What should be taught? How should it be taught? How should instruction and student learning be assessed? Effective and exemplary teachers also understand the essential aspects of effective teaching and how to sustain student engagement while making education relevant. Teachers with solid planning skills encourage students to explore, inquire, and construct knowledge while discovering fundamental knowledge ("Georgia's Teacher Keys Effectiveness System," n.d). For proficient evidence of planning, at a minimum, teachers must understand that no students are the same. Students arrive at school with very different backgrounds, interests, and abilities. The failure to understand students' varying needs and plan accordingly is unprofessional and unethical. The content delivery determines the teacher’s effectiveness ("Georgia's Teacher Keys Effectiveness System," n.d).

Instructional Strategies, Standard 3, requires the following critical elements in instruction: differentiation, variety, cognitive instruction, student engagement, recognition of student learning, and necessary adjustment, questioning, and relevance. Effective teachers promote learning through instructional strategies engaging students in active participation. They also understand that the essential aspect of successful instruction is to sustain student engagement through relevant instruction. Students must be able to explore, inquire, and construct knowledge. Students are highly motivated when learning is authentic. Effective teachers create classroom environments where authentic, rigorous, and engaging instruction occurs daily ("Georgia's Teacher Keys Effectiveness System," n.d).

Standard 4 is Differentiated Instruction. Teachers with sufficient or exemplary ratings on Standard 4 provide appropriate, differentiated content that addresses each student's specific learning needs. The teachers understand that students learn in varying ways and at different speeds. In the differentiated instruction standard, levels III and IV teachers understand diverse learners' knowledge and teach accordingly. Teachers with effective differentiation scores understand precisely what the students should know and be able to do. The teachers have a toolbox of instructional approaches full of student-centered, engaging, and relevant activities ("Georgia's Teacher Keys Effectiveness System," n.d). To differentiate effectively, teachers must have the necessary assessment methods to identify specific learning needs correctly.

Performance Standard 5, Assessment Strategies, outlines a systematic use of valid and appropriate diagnostic, formative, and summative assessment methods. Each type of assessment serves a different purpose in assessing students. Diagnostic assessments determine what the students know before instruction and guide instructional decisions at the unit's beginning. Formative assessment helps teachers adjust instruction throughout the unit as student needs require. Summative assessments typically occur at the end of the period, unit, or chapter and determine the student's attainment of standards. Teachers are expected to practice effective assessment strategies with all types of assessments to deliver instruction to meet the learners' exact needs ("Georgia's Teacher Keys Effectiveness System," n.d).

Performance Standard 6, Assessment Uses, is a natural progression from the prior standard. Following a relevant assessment, the accomplished educator analyzes data from the evaluation to make ongoing educational decisions. Valid assessments allow the teachers to provide timely and appropriate feedback and reinforcement, align instruction to standards, maximize instructional time, and deliver more effective pedagogical methods. Additionally, students may use the assessment to set personal goals. According to the Georgia Department of Education, the evaluation should be a means to an end for instruction and is the ultimate goal of teaching and learning ("Georgia's Teacher Keys Effectiveness System," n.d).

Performance Standard 7 is Positive Learning Environment. This standard recognizes the need for a safe and orderly environment conducive to learning. A positive learning environment encourages respect for all, and students need an engaging environment to maximize learning. Further, appropriate environments allow teachers to monitor student behaviors, keep students on task, and provide engaging scenarios. Educators recognize the value of a positive learning environment and understand it is vital to successful learning outcomes. Elements of a positive classroom environment include discipline, routines, organization, engagement, maximized instructional time, high expectations, and respect for all individuals within the classroom ("Georgia's Teacher Keys Effectiveness System," n.d).

Academically Challenging Environment, Performance Standard 8, guides teachers to create a student-centered instructional environment with high learning levels. Students are self-directed in an academically challenging classroom. The following are attributes of high-quality learning environments: active engagement, authenticity and relevance, collaboration and community, learner autonomy, cognitive complexity, generativity, multiple perspectives, pluralism, reflectivity and metacognitive awareness, self-regulation, ownership, transformation, and productivity. Teachers establishing a clear focus with a well-organized lesson while maintaining on-task behaviors and an appropriate instructional pace meet the requirements for an academically challenging classroom ("Georgia's Teacher Keys Effectiveness System," n.d).

Standard 9 is Professionalism. According to the Georgia Department of Education, there are three essential elements of professionalism: professional standards and ethics, continuous self-professional development, and contributions to the profession. The first element, professional standards and ethics, requires the teacher to adhere to legal and ethical guidelines and professional standards. The teacher must present a professional demeanor and positive interaction and respect the diversity of learners. Regarding continuous professional development, the teacher must remain reflective while continuously acquiring new knowledge through professional renewal as a life-long learner. While contributing to the profession, educators must serve as role models, participate in professional associations, and contribute to the development of other teaching professionals ("Georgia's Teacher Keys Effectiveness System," n.d).

The final standard on which teachers in Georgia are assessed is Communication or TKES Standard 10. Communication is at the core of everyday, effective teaching. Teachers rated as exemplary communicators hold the following inherent understandings regarding communicating, understand the ebb and flow of a classroom, use a wide variety of communication techniques, create relationships with students, and communicate effectively with colleagues, families, and stakeholders ("Georgia's Teacher Keys Effectiveness System," n.d).

Another critical component to maintain evaluation fidelity is understanding the behaviors associated with each standard. Although detailed rubrics are available, key terms are included within each standard. Standard 1 requires teachers to understand the content and build their educational toolbox while embracing a growth mindset. Standards 2 and 3, both instructional-based, call for the teachers to use effective, research-based strategies to deliver lessons based on state standards. Standard 4 challenges the teacher to provide differentiated instruction while supplying the student's appropriate level of support (Leader Keys Effectiveness System: Fact sheets, 2012). Standards 5 and 6 center around assessment and mandate the teacher to implement various authentic assessment strategies while using the data to facilitate leveled instruction. Standards 7 and 8 attempt to build a productive environment by establishing a classroom with clearly stated expectations and maximized instructional time. Standard 9, professionalism, revolves around expected teacher behaviors, i.e., confidentiality, flexibility, and continued professional growth as an educator. The final standard, communication, provides the expectation of clear and appropriate communication of all types to colleagues, parents or guardians, and students (Leader Keys Effectiveness System, 2018).

Beyond the classroom, effective communication is equally important. Because no one person is singularly responsible for educating a student, communicating with colleagues and community members is crucial for optimal learning and student growth. Maintaining appropriate and effective lines of communication between the home and school is also critical for student success. Parents must understand the student's academic progress and the academic expectations of the grade or subject area. They must also know how to support the students to meet educational goals (“Georgia’s Teacher Keys Effectiveness System,” n.d).

Though not a focus of this study, it should be noted that the standards on which administrators are evaluated differ slightly. They include Instructional Leadership, School Climate, Planning and Assessment, Organizational Management, Human Resources Management, Teacher/Staff Evaluation, Professionalism, and Communication and Community Relations (Georgia Department of Education Laps Reference Sheet Performance Standards and Sample Performance Indicators, n.d.).

Following observations during which teachers are evaluated, certain rights are afforded by the evaluation system. For example, all evaluated personnel have the right to see the notes of their evaluations, evaluations must be posted within five working days of observations, staff may request conferences within ten working days to discuss evaluation scores, and finally, evaluated educators have the right to post rebuttals or comments regarding the observations or evaluator’s comments (Georgia’s Teacher Keys Effectiveness System Meaningful feedback professional growth flexibility to innovate, 2018).

The second component of TKES is the student growth measure, which accounts for thirty percent of the Teacher Effectiveness Measure (TEM) score. Since student growth is the basis for the Teacher evaluative system in Georgia, growth models and percentages are an integral component of the TKES system. There are many advantages to using percentiles rather than simple achievement levels. Unlike other measurement tools, student growth percentiles (SGP’s) offer answers to questions such as whether or not the students are on track, if there is growth in all subjects, and how the students are growing relative to their peers rather than merely a snapshot of the current achievement level. Growth percentiles allow educators to know and understand the students' trajectory for achievement over time (*A guide to the Georgia student growth model*, 2014).

Student growth percentiles provide an alternative tool for Georgia administrators in educational decision-making because low-performing schools often fail to meet content mastery on achievement scores (Turnell, 2018). The Georgia Student Growth Model (GSGM) describes how a student performed on a previous Milestones assessment compared to performance on the latest Milestones assessment (*A guide to the Georgia student growth model*, 2014).

SGP's range from 1 to 99. The categories for growth are as follows: low (1-34), typical (35-65), and high (66-99). SGP percentages are utilized in three different ways. The mean SGP is calculated for the teacher and leader's TEM score because it is statistically reliable as compared to the mean and minimizes error while providing a fair representation of overall teacher and leader effectiveness. However, the median SGP is better suited for general conversation and improvement planning necessary in the State Longitudinal Data System (SLDS), Growth Model Visualization Tools. Finally, the percentage of students scoring in the typical and high growth categories quantifies the percentage of students achieving satisfactory growth levels. It is easily compared to similar measures, i.e., students meeting or exceeding standards on portions of the Milestones assessment (*A guide to the Georgia student growth model*, 2014).

Within the TEM scoring guide, mean SGPs are categorized into four levels for ease with integration into the other components of TKES. The ratings are based on four levels. Level I includes mean SGPs less than 30. Mean SGPs greater than or equal to 30 or less than or equal to 40 fall into level II. Level III scores are greater than 40 or less than or equal to 65. Level IV scores are greater than 65. As with other components of the TEM Score, there are various point values associated with the levels of mean SGPs (*A guide to the Georgia student growth model*, 2014).

As with any evaluation system, allowances must be made so that all teachers may be evaluated similarly, regardless of what subject or grade they teach. Since student growth percentiles are only calculated for elementary teachers of grades four and five, other teachers' growth is assessed through the school growth percentile, the SGP, or the district growth percentile. However, if means to determine student growth in all areas become available, the Georgia Department of Education noted individual growth measures would be used accordingly (Georgia’s Teacher Keys Effectiveness System Meaningful Feedback Professional Growth Flexibility to Innovate, 2018).

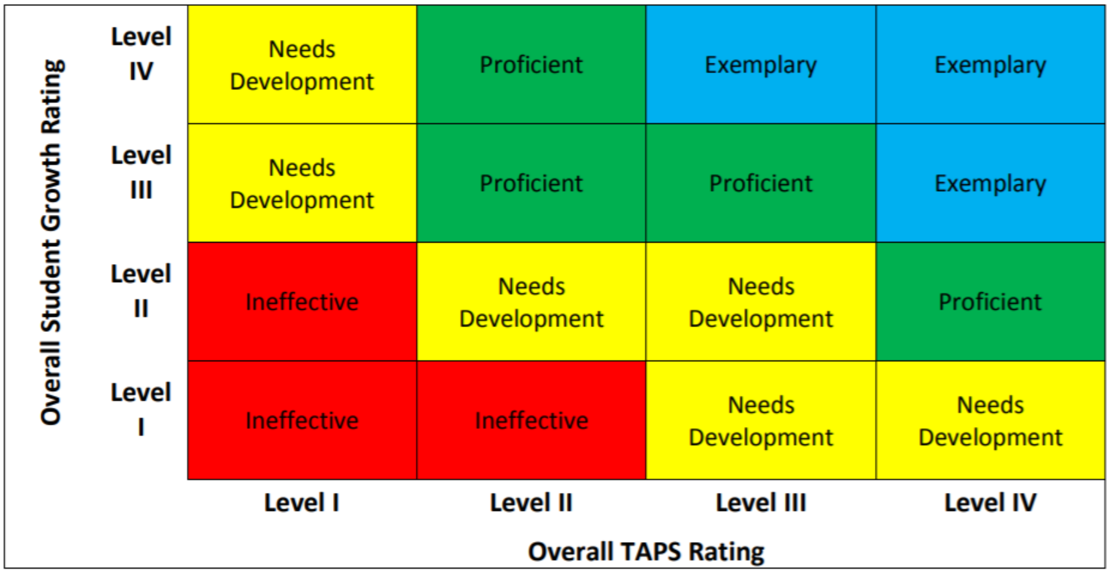
The final component, professional growth, is 20% of the TEM score. Professional Growth is measured by the teacher’s successful completion of their professional growth plan or goal. Teachers new to the profession, district, grade level, or needing remediation have a more intensive plan. More experienced, vetted teachers have more flexibility in working toward their professional learning goals (Georgia’s Teacher Keys Effectiveness System Meaningful Feedback Professional Growth Flexibility to Innovate, 2018).

Following the TEM score calculation, teachers are provided an overall rating: exemplary, proficient, needs development, or ineffective. Failure to receive proficient ratings on TKES evaluations has consequences for educators. Teachers assigned unsatisfactory, ineffective, or needs improvement on summative evaluations within five years are not granted a clear renewable contract. They are placed on the TKES Full Plan. Instead, teachers with two unsatisfactory ratings must apply for a nonrenewable contract until the deficiencies are corrected, as noted in the evaluations. States are required to report teachers with such ratings. Failure to grow in the deficient areas might eventually be a cause for dismissal.

(Georgia's Teacher Keys Effectiveness System Meaningful Feedback Professional Growth Flexibility to Innovate, 2018).

Figure 1

TEM Scoring Guide



Overall Student Growth Rating				
Level IV	Needs Development	Proficient	Exemplary	Exemplary
Level III	Needs Development	Proficient	Proficient	Exemplary
Level II	Ineffective	Needs Development	Needs Development	Proficient
Level I	Ineffective	Ineffective	Needs Development	Needs Development
	Level I	Level II	Level III	Level IV
	Overall TAPS Rating			

(TEM Scoring Guide, 2014)

Summary

Educators need an evaluative tool for continued teacher growth. However, developing and sustaining an effective evaluative tool has been a source of debate for many years. Educational leaders caution against a tool that fails to address research-based teaching practices and encourage professional growth (Marzano et al., 2011). The Georgia Teacher Keys and Effectiveness System was developed based on sound teaching practices (Georgia's Teacher Keys

Effectiveness System Meaningful Feedback Professional Growth Flexibility to Innovate, 2018). Although measures indicate it to be a valid and reliable tool (Elder, Wang, & Cramer, 2015), the connection to student growth and achievement must be established for educators to confidently use evaluation results to make informed decisions that impact American schools.

The following chapters include a detailed description of the remaining components of the study. Chapter 3 describes the methodology. Chapter 4 provides an analysis of the data. The final chapter, chapter 5, describes the conclusions, recommendations, and implications.

Chapter III

METHODOLOGY

This chapter contains the methodology for the study. Included is a description of the research design and a discussion of the independent and dependent variables. The procedures that were necessary to answer the guiding research questions are detailed. Following is a discussion of the target population, sampling procedure, and sample from the Georgia Department of Education (GaDOE). Also included is an explanation of the instruments' validity and reliability, followed by the data collection procedures. Finally, each question's specific quantitative analysis, including specific descriptive and inferential statistics, considerations, and assumptions, are discussed. Chapter 3 concludes with a brief summary of the methodology.

Research Questions

Each of the questions below, including the sub-questions, guided this study:

RQ1: Are summative scores on TKES Standards (Professional Knowledge, Instructional Planning, Instructional Strategies, Differentiated Instruction, Assessment Strategies, Assessment Uses, Positive Learning Environment, Academically Challenging Environment, Professionalism, and Communication) significant predictors of the teacher's SGP level on the Georgia Milestones Assessment System (GMAS)?

a. Are summative scores on TKES Standards significant predictors of the teacher's SGP level (levels I, II, III, or IV) on the fourth grade English/Language Arts portion of the Georgia Milestones Assessment System (GMAS)?
b. Are summative scores on TKES Standards significant predictors of the teacher's SGP level (levels I, II, III, or IV) on the fourth grade Math portion of the Georgia Milestones Assessment System (GMAS)?

c. Are summative scores on TKES Standards significant predictors of the teacher's SGP level (levels I, II, III, and IV) on the fifth grade English/Language Arts portion of the Georgia Milestones Assessment System (GMAS)?

d. Are summative scores on TKES Standards significant predictors of the teacher's SGP level (levels I, II, III, and IV) on the fifth grade Math portion of the Georgia Milestones Assessment System (GMAS)?

RQ2: Are summative scores on TKES Standards (Professional Knowledge, Instructional Planning, Instructional Strategies, Differentiated Instruction, Assessment Strategies, Assessment Uses, Positive Learning Environment, Academically Challenging Environment, Professionalism, and Communication) significant predictors of the teacher's mean scale score on the Georgia Milestones Assessment System (GMAS)?

a. Are summative scores on TKES Standards significant predictors of the teacher's mean scale score on the fourth grade English/Language Arts portion of the Georgia Milestones Assessment System (GMAS)?

b. Are summative scores on TKES Standards significant predictors of the teacher's mean scale score on the fourth grade Math portion of the Georgia Milestones Assessment System (GMAS)?

c. Are summative scores on TKES Standards significant predictors of the teacher's mean scale score on the fifth grade English/Language Arts portion of the Georgia Milestones Assessment System (GMAS)?

d. Are summative scores on TKES Standards significant predictors of the teacher's mean scale score on the fifth grade Math portion of the Georgia Milestones Assessment System (GMAS)?

RQ3: Is there a significant difference in academic setting (departmentalized or self-contained) by level of economically disadvantaged (ED) students on the teacher's mean scale score on the Georgia Milestones Assessment System (GMAS)?

a. Is there a significant difference in academic setting (departmentalized or self-contained) by level of economically disadvantaged (ED) students on the teacher's mean scale score on the fourth grade English/Language Arts portion of the Georgia Milestones Assessment System (GMAS)?
b. Is there a significant difference in academic setting (departmentalized or self-contained) by level of economically disadvantaged (ED) students on the teacher's mean scale score on the fourth grade Math portion of the Georgia Milestones Assessment System (GMAS)?
c. Is there a significant difference in academic setting (departmentalized or self-contained) by level of economically disadvantaged (ED) students on the teacher's mean scale score on the fifth grade English/Language Arts portion of the Georgia Milestones Assessment System (GMAS)?
d. Is there a significant difference in academic setting (departmentalized or self-contained) by level of economically disadvantaged students on the teacher's mean scale score on the fifth grade Math portion of the Georgia Milestones Assessment System (GMAS)?

RQ4: How many reliable and interpretable components are there among the following variables: Professional Knowledge, Instructional Planning, Instructional Strategies, Differentiated Instruction, Assessment Strategies, Assessment Uses, Positive Learning Environment, Academically Challenging Environment, Professionalism, and Communication?

a. How many reliable and interpretable components are there among the following variables: Professional Knowledge, Instructional Planning, Instructional Strategies, Differentiated Instruction, Assessment Strategies, Assessment Uses, Positive Learning Environment, Academically Challenging Environment, Professionalism, and Communication among fourth grade English/Language Arts teachers?

b. How many reliable and interpretable components are there among the following variables: Professional Knowledge, Instructional Planning, Instructional Strategies, Differentiated Instruction, Assessment Strategies, Assessment Uses, Positive Learning Environment, Academically Challenging Environment, Professionalism, and Communication among fourth grade math teachers?

c. How many reliable and interpretable components are there among the following variables: Professional Knowledge, Instructional Planning, Instructional Strategies, Differentiated Instruction, Assessment Strategies, Assessment Uses, Positive Learning Environment, Academically Challenging Environment, Professionalism, and Communication among fifth grade English/Language Arts teachers?

d. How many reliable and interpretable components are there among the following variables: Professional Knowledge, Instructional Planning, Instructional Strategies, Differentiated Instruction, Assessment Strategies, Assessment Uses, Positive Learning Environment, Academically Challenging Environment, Professionalism, and Communication among fourth grade math teachers?

Research Design

The study employed a non-experimental, quantitative design to determine the extent to which TKES scores are predictive of student growth and achievement. Additionally, the study examined the difference the academic setting, self-contained or departmentalized, and levels of the percentage of ED students made on student achievement. The reliability and interpretability of the TKES components were also examined. There was no manipulation of the dependent or independent variables.

The independent variables included the teachers' summative scores on each of the 10 TKES Standards, levels of the percentage of students identified as ED, and academic setting (self-contained or departmentalized). The TKES Standards are as follows: Professional Knowledge, Instructional Planning, Instructional Strategies, Differentiated Instruction, Assessment Strategies, Assessment Uses, Positive Learning Environment, Academically Challenging Environment, Academically Challenging Environment, Professionalism, and Communication. TKES summative scores are used to calculate proficiency levels (I through IV) and are considered ordinal data. The levels of the percent of ED were reported as a percentage between 0 and 100. They were divided into two levels, based on the mean of the subgroup, and were ordinal data. The academic setting, regarded as nominal data, was categorized as self-contained or departmentalized. Self-contained teachers provide instruction on all subjects to the same children without class changes. In a departmentalized setting, teachers only provide instruction on specific subjects to students. Typically, in departmentalized classrooms, one teacher will teach reading, and another teacher will teach math. Other subjects, i.e., social studies and science, usually accompany either ELA or math.

The dependent variables for the first research question were levels of student growth percentiles. The levels of growth percentiles were measured as ordinal data. The levels of student growth were low growth (1-34), typical growth (35-65), and high growth (66-99). The GaDOE calculates growth percentiles for fourth and fifth grade educators based on students' performance within their classrooms (*A Guide to the Georgia Student Growth Model,* 2018). The teacher’s mean scale scores (MSS) on the GMAS served as the dependent variable in question 2 and question 3. The teacher’s MSS is an average of the students’ scale scores. The MSS served as the dependent variable. These data were considered to be on the interval level of measurement.

Participants

This study's target population was fourth and fifth grade English Language Arts and math teachers across Georgia. These teachers were responsible for either ELA or math instruction or both. To have a calculated growth percentile, students must be in the fourth or fifth grade. A teacher's mean growth percentile is based on their students' growth percentiles. Due to the nature of the student growth percentiles, data from the previous year was compared to the following year to determine growth. Because students in Georgia do not take the Milestones Assessment until third grade, student growth percentiles cannot be calculated for a student until fourth grade. Therefore, only teachers in the fourth and fifth grades have student growth percentiles with performance levels (A Guide to the Georgia Student Growth Model, 2014).

The accessible population included a random sample generated by the Georgia Department of Education of fourth and fifth grade English Language Arts and math teachers with student growth percentiles. Approximately 1,800 fourth grade ELA teachers have a growth percentile. In fifth grade ELA, approximately 2,000 teachers have growth percentiles. Similarly, an estimated 1,500 fourth grade teachers and 2,300 fifth grade teachers have growth percentiles in math. Approximately 2,900 fourth grade and 2,300 fifth grade teachers earned student growth percentiles for ELA and math combined (K. Tisdel, personal communication, January 1, 2020).

The Georgia Department of Education (GaDOE) provided the datasets for the study. The GaDOE delivered a random sampling of 4,000 teachers with mean student growth percentiles and mean scale score data, categorized by grade and subject in fourth and fifth grade English Language Arts and math. Additionally, for the same teachers, the GaDOE provided the summative scores for each of the ten standards and the academic setting for each teacher (departmentalized or self-contained), and levels of the percentage of students identified as economically disadvantaged for each respective school. For each grade level, data represented 800 self-contained ELA and math teachers, 600 departmentalized math teachers, and 600 departmentalized ELA teachers. Based on the large sample size, the findings are generalizable to the target population.

Instrumentation

TKES

The GaDOE provided the data for this study. GaDOE provided the TKES scores from the TKES platform. Each teacher in the state receives a summative or overall rating for each of the 10 standards. These standards are Professional Knowledge, Instructional Planning, Instructional Strategies, Differentiated Instruction, Assessment Strategies, Assessment Uses, Positive Learning Environment, Academically Challenging Environment, Academically Challenging Environment, Professionalism, and Communication. Rubrics and look-fors are included in supporting documentation for the rater. Administrators observe teachers up to six times each year, depending on the performance and longevity of the teacher's teaching career (Georgia’s Teacher Keys Effectiveness System, n.d.). Through the totality of the evidence, administrators assign a summative or overall score using the scores on the walkthrough or formative observations as a guide. Scores for the standards are based on a ranking from I to IV. A score of I indicates the mastery of the standard is not evident. A score of II indicates a teacher needs development for mastery of the standard. The expected score, III, indicates a proficient rating. Finally, an IV score indicates the teacher exceeds expectations on the standard (Georgia's Teacher Keys Effectiveness System: Meaningful Feedback Professional Growth Flexibility to Innovate, 2018).

Validity. The validity of the TKES components was established through Kane's (2009, 2013) Arguments-based Approach. The validity was established by outlining the intended uses of the instrument's scores, determining the types of evidence relevant to the uses, gathering the evidence, and assessing the adequacy of the validity evidence to determine if the intended uses are warranted. The instrument was designed with the following purposes: planning of professional development, merit pay decisions, talent management positions, including interventions for teachers and leaders, and renewal/retention and dismissal decisions (Barge, 2013). The TAPS process has explicit instructions for the rater to use the performance standards when rating a teacher. The content validity was established through the comparison of other widely accepted documents. Content validity was established through the close resemblance that the TKES Standards hold to the InTASC Model Core Teaching Standards and Learning Progressions for Teachers (Elder et al., 2015). Although there were potential concerns with the SLO measure, the other elements of TKES were found valid (*Assessing the validity and reliability of the Teacher Keys Effectiveness System (TKES) and the Leader Keys Effectiveness System (LKES) at the Georgia Department of Education*, 2014).

Reliability. Based on a study conducted by the Georgia Center of Assessment in 2014, the TAPS instrument's reliability is high. Polychoric and item correlations were calculated to assess consistency. The ordinal alpha was .95 (*Assessing the validity and reliability of the Teacher Keys Effectiveness System (TKES) and the Leader Keys Effectiveness System (LKES) at the Georgia Department of Education*, 2014). The ordinal alpha is preferred over Cronbach's alpha when data come from items with few response options or show skewness (Gadermann, Guhn, & Zumbo, 2012).

GMAS

GMAS was yet another source of data. The teacher's mean student growth percentiles and mean scale scores, calculated from their students' performance on the Milestones Assessment administration, were used. The Georgia State Legislature formed the GMAS to determine how the students in Georgia are mastering state-mandated standards. Students in grades three through eight and various high school courses must take the Milestones Assessment at the end of courses or grade levels during the spring. The following subject areas' standards are assessed in elementary schools: English Language Arts, math, science, and social studies (Understanding the Georgia Milestones, 2020). Mean scale scores are used to categorize learners into achievement levels.

Table 2

Scale Score Ranges by Achievement Levels

Content Area	Grade Level	Achievement Level 1	Achievement Level 2	Achievement Level 3	Achievemen Level 4
ELA	4	210 to 474	475 to 524	525 to 573	574 to 775
	5	210 to 474	475 to 524	525 to 586	587 to 760
Math	4	270 to 474	475 to 524	525 to 584	585 to 715
	5	265 to 474	475 to 524	525 to 579	580 to 725

Note. Achievement Level for scale score ranges by subject and by grade level. ELA (Milestones ELA), Math (Milestones Math).

The achievement levels are as follows: Beginning learners are classified as level I and need substantial support; developing learners, or level II learners, demonstrate partial proficiency; level III, or proficient learners, are prepared for the next grade level; level IV learners are distinguished with advanced proficiency in the content (Understanding the Georgia Milestones, 2020). Although the scores are initially reported in scale scores, they are transposed into achievement levels for ease of understanding.

Validity. The validity of the GMAS must first be examined through its intended use. The test was designed to measure student progress toward acquiring standards and identify areas for improvement. Therefore, the assessment's content validity rests with the integrity of the process by which the test is vetted. The Georgia Department of Education includes qualified educators from different state areas for test development to ensure validity. Committees of educators then review the content standards that will be assessed by developing the test specifications, the content domain specifications, and the test blueprints. Items specifications are then created for each content area. These specifications later become the Georgia Milestones Assessment Guides, informing all stakeholders of the test's content and assessment method.

Additionally, the test blueprints and content weight documents are posted on the GaDOE website. The test's specific items are written by assessment specialists and vetted by Georgia educators for curriculum suitability, bias, or sensitivity issues. Again, content validity is established by aligning the test items to the material included in the standards. Field tests are conducted to ensure clarity and functionality through field test items. Following this process, committees review item performance and accept, revise, or reject items. Accepted items are banked for inclusion in future tests. Following this process, the test form is created. Each form must assess the same range of content. Tests must be equated using a statistical procedure to ensure equal difficulty (Understanding the Georgia Milestones, 2020).

Once the test is administered, standards are established through a process called standard setting. Educators determine how many total points a student must earn through standard-setting to attain different achievement levels. Finally, test scores are produced, and scores are distributed. Scores are typically reported as scale scores and performance levels. As reported by the Georgia Department of Education, the rigorous process of ensuring assessments is aligned with state standards, and the use of qualified educators is critical to ensuring the validity of the GMAS. Ongoing studies continue to analyze the degree of external validity (Understanding the Georgia Milestones, 2020).

Reliability. The reliability of the 2018 Milestones was measured by Cronbach's alpha reliability coefficient (1951). The measure expressed test scores' consistency as the ratio of actual score variance to the observed total score variance. The ranges are reported from 0 to 1. The average reliabilities in the elementary grades ELA portion of the test range (grades 3, 4, and 5) from .90 to .91. Again, elementary math scores for grades 3, 4, and 5 indicated average reliability of .93. Fifth grade science reliability fell at .92 and social studies at .91. Overall, with all subtests considered, across forms and administrations, the reliabilities range from .88 to .94, indicating high reliability (Georgia Department of Education, 2019).

Data Collection

The Georgia Department of Education provided the data for this study. There were no ethical concerns regarding collecting data in this study, and there was no identifiable information in the data provided by the state. The data sets included, by teacher, the TKES summative scores, student growth percentiles for ELA and math, MSS for the ELA and math sections of the GMAS, academic setting (departmentalized or self-contained), and percentage of ED students. For question 1, summative TKES scores for each standard and category of student growth percentile by the subject were utilized to answer the question. For question 2, the summative TKES Standard scores and the teacher's mean scale score on the Georgia Milestones were examined to answer the research question. The academic setting (departmentalized or self-contained), the percentage of students deemed ED, and the corresponding mean scale score were necessary to answer question 3. For question 4, teachers' scores across all TKES Standards were used to explain correlations among the standards. IRB documentation can be found in Appendix E.

Data Analysis

Data were delivered in an Excel file from the Georgia Department of Education. Data were then be imported into R. R was employed to provide the data analysis and answer the research questions. The purpose of question 1 was to determine if teachers' standard summative TKES scores were predictive of the level of student growth percentiles, low, typical, or high. Ordinal regression was utilized in question 1. The R code used can be found in Appendix A. Question 2 was meant to determine the predictive power of TKES Standard summative scores on student achievement as measured by scale scores on the GMAS. Separate multiple linear regression models were employed to answer research question 2 as seen in the code noted in Appendix B. The purpose of question 3 was to determine if there were significant differences in the scale scores on the ELA and Math portions of the GMAS for students based on academic setting (departmentalized or self-contained) and levels of the percentage of ED students. The code for question 3 is referenced in Appendix C. A factorial ANOVA was utilized to answer this question. The purpose of question 4 was to determine how TKES standards may be reduced to fewer factors or fewer standards. The code for the analysis for question 4 can be found in Appendix D.

Descriptive Statistics

Descriptive statistics were calculated to provide insight and to summarize the characteristics of the data. The n, mean, standard deviation, minimum value, maximum value, skewness, and kurtosis were computed for the interval level variable (GMAS scale scores). The descriptive statistics consisted of frequency and percentages for the categorical variables (academic setting, levels of the percentage of ED students, TKES summative scores, and level of student growth percentiles). These descriptive statistics allowed for a meaningful description of the data.

Inferential Statistics

Through an ordinal regression analysis in question 1, the coefficient estimates with standard errors, z-score, p-value, McFadden R^2, Nagelkerke, Cox-Snell, odds-ratio scale, confidence interval, likelihood ratio test, and a bootstrapping sample were presented. The z-score, along with the p-value, indicated the ability of TKES scores to predict the level of student growth percentiles. Coefficient estimates indicated the probability of student growth percentiles falling within a higher category as TKES scores increase. The p-value indicated statistical significance using .05 as a significance level. The standard error explained how close the estimate might be to the actual coefficient value. The Nagelkerke, Cox-Snell, and McFadden R^2 indicated the variance accounted for between the dependent and independent variables. An odds ratio scale, including confidence intervals, allowed for understanding the odds of achievement level increase for every one-point increase in the TKES scores. The likelihood ratio test compared the fit of the model. Bootstrapping samples were analyzed to provide a more robust representation of the data to increase confidence in the findings.

The multiple regression calculations for research question 2 produced unstandardized and standardized coefficient estimates for the intercept and slope with standard errors, t statistic, degrees of freedom, p-value, R, R^2, R^2_{adj}, and a bootstrapping sample. The coefficient explained the impact the TKES scores have on the mean scale score. The t statistic, degrees of freedom, and the p-value was used to test for statistical significance. The R^2_{adj} value was used to explain the variance accounted for by the independent variables. The bootstrapping method was used to provide a more confident representation of the data.

For research question 3, a factorial ANOVA was utilized, including the Aligned Ranks Transformation approach, Type III sums of squares, F-statistics, degrees of freedom, p-values, and estimated marginal means with standard errors. Additionally, post hoc pairwise comparisons for significant effects (with t statistics, degrees of freedom, and p-values) and graphic-box plots were analyzed. The Type III sum of squares, F statistics, degrees of freedom, and p values provided insight into which individual effects are significantly significant. The F test analyzed the impact of the academic setting and level of economically disadvantaged students on the GMAS scale score. The estimated marginal means aided in understanding the estimated outcome for each setting and level of percentage of economically disadvantaged students. The standard errors analyzed how close the estimates may be to actual mean outcomes. The post hoc pairwise comparisons were used to determine if differences exist in the level of the percentages of ED and academic setting for the teachers' mean score for ELA

The factor analysis used in question 4 included Bartlett's test of sphericity, Kaiser-Meyer-Olkin (KMO) 's overall measure of sampling adequacy, Eigenvalues, determination of correlation matrix, BIC, sample size adjusted BIC, Catell's scree plot, Joliffe's Criterion, Velicer's Map, table of factor loadings, and root mean square residuals for model fit for analysis. Bartlett's test of sphericity tested the overall significance of the correlations within the matrix. The KMO, Eigenvalues, and determinate of the correlation matrix were used in the preliminary analysis to examine the factorability before completing the primary analysis. Joliffe's Criterion, Velicer's Map, BIC, sample size adjusted BIC, and scree plot allowed an understanding of true factors among the various TKES scores. The table of factor loadings indicated the degree to which each TKES score was related to each other. The root means square residual aided in concluding how well the model fit.

Statistical Considerations and Assumptions

Assumptions for using ordinal regression in question 1 included a binary dependent variable. The observations must be independent of each other. There must be little to no multicollinearity indicated by correlations greater than or equal to .90. The final assumption is the proportional odds assumption (Introduction to Generalized Linear Models, n.d.). Statistical considerations were outliers and missing data (Mertler & Vennatta, 2013). The assumptions for the multiple regression analysis in question 2. The assumptions included linearity of the relationship between independent and dependent variables, or TKES scores and percentage of students scoring in the GMAS levels II, III, and IV, homoscedasticity of the variance of the residuals, independence of the observations, and normality of the residuals of the model. The assumptions concerning the residuals included the mean of the residuals for each observation on the dependent variable over many replications being zero, errors associated with any single observation on the dependent variable were independent errors related to any other observation on the dependent variable, the errors were not correlated with the independent variables, the variance of the residuals across all values of the independent variables was constant, and the errors were normally distributed. Statistical considerations were outliers and missing data. The assumptions for question 3 (factorial ANOVA) were the independence of the observations, normality of the residuals, and homoscedasticity of the residuals' variance. The assumptions for question 4 (factor analysis) were multivariate normality and linearity. Statistical considerations included missing data and outliers (Mertler & Vernatta, 2013).

Summary

This chapter describes the proposed methodology on which the study is based. Following the research questions' statement, the non-experimental, quantitative design was described, including the independent and dependent variables' measurement levels. The independent variables included the TKES summative scores, levels of the percentage ED students, and academic setting. The student growth percentiles and the GMAS scale scores were the dependent variables. The target population and sampling population included fourth and fifth grade English Language Arts and math teachers in Georgia. The accessible population and random sample provided by the Georgia Department of Education were also discussed.

The chapter also included a description of the Teacher Keys Effectiveness System and the Georgia Milestones Assessment System and a discussion of each instrument's corresponding validity and reliability. The method through which the Georgia Department of Education delivered the data was noted. The data was imported into R to conduct the complex analysis. A discussion of the descriptive statistics for each question followed. The statistical procedures, including ordinal regression, linear regression, factorial ANOVA, and factor analysis, were identified and discussed. The chapter concluded with considerations and assumptions for each statistical analysis.

The remaining elements of the study are described in chapters 4 and 5. Chapter 4 provides an analysis of the data. Chapter 5 details the conclusions, interpretations, and recommendations.

Chapter IV

RESULTS

The purpose of this study was to identify the relationship between individual TKES Standard scores, student growth percentiles, and student achievement scores for students in fourth and fifth grade ELA and math to determine if specific standards are predictive of student growth and achievement. Additionally, the study also determined whether academic setting, self-contained or departmentalized, and the levels of the percentage of ED students predicted student achievement. This study's results will allow educational leaders to isolate the standards with the most significant impact. Administrators may use the results from this study to make more informed hiring decisions. Using the evidence from this study, educators may make these decisions with confidence following further validation of the reliability and validity of the TKES instrument.

The following questions were answered in this study:

RQ1: Are summative scores on TKES Standards (Professional Knowledge, Instructional Planning, Instructional Strategies, Differentiated Instruction, Assessment Strategies, Assessment Uses, Positive Learning Environment, Academically Challenging Environment, Professionalism, and Communication) significant predictors of the teacher's SGP level on the Georgia Milestones Assessment System (GMAS)?

a. Are summative scores on TKES Standards significant predictors of the teacher's SGP level (levels II, III, or IV) on the fourth grade English/Language Arts portion of the Georgia Milestones Assessment System (GMAS)?

b. Are summative scores on TKES Standards significant predictors of the teacher's SGP level (levels I, II, III, or IV) on the fourth grade Math portion of the Georgia Milestones Assessment System (GMAS)?

c. Are summative scores on TKES Standards significant predictors of the teacher's SGP level (levels I, II, III, and IV) on the fifth grade English/Language Arts portion of the . Georgia Milestones Assessment System (GMAS)?

d. Are summative scores on TKES Standards significant predictors of the teacher's SGP level (levels I, II, III, and IV) on the fifth grade Math portion of the Georgia Milestones Assessment System (GMAS)?

RQ2: Are summative scores on TKES Standards (Professional Knowledge, Instructional Planning, Instructional Strategies, Differentiated Instruction, Assessment Strategies, Assessment Uses, Positive Learning Environment, Academically Challenging Environment, Professionalism, and Communication) significant predictors of the teacher's mean scale score on the Georgia Milestones Assessment System (GMAS)?

a. Are summative scores on TKES Standards significant predictors of the teacher's mean scale score on the fourth grade English/Language Arts portion of the Georgia Milestones Assessment System (GMAS)?

b. Are summative scores on TKES Standards significant predictors of the teacher's mean scale score on the fourth grade Math portion of the Georgia Milestones Assessment System (GMAS)?

c. Are summative scores on TKES Standards significant predictors of the teacher's mean scale score on the fifth grade English/Language Arts portion of the Georgia Milestones Assessment System (GMAS)?

d. Are summative scores on TKES Standards significant predictors of the teacher's mean scale score on the fifth grade Math portion of the Georgia Milestones Assessment System (GMAS)?

RQ3: Is there a significant difference in academic setting (departmentalized or self-contained) by level of economically disadvantaged students on the teacher's mean scale score on the Georgia Milestones Assessment System (GMAS)?

a. Is there a significant difference in academic setting (departmentalized or self-contained) by level of economically disadvantaged students on the teacher's mean scale score on the fourth grade English/Language Arts portion of the Georgia Milestones Assessment System (GMAS)?
b. Is there a significant difference in academic setting (departmentalized or self-contained) by level of economically disadvantaged students on the teacher's mean scale score on the fourth grade Math portion of the Georgia Milestones Assessment System (GMAS)?
c. Is there a significant difference in academic setting (departmentalized or self-contained) by level of economically disadvantaged students on the teacher's mean scale score on the fifth grade English/Language Arts portion of the Georgia Milestones Assessment System (GMAS)?
d. Is there a significant difference in academic setting (departmentalized or self-contained) by level of economically disadvantaged students on the teacher's mean scale score on the fifth grade Math portion of the Georgia Milestones Assessment System (GMAS)?

RQ4: How many reliable and interpretable components are there among the following variables: Professional Knowledge, Instructional Planning, Instructional Strategies, Differentiated Instruction, Assessment Strategies, Assessment Uses, Positive Learning Environment, Academically Challenging Environment, Professionalism, and Communication?

a. How many reliable and interpretable components are there among the following variables: Professional Knowledge, Instructional Planning, Instructional Strategies, Differentiated Instruction, Assessment Strategies, Assessment Uses, Positive Learning Environment, Academically Challenging Environment, Professionalism, and Communication among fourth grade English/Language Arts teachers?

b. How many reliable and interpretable components are there among the following variables: Professional Knowledge, Instructional Planning, Instructional Strategies, Differentiated Instruction, Assessment Strategies, Assessment Uses, Positive Learning Environment, Academically Challenging Environment, Professionalism, and Communication among fourth grade math teachers?

c. How many reliable and interpretable components are there among the following variables: Professional Knowledge, Instructional Planning, Instructional Strategies, Differentiated Instruction, Assessment Strategies, Assessment Uses, Positive Learning Environment, Academically Challenging Environment, Professionalism, and Communication among fifth grade English/Language Arts teachers?

d. How many reliable and interpretable components are there among the following variables: Professional Knowledge, Instructional Planning, Instructional Strategies, Differentiated Instruction, Assessment Strategies, Assessment Uses, Positive Learning Environment, Academically Challenging Environment, Professionalism, and Communication among fourth grade math teachers?

This chapter presents a description of the findings of this study. First, there is an explanation of the population and sample. Following are the descriptive statistics for each variable used in the study accompanied identified by the question in which the variable is analyzed. Then results of the statistical analyses are presented for each research question along with appropriate tables for the given statistical procedure. The chapter concludes with a summary of the results.

Population and Sample

The target population for this study consisted of all fourth grade and fifth grade English Language Arts and Math teachers in public schools in Georgia. Data consisted of a random sample of 4,000 fourth and fifth grade teachers in Georgia with student growth percentiles. For each grade, data represented 800 self-contained teachers and 600 departmentalized teachers for both ELA and math. The departmentalized teachers were responsible solely for the instruction in the ELA or Math segment of the instructional day but taught the same subject to more than one group of students. Conversely, the self-contained teachers were responsible for both ELA and math instruction. The GaDOE provided the ELA and math mean SGP, scale score on the 2019 GMAS, percentage of the levels of ED students represented in the school, academic setting (departmentalized or self-contained) for each teacher, teacher scores for each TKES Standard, and percentage of students scoring in levels II, III, and IV on the GMAS.

Descriptive Statistics

Following is a description of the descriptive statistics for variables used in this study. The analysis for questions 1, 2, and 4 utilized TKES Standard score data. Levels of student growth percentiles (SGP's) were used in the analysis for question 1. Teachers' mean scale scores were analyzed in questions 2 and 3. The percentage of ED students and setting (departmentalized or self-contained) were used in the analysis for question 3. Based on those scores, teachers are ranked as ineffective (level 1), needs development (level II), proficient (level III), and exemplary (level IV). For this study, due to low numbers in the lowest categories, the ineffective and needs development categories (levels I and II) were combined.

In Table 3, the TKES Standard scores in fourth grade ELA are noted. Standards 1 and 5 have the lowest number of teachers reported in the Ineffective/Needs Improvement category, with percentages less than 1%. Conversely, Standard 8 has the highest percentage of teachers in the Ineffective/Needs Improvement category. Standards 2, 4, 5, 7, and 8 have the highest percentages of students in the Proficient categories, with over 80% of teachers falling in this category. Standards 7 and 9 have the highest percentages of teachers in the exemplary category.

Table 3

Fourth Grade ELA Teachers Scoring at Ineffective/Needs Improvement, Proficient, or Exemplary by Professional Standard

TKES Categories	Ineffective/ Needs Improvement	Proficient	Exemplary	*Mdn*	*M*	*SD*
PS 1	10 (0.71%)	1054 (75.29%)	336 (24%)	3.00	3.23	0.44
PS 2	29 (2.07%)	1124 (80.29%)	247 (17.64%)	3.00	3.16	0.42
PS 3	27 (1.57%)	1040 (74.29%)	333 (23.79%)	3.00	3.22	0.46
PS 4	20 (1.43%)	1207 (86.21%)	173 (12.36%)	3.00	3.11	0.35
PS 5	8 (0.57%)	1295 (92.50%)	97 (6.93%)	3.00	3.06	0.27
PS 6	16 (1.14%)	1254 (89.57%)	130 (9.29%)	3.00	3.08	0.31
PS 7	21 (1.50%)	867 (61.93%)	512 (36.57%)	3.00	3.35	0.51
PS 8	42 (3.00%)	1145 (81.79%)	213 (15.21%)	3.00	3.12	0.41
PS 9	22 (1.57%)	825 (58.93%)	553 (39.50%)	3.00	3.38	0.52
PS 10	15 (1.07%)	1111 (79.36%)	274 (19.57%)	3.00	3.19	0.42

Note. PS 1 (Professional Standard 1), PS 2 (Professional Standard 2), PS 3 (Professional Standard 3), PS 4 (Professional Standard 4), PS 5 (Professional Standard 5), PS 6 (Professional Standard 6), PS 7 (Professional Standard 7), PS 8 (Professional Standard 8), PS 9 (Professional Standard 9), PS 10 (Professional Standard 10).

Fourth grade math scores are indicated in Table 4. In this grade level and subject, Standards 1, 5, 6, and 10 have the lowest percentage of students in the lowest category, while the highest number of teachers in the category is indicated in Standard 8. The highest percentage of teachers in the Proficient category can be found in Standards 5, 6, and 7. Both Standards 7 and 9 have the highest percentage of students in the exemplary category.

Table 4

Fourth Grade Math Teachers Scoring at Ineffective/Needs Improvement, Proficient, or Exemplary by Professional Standard

TKES Categories	Ineffective/ Needs Improvement	Proficient	Exemplary	*Mdn*	*M*	*SD*
PS 1	9 (0.64%)	1015 (72.50%)	376 (26.86%)	3.00	3.23	0.45
PS 2	26 (1.86%)	1099 (78.50%)	275 (19.64%)	3.00	3.18	0.43
PS 3	27 (1.93%)	1010 (72.14%)	363 (25.93%)	3.00	3.24	0.47
PS 4	24 (1.71%)	1177 (84.07%)	199 (14.21%)	3.00	3.13	0.38
PS 5	11 (0.79%)	1265 (90.36%)	124 (8.86%)	3.00	3.08	0.30
PS 6	12 (0.86%)	1229 (87.79%)	159 (9.21%)	3.00	3.11	0.33
PS 7	20 (1.43%)	843 (60.21%)	537 (38.36%)	3.00	3.37	0.51
PS 8	36 (2.57%)	1111 (79.36%)	253 (18.07%)	3.00	3.16	0.43
PS 9	23 (1.64%)	816 (58.29%)	561 (40.07%)	3.00	3.38	0.52
PS 10	15 (0.34%)	1090 (77.86%)	295 (21.07%)	3.00	3.20	0.43

Note. PS 1 (Professional Standard 1), PS 2 (Professional Standard 2), PS 3 (Professional Standard 3), PS 4 (Professional Standard 4), PS 5 (Professional Standard 5), PS 6 (Professional Standard 6), PS 7 (Professional Standard 7), PS 8 (Professional Standard 8), PS 9 (Professional Standard 9), PS 10 (Professional Standard 10).

As indicated in Table 5, less than 1% of teachers scored in the ineffective/needs development category for Standards 1, 5, and 10 in fifth grade ELA. While no categories had over 3% at the lowest level, Standard 8 had the largest group of teachers (2.71%) scoring at that level. More than 80% of the teachers in the sample had ratings in the Proficient category for Standards 2, 4, 5, 6, 7, and 8. Over 30% of teachers rated as exemplary in Standards 7 and 9.

Table 5

Fifth Grade ELA Teachers Scoring at Ineffective/Needs Improvement, Proficient, or Exemplary by Professional Standard

TKES Categories	Ineffective/ Needs Improvement	Proficient	Exemplary	*Mdn*	*M*	*SD*
PS 1	7 (0.50%)	1003 (71.64%)	390 (27.86%)	3.00	3.27	0.46
PS 2	31 (2.21%)	1134 (81%)	235 (16.79%)	3.00	3.15	0.41
PS 3	22 (1.57%)	1027 (73.36%)	351 (25.07%)	3.00	3.24	0.46
PS 4	29 (2.07%)	1210 (84.43%)	161 (11.50%)	3.00	3.09	0.36
PS 5	7 (0.50%)	1297 (92.64%)	96 (6.86%)	3.00	3.06	0.26
PS 6	19 (1.46%)	1234 (88.14%)	147 (10.50%)	3.00	3.09	0.33
PS 7	27 (1.93%)	862 (61.57%)	511 (36.50%)	3.00	3.35	0.51
PS 8	38 (2.71%)	1123 (80.21%)	239 (17.07%)	3.00	3.14	0.42
PS 9	16 (1.14%)	794 (56.71%)	590 (42.14%)	3.00	3.41	0.51
PS 10	12 (0.86%)	1104 (78.86%)	284 (20.29%)	3.00	3.19	0.42

Note. PS 1 (Professional Standard 1), PS 2 (Professional Standard 2), PS 3 (Professional Standard 3), PS 4 (Professional Standard 4), PS 5 (Professional Standard 5), PS 6 (Professional Standard 6), PS 7 (Professional Standard 7), PS 8 (Professional Standard 8), PS 9 (Professional Standard 9), PS 10 (Professional Standard 10).

Table 6 represents the fifth grade math teachers in the sample. Less than 1% of teachers were rated as Ineffective/Needs Improvement in Standards 1, 5, and 10. For that category, Standard 8 had the highest number with 3% of teachers. Over 80% of teachers performed at the Proficient level in Standards 4, 5, and 6. More teachers scored in the exemplary level for Standards 1, 7, and 9 than other standards.

An analysis of each table indicates similar patterns across the grade levels and subjects. For example, over 90% of teachers were given a Proficient rating on Standard 5 across all subgroups. Standard 5 also held the lowest percentage of teachers scoring at the exemplary level. The percentage of teachers scoring at the Proficient level for Standard 6 was above 86% in fourth and fifth grade ELA and Math. Standards 7 and 9 consistently had the highest percentage of teachers at the exemplary level. Across all grade levels and subjects, the number of teachers given an Ineffective/Needs Improvement was highest in Standard 8. Standards 7 and 9 had the greatest standard deviation at all levels and subjects, while Standard 5 had the lowest standard deviation.

Table 6

Fifth Grade Math Teachers Scoring at Ineffective/Needs Improvement, Proficient, or Exemplary by Professional Standard

TKES Categories	Ineffective/ Needs Improvement	Proficient	Exemplary	*Mdn*	*M*	*SD*
PS 1	12 (0.86%)	964 (68.86%)	424 (30.29%)	3.00	3.29	0.47
PS 2	41 (2.93%)	1117 (79.79%)	242 (17.29%)	3.00	3.14	0.43
PS 3	35 (2.50%)	1011 (72.21%)	354 (25.29%)	3.00	3.23	0.48
PS 4	30 (2.14%)	1192 (85.14%)	178 (12.71%)	3.00	3.11	0.37
PS 5	9 (0.64%)	1262 (90.14%)	129 (9.21%)	3.00	3.09	0.30
PS 6	24 (1.71%)	1217 (86.93%)	159 (11.36%)	3.00	3.10	0.35
PS 7	31 (2.21%)	859 (61.36%)	510 (36.43%)	3.00	3.34	0.52
PS 8	42 (3.00%)	1090 (77.86%)	268 (19.14%)	3.00	3.16	0.44
PS 9	24 (1.71%)	807 (57.64%)	569 (40.64%)	3.00	3.39	0.52
PS 10	13 (0.93%)	1102 (78.71%)	285 (20.36%)	3.00	3.19	0.42

Note. PS 1 (Professional Standard 1), PS 2 (Professional Standard 2), PS 3 (Professional Standard 3), PS 4 (Professional Standard 4), PS 5 (Professional Standard 5), PS 6 (Professional Standard 6), PS 7 (Professional Standard 7), PS 8 (Professional Standard 8), PS 9 (Professional Standard 9), PS 10 (Professional Standard 10).

In Table 7, polychoric and polyserial correlations were generated. Polychoric correlations were generated among the TKES Standards. In fourth grade ELA, moderate to strong polychoric correlations were found between Standard 1 and Standard 3, *r*(1398) = .70, *p* <. 05, and Standard 3 and 4, *r*(1398) = .70, *p* <. 05. Standard 9 and 10 also had a moderate to strong correlation, *r*(1398) = .73, *p* <. 05, along with Standard 5 and 6 and Standard 3 and 8, *r*(1398) = .76, *p* <. 05. Standard 5 and Standard 7, *r*(1398) = .46, *p* <. 05, Standard 5 and 9, *r*(1398) = .46, , *p* <. 05, and Standard 8 and 9, again with *r*(1398) = .46, *p* <. 05, had weak to moderate correlations. Polyserial correlations were used to measure the correlations between the standards and the MSS. Standard 8, *r*(1398) = .40, *p* <. 05, Standard 7, *r*(1398) = .33, *p* <. 05, and Standard 3, *r*(1398) = .31, *p* <. 05, and Standard 2, *r*(1398) = .30, *p* <. 05, had weak to moderate polyserial correlations to MSS. With the exception of these standards, the remaining correlations between standards and MSS were weak.

Table 7

Polychoric and Polyserial Correlations among TKES Standard Scores and Mean Scale Score (MSS) for Fourth Grade ELA

	MSS	PS 1	PS 2	PS 3	PS 4	PS 5	PS 6	PS 7	PS 8	PS 9	PS 10
MSS											
PS 1	.29										
PS 2	.30	.61									
PS 3	.31	.70	.64								
PS 4	.25	.57	.62	.70							
PS 5	.25	.60	.54	.54	.52						
PS 6	.22	.55	.61	.56	.65	.76					
PS 7	.33	.61	.63	.68	.66	.46	.49				
PS 8	.40	.66	.65	.76	.65	.60	.60	.64			
PS 9	.24	.55	.53	.49	.48	.46	.52	.60	.46		
PS 10	.29	.56	.53	.53	.52	.51	.52	.66	.52	.73	

Note. MSS (Mean Scale Score), PS 1 (Professional Standard 1), PS 2 (Professional Standard 2), PS 3 (Professional Standard 3), PS 4 (Professional Standard 4), PS 5 (Professional Standard 5), PS 6 (Professional Standard 6), PS 7 (Professional Standard 7), PS 8 (Professional Standard 8), PS 9 (Professional Standard 9), PS 10 (Professional Standard 10).

For fourth grade math (Table 8), correlations were also measured between standards using polychoric correlations. Standard 9 and Standard 10, *r*(1398) = .70, *p* < .05, Standard 4 and Standard 8, *r*(1398) = .71, *p* < .05, and Standard 3 and 8, *r*(1398) = .78, *p* < .05, had moderate to strong correlations. Standards 5 and 6 had a strong correlation, *r*(1398) = .84, *p* < .05. Standards 3 and 9, *r*(1398) = .45, *p* < .05 and Standards 8 and 9, *r*(1398) = .43, *p* < .05, had a weak to moderate correlation. Again, through the use of polyserial correlations, Standard 8, *r*(1398) = .37, *p* <. 05, Standard 2, *r*(1398) = .36, *p* <. 05, Standard 1, *r*(1398) = .35, *p* <. 05, Standard 5, *r*(1398) = .33, *p* <. 05, and Standard 7, *r*(1398) = .30, *p* <. 05, had weak to moderate polyserial correlations to MSS. Weak correlations were found between the remaining standards and the mean scale score.

Table 8

Polychoric and Polyserial Correlations among TKES Standard Scores and Mean Scale Score (MSS) for Fourth Grade Math

	MSS	PS 1	PS 2	PS 3	PS 4	PS 5	PS 6	PS 7	PS 8	PS 9	PS 10
MSS											
PS 1	.35										
PS 2	.36	.63									
PS 3	.29	.68	.65								
PS 4	.29	.60	.62	.69							
PS 5	.33	.65	.67	.57	.60						
PS 6	.27	.62	.66	.56	.68	.84					
PS 7	.30	.60	.64	.69	.64	.52	.52				
PS 8	.37	.69	.69	.78	.71	.66	.57	.64			
PS 9	.26	.57	.57	.45	.49	.58	.54	.57	.43		
PS 10	.27	.59	.58	.53	.53	.60	.48	.66	.48	.70	

Note. MSS (Mean Scale Score), PS 1 (Professional Standard 1), PS 2 (Professional Standard 2), PS 3 (Professional Standard 3), PS 4 (Professional Standard 4), PS 5 (Professional Standard 5), PS 6 (Professional Standard 6), PS 7 (Professional Standard 7), PS 8 (Professional Standard 8), PS 9 (Professional Standard 9), PS 10 (Professional Standard 10).

Table 9 represents the fifth grade ELA data. In this table, Standards 5 and 6 showed the a moderate to strong polychoric correlation, $r(1398) = .78, p < .05$. Standard 3 and 8, $r(1398) = .75, p < .05$, and Standard 3 and 7, $r(1398) = .70, p < .05$, also had a moderate to strong correlation. Standard 4 and Standard 9, $r(1398) = .43, p < .05$, Standard 6 and Standard 9, $r(1398) = .41, p < .05$, and Standards 4 and 10, $r(1398) = .40, p < .05$, had weak to moderate correlations. Standard 8, again, had a weak to moderate polyserial correlation to the MSS, $r(1398) = .35, p < .05$. Again, using polyserial correlations, Standard 8, $r(1398) = .35$, p <. 05, Standard 9, *r*(1398) = .32, *p* <. 05, and Standard 3, *r*(1398) = .30, *p*<. 05 had weak to moderate polyserial correlations to MSS. The remaining standards had weak correlations to the MSS.

Table 9

Polychoric and Polyserial Correlations among TKES Standard Scores and Mean Scale Score (MSS) for Fifth Grade ELA

	MSS	PS 1	PS 2	PS 3	PS 4	PS 5	PS 6	PS 7	PS 8	PS 9	PS 10
MSS											
PS 1	.26										
PS 2	.27	.61									
PS 3	.30	.64	.66								
PS 4	.23	.57	.65	.67							
PS 5	.25	.66	.66	.60	.61						
PS 6	.22	.62	.60	.60	.65	.78					
PS 7	.29	.58	.55	.70	.55	.55	.47				
PS 8	.35	.66	.65	.75	.65	.60	.60	.60			
PS 9	.32	.52	.53	.51	.43	.45	.41	.61	.50		
PS 10	.21	.57	.53	.48	.40	.54	.51	.58	.46	.67	

Note. MSS (Mean Scale Score), PS 1 (Professional Standard 1), PS 2 (Professional Standard 2), PS 3 (Professional Standard 3), PS 4 (Professional Standard 4), PS 5 (Professional Standard 5), PS 6 (Professional Standard 6), PS 7 (Professional Standard 7), PS 8 (Professional Standard 8), PS 9 (Professional Standard 9), PS 10 (Professional Standard 10).

In fifth grade math (see Table 10), Standard 1 and 3 were moderately correlated, *r*(1398) = .72, *p* < .05. Standard 3 and 8, *r*(1398) = .81, *p* < .05, and Standard 5 and 6, *r*(1398) = .82, *p* < .05, had strong polychronic correlations. For lower correlation values, Standard 6 and Standard 9, *r*(1398) = .42, *p* < .05, and Standards 6 and 7, *r*(1398) = .41, *p* < .05 had weak to moderate correlations. Standard 8, *r*(1398) = .38, *p* <. 05, Standard 1, *r*(1398) = .33, *p* <. 05, Standard 3, *r*(1398) = .31, *p* <. 05, Standard 5, *r*(1398) = .31, *p* <. 05, Standard 7, *r*(1398) = .31, *p* <. 05, and Standard 6, *r*(1398) = .30, *p* <. 05, had weak to moderate polyserial correlations to MSS. The remaining standards had weak correlations to the mean scale score.

Table 10

Polychoric and Polyserial Correlations among TKES Standard Scores and Mean Scale Score (MSS) for Fifth Grade Math

	MSS	PS 1	PS 2	PS 3	PS 4	PS 5	PS 6	PS 7	PS 8	PS 9	PS 10
MSS											
PS 1	.33										
PS 2	.27	.66									
PS 3	.31	.72	.68								
PS 4	.24	.55	.66	.65							
PS 5	.31	.62	.66	.66	.65						
PS 6	.30	.62	.62	.63	.65	.82					
PS 7	.31	.62	.58	.68	.56	.51	.41				
PS 8	.38	.69	.69	.81	.68	.64	.63	.63			
PS 9	.27	.55	.52	.53	.48	.49	.42	.60	.51		
PS 10	.22	.58	.59	.47	.44	.51	.47	.58	.48	.69	

Note. MSS (Mean Scale Score), PS 1 (Professional Standard 1), PS 2 (Professional Standard 2), PS 3 (Professional Standard 3), PS 4 (Professional Standard 4), PS 5 (Professional Standard 5), PS 6 (Professional Standard 6), PS 7 (Professional Standard 7), PS 8 (Professional Standard 8), PS 9 (Professional Standard 9), PS 10 (Professional Standard 10).

Student growth percentiles (SGP) were also used in the analysis for question 1. SGP's are measures of a student's growth from one testing year to another. Teachers are assigned a mean SGP based on the growth of the students in their classroom. Only fourth and fifth grade teachers are assigned a mean SGP because SGP's are based on growth from one year to the next, and students take the GMAS for the first time in third grade. Accordingly, the fourth grade year is the earliest SGP's can be calculated. SGP's are categorized in levels I, II, III, and IV. Level I represents SGP's less than 30. Level II SGP's are greater than or equal to 30 or less than 40. Level III scores are greater than 40 but less than or equal to 65. SGP's greater than 65 are Level IV.

In Table 11, the number and percentage of teachers falling into each SGP rating are included. Also noted are the median, mean, standard deviation, minimum, and maximum values. The median across the grade levels and subjects fell between 49.78 and 51.05. Grade 5 Math has the lowest median among the four datasets. More teachers in this grade level and subject fell within the Ineffective and Needs Improvement range for SGP's. Overall, fourth grade teachers had higher median SGP's than fifth grade teachers.

Table 11

SGP Ratings for Teachers by Grade Level and Subject

SGP Categories	Ineffective	Needs Development	Proficient	Exemplary	*Mdn*	*M*	*SD*	Min	Max
Grade 4 ELA	17 (1.21%)	160 (11.43%)	1135 (81.07%)	88 (6.29%)	50.38	50.46	46.91	24.08	85.38
Grade 4 Math	52 (3.71%)	211 (15.07%)	992 (70.86%)	145 (10.36%)	51.05	50.84	62.32	19.44	87.29
Grade 5 ELA	10 (0.71%)	148 (10.57%)	1165 (83.21%)	77 (5.50%)	50.24	50.63	43.03	22.50	81.00
Grade 5 Math	74 (5.29%)	216 (15.43%)	947 (67.64%)	163 (11.64%)	49.78	50.16	67.96	17.29	85.80

Note. Grade 4 ELA (Grade 4 ELA GMAS), Grade 4 Math (Grade 4 Math GMAS), Grade 5 ELA (Grade 5 ELA GMAS), Grade 5 Math (Grade 5 Math GMAS).

GMAS mean scale scores were used in the analysis for questions 2 and 3. Table 12 provides the mean, median, and standard deviation. The minimum, maximum, kurtosis, and skew values for the mean scale scores are also provided. Higher mean scale scores were found in fourth grade math (*M* = 524.62) than the other grade levels and subjects. Fifth grade ELA and math mean scores were very similar, while there was a greater difference in mean scale scores between ELA and math in fourth grade. The standard deviation for fourth grade math was 29.39. More variance was found in fifth grade math (*SD* = 32.80).

Table 12

GMAS Teacher's Mean Scale Scores (MSS) in Fourth and Fifth Grade ELA and Math

Grade/Subject	*Mdn*	*M*	*SD*	Min	Max	*Skew*	*Kurtosis*
Grade 4 ELA	507.78	508.56	29.39	410.00	602.78	0.79	0.00
Grade 4 Math	522.14	524.62	30.02	450.53	636.93	0.80	0.51
Grade 5 ELA	511.5	514.42	32.04	418.11	624.93	0.86	0.07
Grade 5 Math	509.8	514.85	32.80	432.12	670.70	0.81	1.06

Note. Grade 4 ELA (Grade 4 ELA GMAS), Grade 4 Math (Grade 4 Math GMAS), Grade 5 ELA (Grade 5 ELA GMAS), Grade 5 Math (Grade 5 Math GMAS).

The percentage of ED students was used for analysis in research question 3. Each ED category had a minimum of zero and a maximum of 100. The statistics for this population are referenced in Table 13. The ED levels were divided into upper and lower categories based on the median for each grade level. Fourth grade ELA ED groups were comprised of 699 in the lower ED group and 701 in the upper group. Conversely, in fifth grade math, the lower category was comprised of 701 students and 699 in the higher ED group. Fifth grade ELA was divided equally with 700 students in each category of upper and lower ED groups. There were 701 students in the lower category in fifth grade math and 699 in the upper category.

Table 13

SGP Information for Economically Disadvantaged (ED) Students

Grade/Subject	*Mdn*	*M*	*SD*	*Skew*	*Kurtosis*	*SEM*
Grade 4 ELA	73.68	67.54	30.82	-0.54	-0.96	0.82
Grade 4 Math	74.47	67.63	31.44	-0.57	-0.97	0.84
Grade 5 ELA	75.56	67.71	31.56	-0.61	-0.96	0.84
Grade 5 Math	76.00	67.88	31.73	-.062	-0.94	0.85

Note. Grade 4 ELA (Grade 4 ELA GMAS), Grade 4 Math (Grade 4 Math GMAS), Grade 5 ELA (Grade 5 ELA GMAS), Grade 5 Math (Grade 5 Math GMAS).

The academic setting was also used in the analysis for question 3. Teachers were categorized as either self-contained or departmentalized teachers. Self-contained teachers taught all subjects to one group of students throughout the day. Departmentalized teachers only taught reading or math to different groups of students. Each dataset for question 3 included 600 departmentalized teachers and 800 self-contained teachers.

Results by Question

Research Question 1

Are summative scores on TKES Standards (Professional Knowledge, Instructional Planning, Instructional Strategies, Differentiated Instruction, Assessment Strategies, Assessment Uses, Positive Learning Environment, Academically Challenging Environment, Professionalism, and Communication) significant predictors of the teacher's SGP level on the Georgia Milestones Assessment System (GMAS)?

Fourth Grade ELA

a. Are summative scores on TKES Standards significant predictors of the teacher's SGP level (levels I, II, III, or IV) on the fourth grade English/Language Arts portion of the Georgia Milestones Assessment System (GMAS)?

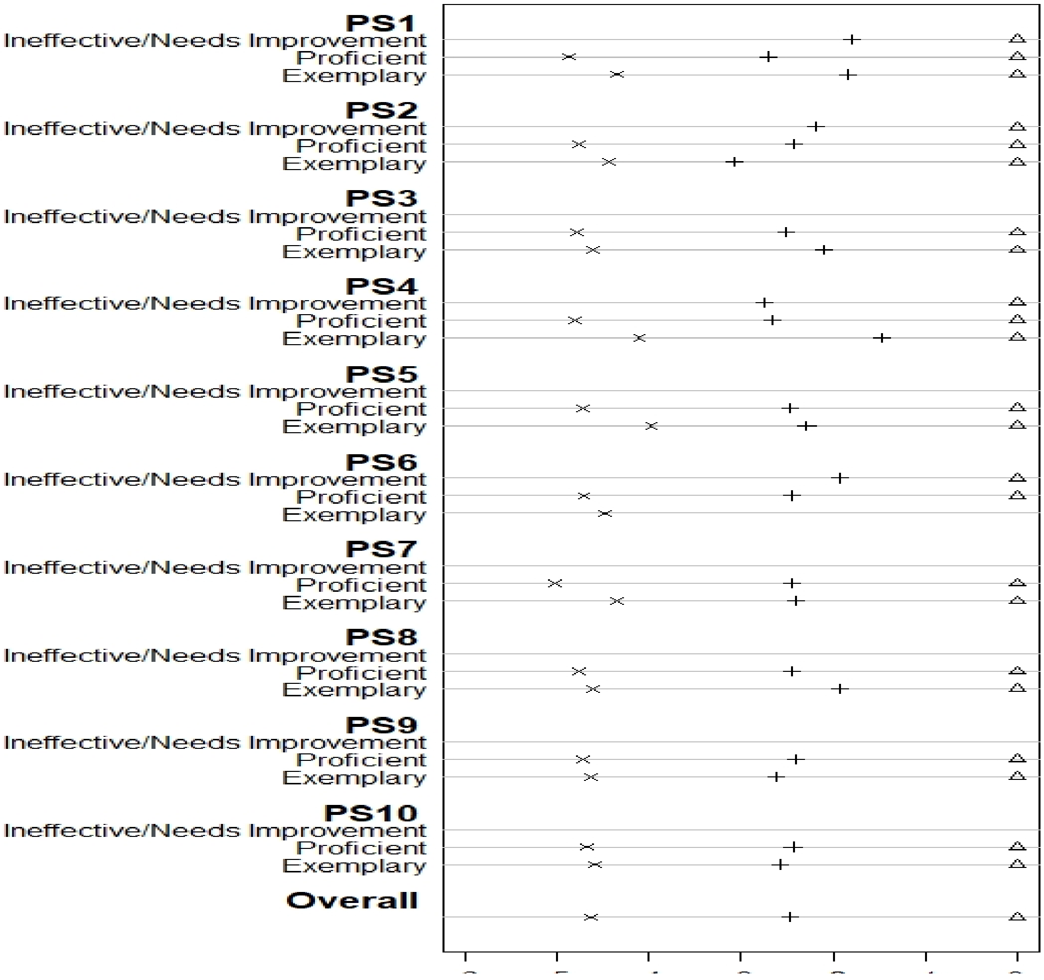
There are four assumptions that must be met for ordinal logistic regression. The first assumption was that the dependent variable was binary. That assumption was met. Next, the observations must be independent of each other. The assumption of independent observations was also met. Next, no correlations were greater than or equal to .90, indicating extreme collinearity. The assumption of little or no multicollinearity was also met.

The fourth assumption of ordinal regression is the proportional odds assumption. Binary logit regressions were utilized as a robustness check for the proportional odds assumption underlying the ordinal regression model. If the proportional odds assumption were precisely true, the coefficients in each of the regressions would be proportional. However, the coefficients are relatively different across each binary logistic regression. For example, in the regression using the binary variable for SGP level > = 2, the Standard 1 at the proficient and exemplary level coefficients are 2.36 and 2.20, respectively. However, in the regression using a binary variable for the SGP levels > = 4, these coefficients are 12.32 and 13.56, respectively.

To check the proportional odds assumption, a visual examination is helpful. For example, proficiency on Standard 1 has the same effect on the odds of obtaining ineffective versus needs development as it does needs development versus proficient. In Figure 2, the results for each binary regression employ different cut points of the dependent variable. If the proportional odds assumption were met precisely, each symbol, representing a coefficient from a binary logistic regression with a different cut point, would be in a straight line down the plot. There was some variance noted, especially for the lower SGP levels. However, it was determined the proportional odds assumption was also met.

Figure 2

Proportional Odds Assumption: Fourth Grade ELA



A likelihood ratio test was used to compare the fit of this ordinal regression model to an intercept-only model, $X^2(20, N = 1400) = 69.99$, $p<.001$. Therefore, there is a statistical difference between the intercept-only and the full models. The full ordinal regression model was a better fit and more conclusive than the intercept-only model. The Cox-Snell, Nagelkerke, and McFadden pseudo-R-squared methods yielded values of $r^2 = 5\%$, 7%, and 4%, respectifourvely, indicating the variance accounted for between the TKES summative scores and teacher's SGP level.

Table 14 shows the results from the ordinal logistic regression analysis for fourth grade ELA. Of the twenty predictor variables, four were found to be statistically significant. Standard 3 at the proficient level had a negative impact on student growth, z = -1.99, *p*<.05, Odds Ratio = 0.31, 95% CI [0.10, 0.96]. Standard 4 at the proficient level (z = 2.53, *p*<.05, Odds Ratio = 4.15, 95% CI [1.35,12.53]), Standard 4 at the exemplary level (z = 2.36, *p*<.05, Odds Ratio = 4.31, 95% CI [1.26, 14.52]), and Standard 8 at the exemplary level (z = 1.99, *p*<.05, Odds Ratio = 2.53, 95% CI [1.00, 6.25]) were found to have a positive impact on student growth. Standards 4 at proficient and ixemplary Levels and Standard 8 at the exemplary level are associated with 4.14, 4.30, and 2.53, respectively, times higher odds of obtaining a higher SGP level than obtaining an ineffective/needs Improvement on each standard. The weakest predictor was Standard 5 at the proficient level (z = -.92, *p*<.05, Odds Ratio = 0.41, 95% CI [0.06, 2.48]). Using the model to predict in-sample SGP levels, approximately 81% of the SGP values were classified correctly.

Table 14

Variables Used to Predict SGP's Utilizing Ordinal Logistic Regression in Fourth Grade ELA

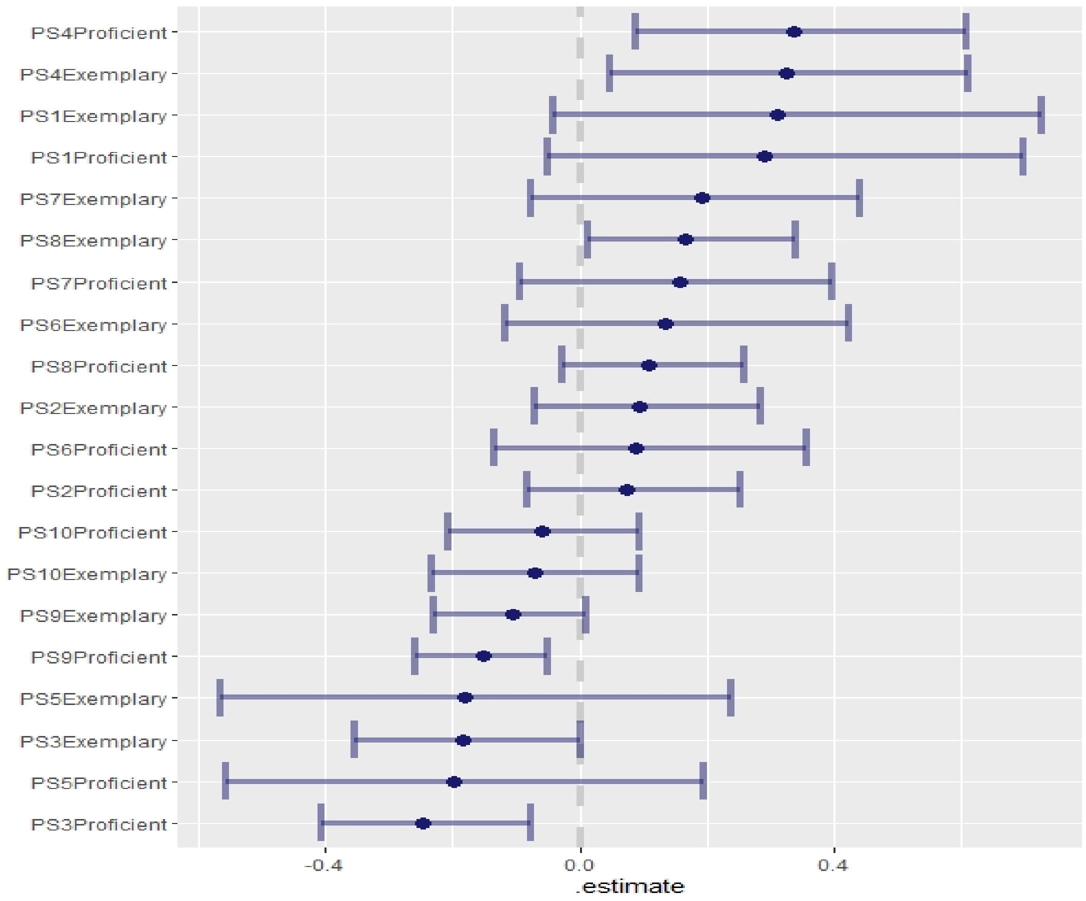
							95% Confidence Interval	
Predictor	Log Odds	*SE*	*Z*	Pr(>|z|)		OR	*Lower*	*Upper*
PS 1 Proficient	1.116	0.788	1.434	0.152		3.053	0.629	13.684
PS 1 Exemplary	1.321	0.804	1.643	0.100		3.745	0.736	17.680
PS 2 Proficient	0.405	0.484	0.837	0.402		1.499	0.559	3.743
PS 2 Exemplary	0.500	0.532	0.940	0.347		1.649	0.563	4.555
PS 3 Proficient	-1.166	0.588	-1.994	0.046	*	0.312	0.099	0.964
PS 3 Exemplary	-0.779	0.624	-1.249	0.212		0.459	0.136	1.538
PS 4 Proficient	1.422	0.564	2.525	0.012	*	4.147	1.352	12.525
PS 4 Exemplary	1.460	0.620	2.355	0.019	*	4.306	1.263	14.520
PS 5 Proficient	-0.884	0.961	-0.920	0.358		0.413	0.061	2.475
PS 5 Exemplary	-0.749	1.014	-0.739	0.460		0.473	0.623	3.174
PS 6 Proficient	0.389	0.641	0.607	0.544		1.476	0.402	4.990
PS 6 Exemplary	0.587	0.705	0.833	0.405		1.799	0.436	6.937
PS 7 Proficient	0.706	0.574	1.232	0.218		2.027	0.638	6.097
PS 7 Exemplary	0.900	0.599	1.503	0.133		2.461	0.739	7.798
PS 8 Proficient	0.556	0.394	1.413	0.158		1.744	0.788	3.707
PS 8 Exemplary	0.928	0.468	1.985	0.047	*	2.530	0.997	6.254
PS 9 Proficient	-0.797	0.560	-1.423	0.155		0.451	0.152	1.328
PS 9 Exemplary	-0.561	0.581	-0.965	0.335		0.571	0.185	1.757
PS 10 Proficient	-0.345	0.704	-0.490	0.624		0.708	0.180	2.730
PS 10 Exemplary	-0.442	0.735	-0.601	0.548		0.643	0.154	2.642

Note: p< 0.001 '***', p<.01 '**', p<.05 '*', PS 1 Proficient (Professional Standard 1 Proficient Level), PS 1 Exemplary (Professional Standard 1 Exemplary Level), PS 2 Proficient (Professional Standard 2 Proficient Level), PS 2 Exemplary (Professional Standard 2 Exemplary Level), PS 3 Proficient (Professional Standard 3 Proficient Level), PS 3 Exemplary (Professional Standard 3 Exemplary Level), PS 4 Proficient (Professional Standard 4 Proficient Level), PS 4 Exemplary (Professional Standard 4 Exemplary Level), PS 5 Proficient (Professional Standard 5 Proficient Level), PS 5 Exemplary (Professional Standard 5 Exemplary Level), PS 6 Proficient (Professional Standard 6 Proficient Level), PS 6 Exemplary (Professional Standard 6 Exemplary Level), PS 7 Proficient (Professional Standard 7 Proficient Level), PS 7 Exemplary (Professional Standard 7 Exemplary Level), PS 8 Proficient (Professional Standard 8 Proficient Level), PS 8 Exemplary (Professional Standard 8 Exemplary Level), PS 9 Proficient (Professional Standard 9 Proficient Level), PS 9 Exemplary (Professional Standard 9 Exemplary Level), PS 10 Proficient (Professional Standard 10 Proficient Level), PS 10 Exemplary (Professional Standard 10 Exemplary Level).

Figure 3 provides a graphic representation of the 1,000 bootstrapping samples of the dataset, rather than one run of the data as represented in Table 14. The data replication through the bootstrapping sample allows for a more accurate and robust representation of the data set. Therefore, there is a higher level of confidence in the outcome, as represented in Figure 3. In addition to the standards found significant using ordinal logistic regression (Standard 4 at the exemplary and proficient level, Standard 8 at the exemplary level, and Standard 3 at the proficient level), Standard 9 at the proficient level, and Standard 3 at the exemplary level were also found to have a negative impact on SGP's using the bootstrapping technique. Again, these conclusions were based on 1,000 bootstrapping samples, allowing for a high level of confidence in the results.

Figure 3

Bootstrapping Sample: Fourth Grade ELA



Fourth Grade Math

b. Are summative scores on TKES Standards significant predictors of the teacher's SGP level (levels I, II, III, or IV) on the fourth grade Math portion of the Georgia Milestones Assessment System (GMAS)?

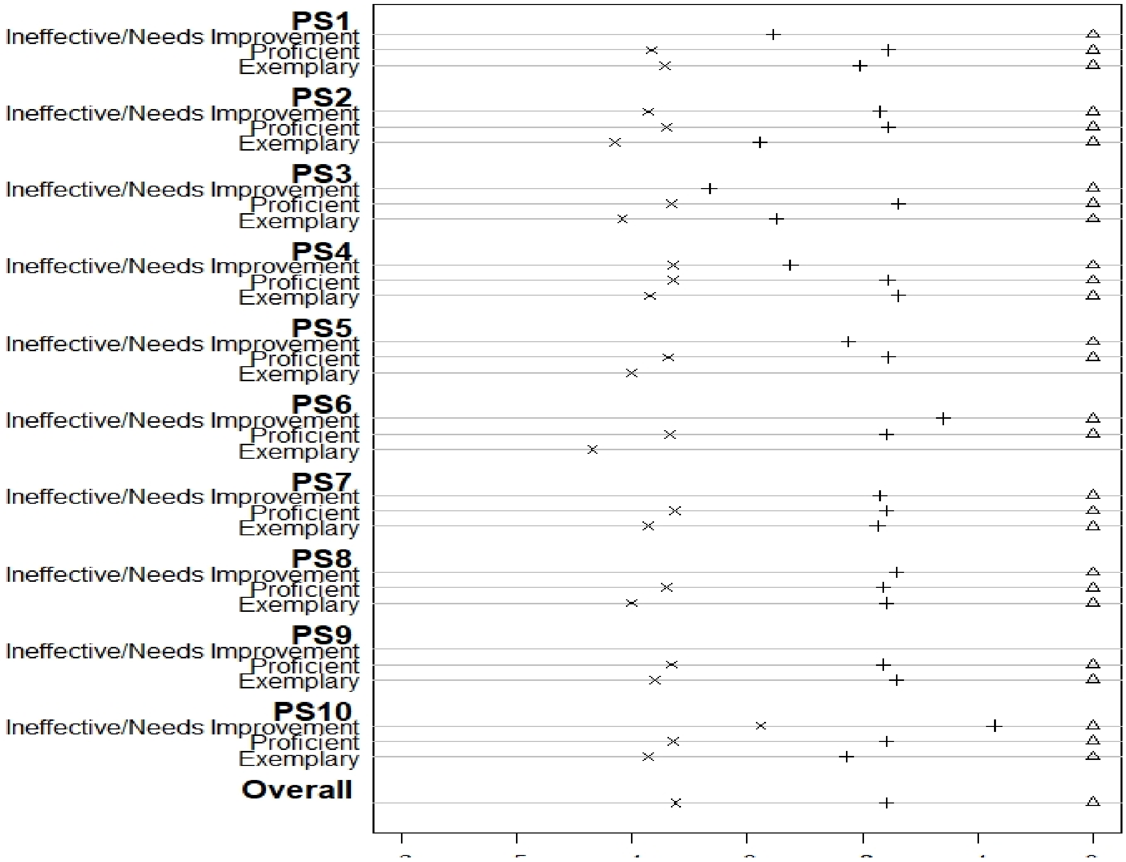
The first of four assumptions for ordinal logistic regression is that the dependent variable was binary. That assumption was met. Next, the observations must be independent of each other. The assumption of independent observations was also met. Next, no correlations were greater than or equal to .90, indicating extreme collinearity. The assumption of little or no multicollinearity was also met.

The final assumption which must be met is the proportional odds assumption. Binary logit regressions using each independent variable in the main ordinal logistic regression model were run as a robustness check for the proportional odds assumption underlying the ordinal logistic regression model. To construct the binary dependent variables, the value of the dependent variable was classified as if an SGP is greater than or equal to a given level and 0 otherwise. For example, an indicator variable for whether or not the math SGP level is greater than or equal to two was constructed. Then the logit was used to regress this binary variable on the TKES Standard 1 levels. The same was done for SGP levels being greater than or equal to 3 and 4, respectively. If the proportional odds assumption were precisely true, the coefficients in each regression would be the same. However, the coefficients are relatively different across each binary logistic regression. For example, in the example mentioned above, in the regression using the binary variable for SGP level greater than or equal to 2, the Standard 1 at the proficient and exemplary level coefficients are 0.96 and 2.22, respectively. However, in the regression using a binary variable for SGP levels greater than or equal to 4, these coefficients are 13.01 and 14.13, respectively.

A visual representation was used to examine the proportional odds assumption for each standard. In Figure 4, the plot visually allowed an understanding of the proportional odds assumption, plotting the results of each of these binary regressions that employ different cut-points of the dependent variable. If the proportional odds assumption were met precisely, each symbol, representing a coefficient from a binary logistic regression with a different cut point, would be in a straight line down the plot. There is some variance, especially for the lower SGP levels. Although caution should be used when interpreting the results, this assumption was met.

Figure 4

Proportional Odds Assumption: Fourth Grade Math



A likelihood ratio test to compare the fit of this ordinal regression model to an intercept-only model was used to understand the overall goodness-of-fit of the model, X^2(20, N = 1400) = 133.14, *p*<.001. The null hypothesis was accepted because the intercept-only model did not perform as well as the performed model. The full regression model was a better fit. R-squared values of r^2 = 9%, 11%, and 5% for the Cox-Snell, Nagelkerke, and McFadden pseudo R-squared methods were computed, respectively, indicating the variance accounted for between the TKES summative scores and teacher's SGP level in fourth grade math.

Table 15 indicates the results from the ordinal logistic regression analysis for fourth grade Math. Of the twenty predictor variables, three were found to be statistically significant. The ordinal logistic model indicated only Standard 1 at the exemplary level (z = 2.25, *p*<.05, Odds Ratio = 6.12, 95% CI [1.24, 30.68]), Standard 3 at the proficient level (z = 2.32, *p*<.05, Odds Ratio = 2.89, 95% CI [1.16, 7.08]), and Standard 3 at the exemplary level (z = 2.66, *p*<.05, Odds Ratio = 3.70, 95% CI [1.40, 9.64]) have positive and statistically significant associations with increased SGP levels at the 95% level of confidence. Standard 1 and Standard 3 at the exemplary level were found to have the greatest impact on student growth. Standard 1 (exemplary), Standard 3 (proficient), and Standard 3 (exemplary) are associated with 6.12, 2.89, and 3.70, times higher odds of obtaining a higher SGP level than receiving ineffective/needs eevelopment on each standard. The weakest predictors were Standard 2 at the ppoficient Level (z = -.05, *p*<.05, Odds Ratio = 1.00, 95% CI [0.38, 2.41]) and Standard 8 at the proficient level (z = -.16, *p*<.05, Odds Ratio = 0.93, 95% CI [0.40, 2.11]). Using the model to predict the in-sample SGP levels, approximately 70% of the SGP levels were classified correctly.

Table 15

Variables Used to Predict SGP's Utilizing Ordinal Logistic Regression in Fourth Grade Math

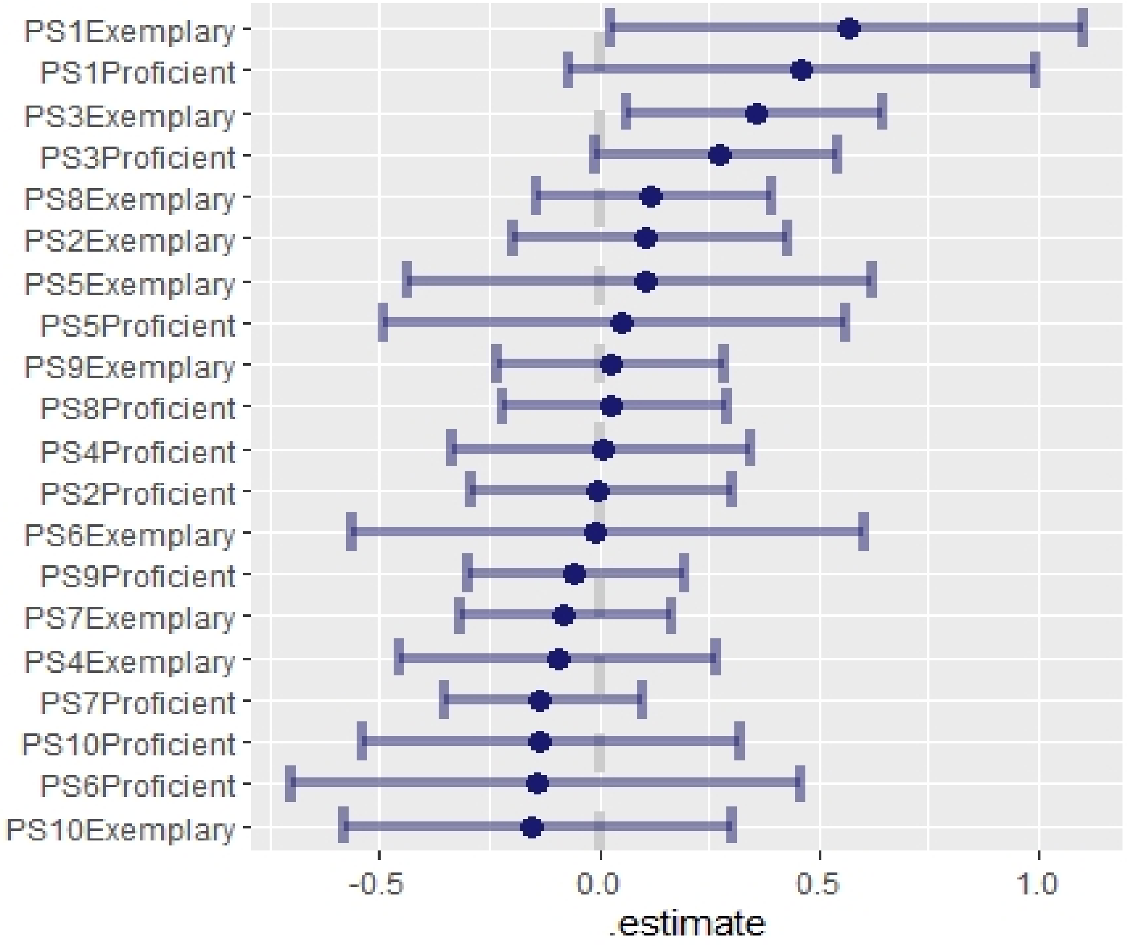
							95% Confidence Interval	
Predictor	Log Odds	*SE*	*Z*	Pr(>|z|)		OR	*Lower*	*Upper*
PS 1 Proficient	1.373	0.788	1.742	0.082		3.948	0.829	19.132
PS 1 Exemplary	1.811	0.806	2.247	0.025	*	6.119	1.241	30.680
PS 2 Proficient	-0.024	0.473	-0.050	0.960		0.997	0.376	2.414
PS 2 Exemplary	0.387	0.510	0.759	0.448		1.473	0.531	3.924
PS 3 Proficient	1.062	0.458	2.319	0.020	*	2.893	1.163	7.075
PS 3 Exemplary	1.308	0.491	2.666	0.008	**	3.698	1.399	9.642
PS 4 Proficient	0.038	0.505	0.076	0.939		1.039	0.376	2.736
PS 4 Exemplary	-0.310	0.544	-0.570	0.569		0.734	0.247	2.089
PS 5 Proficient	0.386	0.837	0.461	0.645		1.471	0.265	7.222
PS 5 Exemplary	0.567	0.875	0.648	0.517		1.763	0.297	9.368
PS 6 Proficient	-0.825	0.776	-1.064	0.287		0.438	0.098	1.967
PS 6 Exemplary	-0.347	0.811	-0.429	0.668		0.706	0.147	4.406
PS 7 Proficient	-0.371	0.530	-0.699	0.485		0.690	0.239	1.908
PS 7 Exemplary	-0.170	0.550	-0.309	0.757		0.844	0.282	2.428
PS 8 Proficient	-0.068	0.422	-0.161	0.872		0.934	0.403	2.106
PS 8 Exemplary	0.338	0.467	0.724	0.469		1.402	0.555	3.464
PS 9 Proficient	-0.110	0.467	-0.237	0.813		0.895	0.350	2.182
PS 9 Exemplary	0.241	0.483	0.499	0.617		1.273	0.483	3.208
PS 10 Proficient	-0.645	0.634	-1.016	0.309		0.525	0.152	1.784
PS 10 Exemplary	-0.744	0.658	-1.131	0.258		0.475	0.131	1.695

Note: p< 0.001 ‘***’, p<.01 ‘**’, p<.05 ‘*’, PS 1 Proficient (Professional Standard 1 Proficient Level), PS 1 Exemplary (Professional Standard 1 Exemplary Level), PS 2 Proficient (Professional Standard 2 Proficient Level), PS 2 Exemplary (Professional Standard 2 Exemplary Level), PS 3 Proficient (Professional Standard 3 Proficient Level), PS 3 Exemplary (Professional Standard 3 Exemplary Level), PS 4 Proficient (Professional Standard 4 Proficient Level), PS 4 Exemplary (Professional Standard 4 Exemplary Level), PS 5 Proficient (Professional Standard 5 Proficient Level), PS 5 Exemplary (Professional Standard 5 Exemplary Level), PS 6 Proficient (Professional Standard 6 Proficient Level), PS 6 Exemplary (Professional Standard 6 Exemplary Level), PS 7 Proficient (Professional Standard 7 Proficient Level), PS 7 Exemplary (Professional Standard 7 Exemplary Level), PS 8 Proficient (Professional Standard 8 Proficient Level), PS 8 Exemplary (Professional Standard 8 Exemplary Level), PS 9 Proficient (Professional Standard 9 Proficient Level), PS 9 Exemplary (Professional Standard 9 Exemplary Level), PS 10 Proficient (Professional Standard 10 Proficient Level), PS 10 Exemplary (Professional Standard 10 Exemplary Level).

Figure 5 provides a graphic representation of the bootstrapping sample for fourth grade math, again based on 1,000 replications of the data. Although Standard 1 at the exemplary level, Standard 3 at the proficient level, and Standard 3 at the exemplary level were found significant in the ordinal regression model, only Standard 1 and Standard 3 at the exemplary level were identified as significant predictors of student growth using bootstrap sampling. No standards were identified as having a greater chance of negatively impacting SGP levels. Due to the 1,000 replications of data with bootstrap samples, there is a high level of confidence that Standards 1 and 3 at the exemplary level have a significant impact on student achievement in fourth grade math.

Figure 5

Bootstrapping Sample: Fourth Grade Math



Fifth Grade ELA

c. Are summative scores on TKES Standards significant predictors of the teacher's SGP level (levels I, II, III, and IV) on the fifth grade English/Language Arts portion of the Georgia Milestones Assessment System (GMAS)?

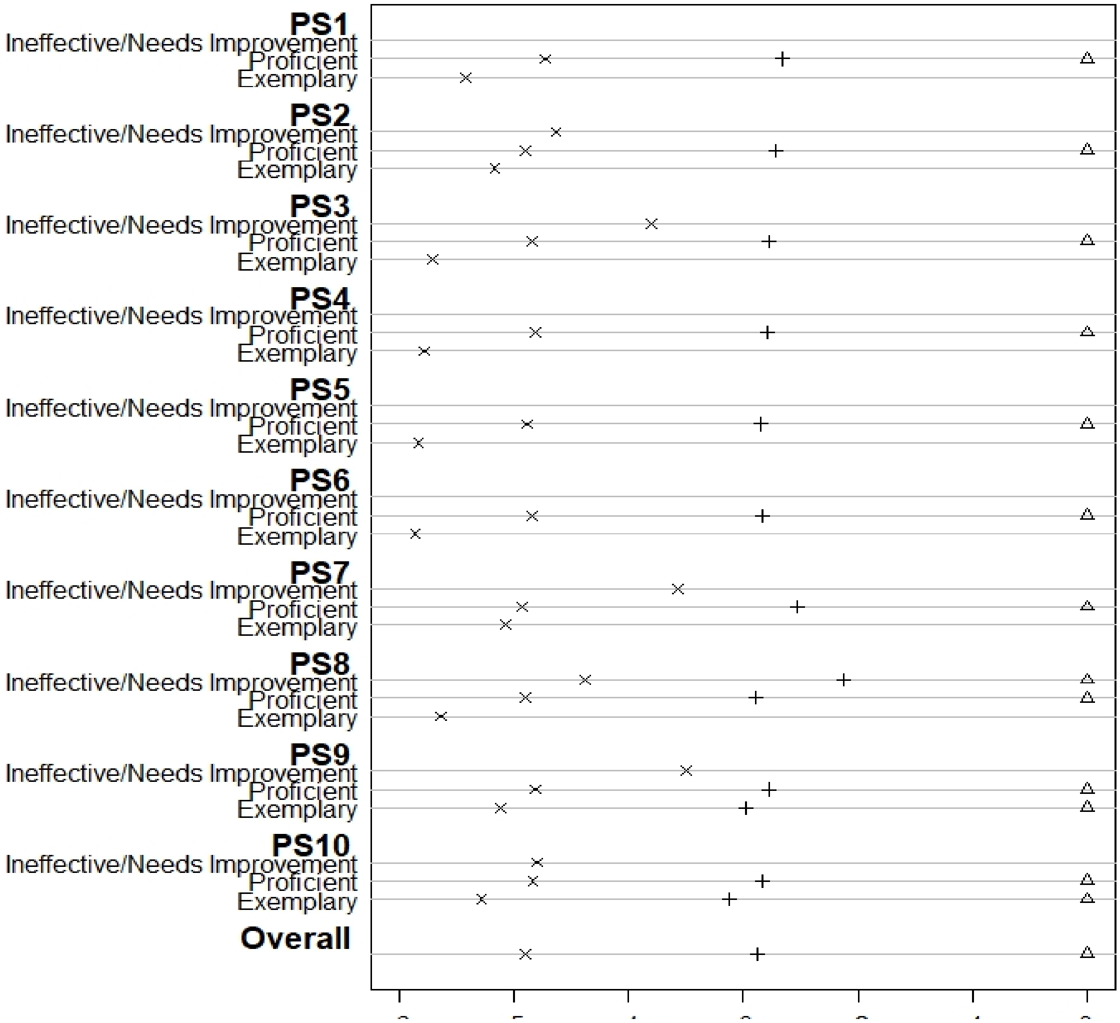
The four assumptions which must be met for ordinal logistic regression include the binary independent variable, independent observations, and no correlations were greater than or equal to .90, indicating extreme collinearity. The assumptions were met.

The final assumption is the proportional odds assumption. Binary logit regressions were run on this dataset using each independent variable in the primary ordinal logistic regression model. The value of the dependent variable was classified as 1 if an SGP was greater than or equal to the given value and 0 otherwise to construct the binary dependent variable. In the regression for SGP level > = 2, Standard 1 at the proficient and exemplary levels, coefficients are 0.96 and 2.22. Alternatively, in the regression using a binary variable for SGP levels > = 4, the coefficients are 13.01 and 14.13.

Figure 6 allows for a visual representation helpful in understanding the proportional odds assumption. As with the other datasets, there is some variance, especially for lower SGP levels. Although the proportional odds assumption was met, caution should be used when interpreting the coefficients.

Figure 6

Proportional Odds Assumption: Fifth Grade ELA



As with the fourth grade ELA and math datasets, an ordinal regression analysis was executed for the fifth grade ELA dataset. Unlike the fourth grade samples, no TKES Standards were found to have a statistically significant association with increased SGP levels. A likelihood ratio test was used to compare the fit of the intercept-only model and the full model, X^2(20, N = 1400) = 45.13, p<.001). There was a statistical difference between the models. The intercept-only model does not perform as well as the ordinal regression analysis. The values of r^2 = 9%, 11%, and 5% with the Cox-Snell, Nagelkerke, and McFadden pseudo-R-squared methods were calculated, indicating the variance accounted for between the TKES summative scores and teacher's SGP level.

Table 16 displays the results from the ordinal logistic regression analysis for fifth grade ELA. None of the predictor variables were found to have a statistical impact on SGP's. The table may also be used to predict the most influential and weakest predictors of SGP's in fifth grade ELA. The weakest predictors were Standard 10 at the proficient Level (z = -1.62, *p*<.05, Odds Ratio = 0.25, 95% CI [0.05, 1.36]) and Standard 10 at the exemplary level (z = -1.50, *p*<.05, Odds Ratio = 0.27, 95% CI [0.05, 1.49]). The strongest predictors of SGP's were Standard 9 at the proficient level (z = 1.81, *p*<.05, Odds Ratio = 3.25, 95% CI [0.84, 11.25]) and Standard 9 at the exemplary level (z = 1.67, *p*<.05, Odds Ratio = 3.06, 95% CI [0.77, 10.95]). Approximately 70% of the SGP levels may be classified correctly.

Table 16

Variables Used to Predict SGP's Utilizing Ordinal Logistic Regression in Fifth Grade ELA

						95% Confidence Interval	
Predictor	Log Odd	*SE*	*Z*	Pr(>|z|)	OR	*Lower*	*Upper*
PS 1 Proficient	0.319	1.196	0.267	0.790	1.376	0.119	13.121
PS 1 Exemplary	0.004	1.211	0.003	0.998	1.004	0.084	9.838
PS 2 Proficient	0.542	0.557	0.972	0.331	1.719	0.552	4.944
PS 2 Exemplary	0.559	0.603	0.927	0.354	1.750	0.517	5.538
PS 3 Proficient	-0.898	0.744	-1.207	0.228	0.407	0.094	1.700
PS 3 Exemplary	-0.299	0.775	-0.386	0.700	0.742	0.162	3.302
PS 4 Proficient	0.551	0.620	0.890	0.374	1.736	0.486	5.562
PS 4 Exemplary	0.432	0.669	0.646	0.518	1.541	0.395	5.483
PS 5 Proficient	-0.347	-0.347	-0.332	0.740	0.707	0.084	4.802
PS 5 Exemplary	-0.113	-0.113	-0.102	0.918	0.893	0.096	6.832
PS 6 Proficient	-0.046	-0.046	-0.065	0.948	0.955	0.229	3.544
PS 6 Exemplary	0.257	0.257	0.338	0.735	1.293	0.281	5.403
PS 7 Proficient	0.725	0.725	1.362	0.173	2.065	0.690	5.634
PS 7 Exemplary	0.896	0.896	1.605	0.109	2.450	0.782	7.060
PS 8 Proficient	-0.217	-0.217	-0.386	0.700	0.805	0.261	2.342
PS 8 Exemplary	0.300	0.300	0.492	0.622	1.350	0.401	4.346
PS 9 Proficient	1.178	1.178	1.805	0.071	3.247	0.843	11.249
PS 9 Exemplary	1.118	1.118	1.673	0.094	3.059	0.772	10.948
PS 10 Proficient	-1.374	-1.374	-1.616	0.106	0.253	0.051	1.356
PS 10 Exemplary	-1.311	-1.311	-1.502	0.133	0.269	0.052	1.489

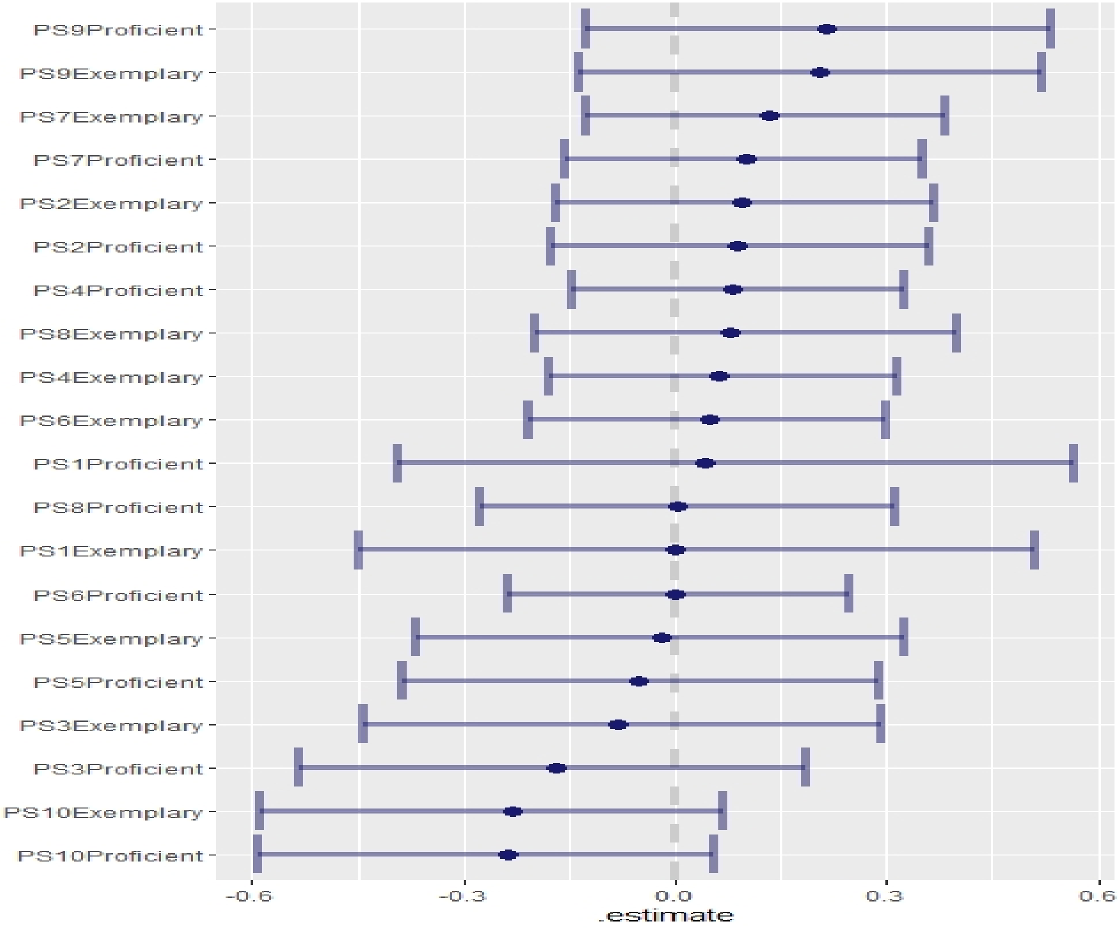
Note: p< 0.001 '***', p<.01 '**', p<.05 '*', PS 1 Proficient (Professional Standard 1 Proficient Level), PS 1 Exemplary (Professional Standard 1 Exemplary Level), PS 2 Proficient (Professional Standard 2 Proficient Level), PS 2 Exemplary (Professional Standard 2 Exemplary Level), PS 3 Proficient (Professional Standard 3 Proficient Level), PS 3 Exemplary (Professional Standard 3 Exemplary Level), PS 4 Proficient (Professional Standard 4 Proficient Level), PS 4 Exemplary (Professional Standard 4 Exemplary Level), PS 5 Proficient (Professional Standard 5 Proficient Level), PS 5 Exemplary (Professional Standard 5 Exemplary Level), PS 6 Proficient (Professional Standard 6 Proficient Level), PS 6 Exemplary (Professional Standard 6 Exemplary Level), PS 7 Proficient (Professional Standard 7 Proficient Level), PS 7 Exemplary (Professional Standard 7 Exemplary Level), PS 8 Proficient (Professional Standard 8 Proficient Level), PS 8 Exemplary (Professional Standard 8 Exemplary Level), PS 9 Proficient (Professional Standard 9 Proficient Level), PS 9 Exemplary (Professional Standard 9 Exemplary Level), PS 10 Proficient (Professional Standard 10 Proficient Level), PS 10 Exemplary (Professional Standard 10 Exemplary Level).

Figure 7 provides a graphic representation of the 1,000 bootstrapping samples. According to the ordinal regression model and bootstrap sampling, no levels of performance for any TKES

Standards have a greater chance of positively or negatively impacting SGP levels. Based on the number of bootstrap samples, the results are highly likely to be representative of the data.

Figure 7

Bootstrapping Sample: Fifth Grade ELA



Fifth Grade Math

d. Are summative scores on TKES Standards significant predictors of the teacher's SGP level (levels I, II, III, and IV) on the fifth grade Math portion of the Georgia Milestones Assessment System (GMAS)?

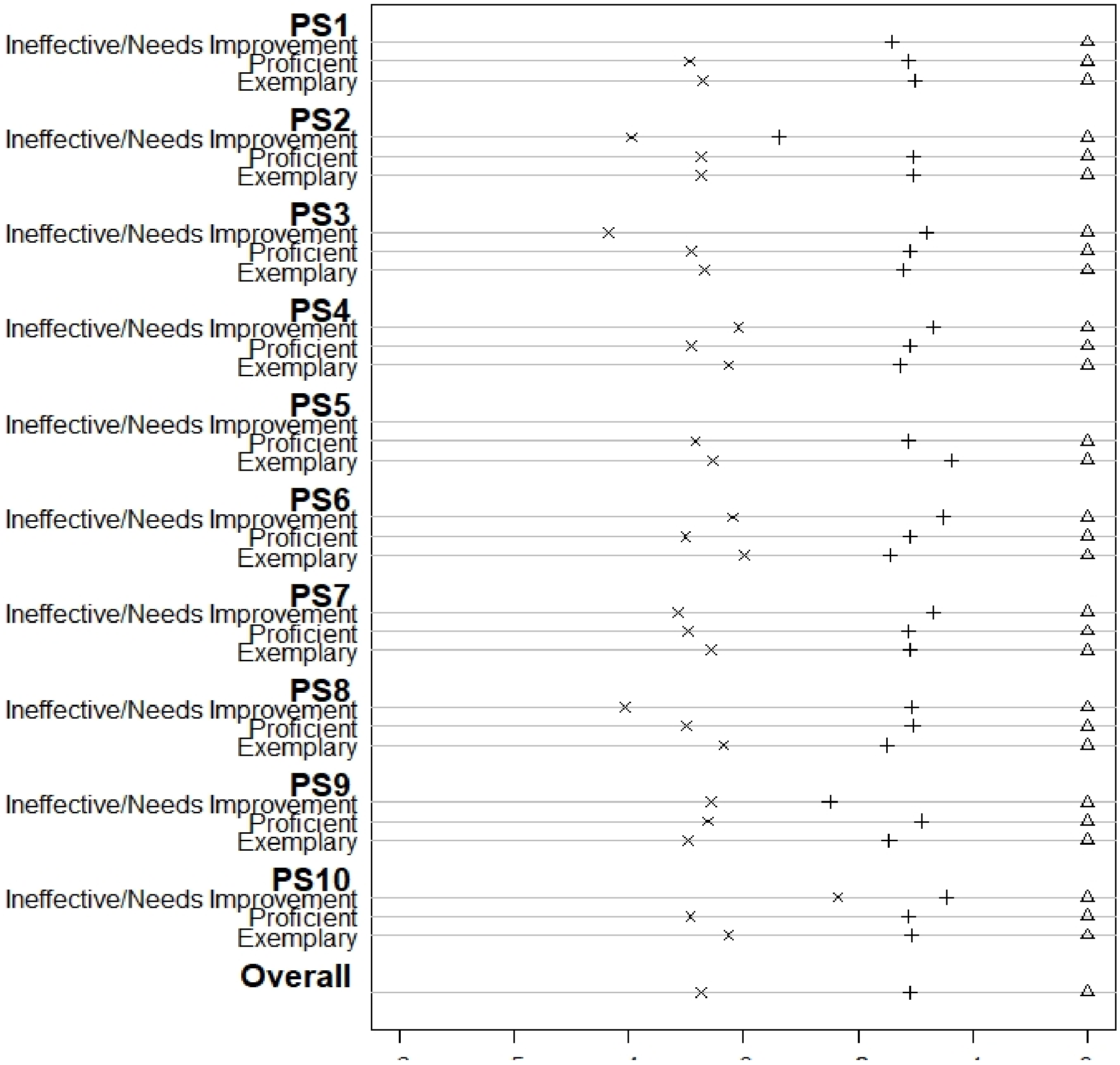
Assumptions for ordinal logistic regression were also met in this sample. The dependent variable was binary, and there were independent observations. All correlations were less than .90, indicating extreme collinearity.

Through the robustness check for the proportional odds assumption, relatively different coefficients were noted. For example, in the regression using the binary variable for SGP level > = 2, the coefficients for Standard 1 at the proficient and exemplary levels are -16.97 and 0. However, in the regression using a binary variable for SGP levels > = 4, the coefficients are 12.78 and 12.59, respectively.

Figure 8 provides a visual representation of the results using different cut-points of the dependent variable. As indicated below, the coefficients in each regression are not equal. There is no straight line down the plot, and some variance is indicated. The assumption was met for the proportional odds assumption.

Figure 8

Proportional Odds Assumption: Fifth Grade Math



To understand the overall goodness-of-fit of the model, the likelihood ratio test was used to compare the fit of this ordinal regression model to an intercept-only model, X^2(20, N = 1400) = 64.90, p<.001. The full ordinal regression model is a better fit. The full model provides more information than the intercept-only model. The Cox-Snell, Naglekerke, and McFadden pseudo-R-squared methods yielded the following: r^2 = 3%, 4%, and 3%. These values indicated the amount of variance accounted for between the TKES summative scores and teacher's SGP level in fifth grade math.

Table 17 displays the results from the ordinal logistic regression analysis for fifth grade math. No statistically significant SGP impact by TKES Standard was evident in the fifth grade math data sample. Although no TKES Standards were significantly significant, the weakest predictors were Standard 5 at the exemplary level (z = -.08, *p*<.05, Odds Ratio = 0.94, 95% CI [0.18, 4.49]) and Standard 9 at the exemplary level (z = -.14, *p*<.05, Odds Ratio = 0.93, 95% CI [0.32, 2.63]). The strongest predictors of SGP's were Standard 3 at the exemplary level (z = 1.74, *p*<.05, Odds Ratio = 2.39, 95% CI [0.88, 6.35]) and Standard 6 at the exemplary level (z = 1.60, *p*<.05, Odds Ratio = 2.50, 95% CI [0.80, 7.63]). Using the model to predict the in-sample SGP levels, approximately 83% of the SGP levels are classified correctly.

Table 17

Variables Used to Predict SGP's Utilizing Ordinal Logistic Regression in Fifth Grade Math

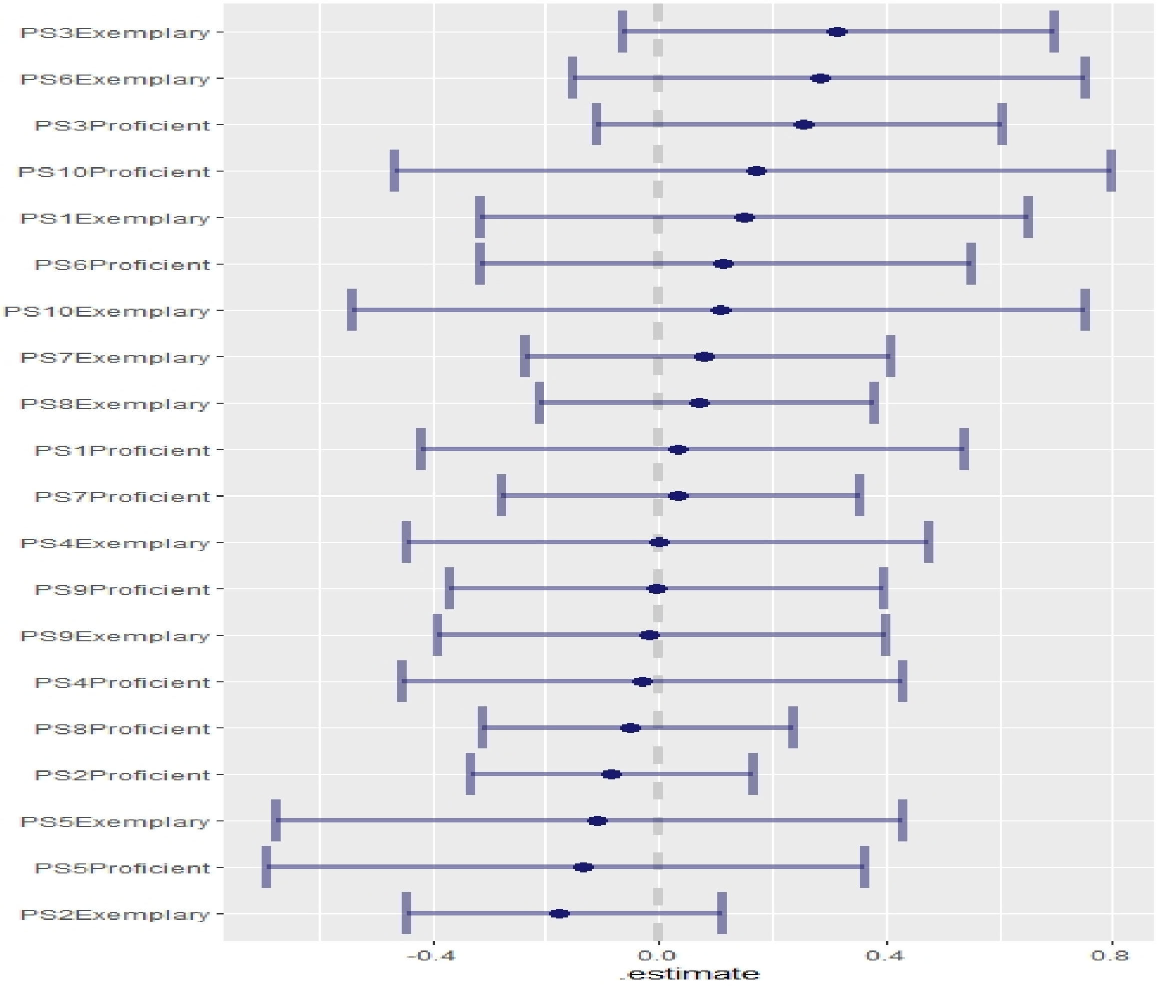
						95% Confidence Interval	
Predictor	Log Odds	*SE*	*Z*	Pr(>|z|)	OR	*Lower*	*Upper*
PS 1 Proficient	0.201	0.659	0.305	0.760	1.223	0.325	4.375
PS 1 Exemplary	0.605	0.677	0.894	0.371	1.832	0.471	5.799
PS 2 Proficient	-0.112	0.424	-0.264	0.792	0.894	0.386	2.032
PS 2 Exemplary	-0.461	0.462	-0.998	0.318	0.631	0.253	1.545
PS 3 Proficient	0.673	0.472	1.426	0.154	1.960	0.765	4.896
PS 3 Exemplary	0.873	0.502	1.738	0.082	2.394	0.882	6.353
PS 4 Proficient	-0.297	0.550	-0.540	0.589	0.743	0.253	2.170
PS 4 Exemplary	-0.190	0.584	-0.326	0.745	0.827	0.263	2.583
PS 5 Proficient	-0.208	0.782	-0.266	0.790	0.812	0.169	3.613
PS 5 Exemplary	-0.067	0.819	-0.082	0.934	0.935	0.181	4.491
PS 6 Proficient	0.315	0.530	0.594	0.552	1.370	0.477	3.811
PS 6 Exemplary	0.916	0.576	1.592	0.111	2.499	0.798	7.631
PS 7 Proficient	0.072	0.436	0.166	0.868	1.075	0.450	2.485
PS 7 Exemplary	0.234	0.454	0.516	0.606	1.264	0.511	3.036
PS 8 Proficient	-0.120	0.426	-0.282	0.778	0.887	0.383	2.026
PS 8 Exemplary	0.279	0.465	0.600	0.549	1.322	0.529	3.265
PS 9 Proficient	0.029	0.526	0.056	0.956	1.030	0.363	2.853
PS 9 Exemplary	-0.076	0.538	-0.141	0.888	0.927	0.320	2.630
PS 10 Proficient	0.415	0.706	0.588	0.557	1.515	0.368	5.896
PS 10 Exemplary	0.252	0.723	0.349	0.727	1.287	0.303	5.189

Note: p< 0.001 '***', p<.01 '**', p<.05 '*', PS 1 Proficient (Professional Standard 1 Proficient Level), PS 1 Exemplary (Professional Standard 1 Exemplary Level), PS 2 Proficient (Professional Standard 2 Proficient Level), PS 2 Exemplary (Professional Standard 2 Exemplary Level), PS 3 Proficient (Professional Standard 3 Proficient Level), PS 3 Exemplary (Professional Standard 3 Exemplary Level), PS 4 Proficient (Professional Standard 4 Proficient Level), PS 4 Exemplary (Professional Standard 4 Exemplary Level), PS 5 Proficient (Professional Standard 5 Proficient Level), PS 5 Exemplary (Professional Standard 5 Exemplary Level), PS 6 Proficient (Professional Standard 6 Proficient Level), PS 6 Exemplary (Professional Standard 6 Exemplary Level), PS 7 Proficient (Professional Standard 7 Proficient Level), PS 7 Exemplary (Professional Standard 7 Exemplary Level), PS 8 Proficient (Professional Standard 8 Proficient Level), PS 8 Exemplary (Professional Standard 8 Exemplary Level), PS 9 Proficient (Professional Standard 9 Proficient Level), PS 9 Exemplary (Professional Standard 9 Exemplary Level), PS 10 Proficient (Professional Standard 10 Proficient Level), PS 10 Exemplary (Professional Standard 10 Exemplary Level).

Figure 9 provides a graphic representation of the bootstrapping samples. Again, in agreement with the ordinal regression analysis, the bootstrap sampling indicated no levels of performance for any TKES Standard have a greater chance of positively or negatively impacting SGP levels. Results through the robust bootstrap method allow for a high level of confidence in the results indicated in Figure 9.

Figure 9

Bootstrapping Sample: Fifth Grade Math



Research Question 2

Are summative scores on TKES standards (Professional Knowledge, Instructional Planning, Instructional Strategies, Differentiated Instruction, Assessment Strategies, Assessment Uses, Positive Learning Environment, Academically Challenging Environment, Professionalism, and Communication) significant predictors of the teacher's mean scaled score on the Georgia Milestones Assessment System (GMAS)?

Fourth Grade ELA

a. Are summative scores on TKES Standards significant predictors of the teacher's mean scaled score on the fourth grade English/Language Arts portion of the Georgia Milestones Assessment System (GMAS)?

The goal of this sub-question was to determine if summative scores on TKES Standards are significant predictors of the teachers' mean scale score on the fourth grade English Language Arts portion of the Georgia Milestones Assessment System. The independent variable was transformed to have three levels for each standard, including level 2 (needs development), level 3 (proficient), and level 4 (distinguished). Level 3 was made the reference level, while levels 2 and 4 values are presented in the summary estimates table for the fitted model. The dependent variable, or mean scale score, was on the interval level of measurement. The teacher's mean scale score was calculated by averaging the student's mean scale scores in their classes. The statistical method adopted for this question was multiple regression analysis.

Statistical considerations were missing data and outliers. There was no missing data in the dataset. There were five values identified as potential outliers. *Z*-score values fell between -3.35 and 3.20. Skew (.21) and kurtosis (.00) values were calculated. TKES Scores were treated as ordinal level variables, and GMAS MSS values were treated as interval level variables.

Assumptions must be met for multiple regression before completing the analysis. Initially, the assumption of normality was not met. The normality assumption was checked using the Shapiro-Wilk test on the raw data. The data were not found to be normally distributed (W(1398) = 1.00, *p* = .003). Results from the Jarque Bera Test on the raw data also indicated a lack of normality (χ^2 (1398) = 10.097, (*p* = .006). Hence, there was a need for data transformation. The dataset was transformed using Box-Cox transformation. The distribution of the resulting variable was checked using the Shapiro-Wilk test as conducted with the raw data. The results indicated that the data were normally distributed (W(1398) = 1.00, *p* = .72). Additionally, the Jarque Bera Test indicated normality on the transformed data *(*χ^2 (1398) = 0.47, *p* = .80). Because the data transformation helped achieve a normal distribution for the dependent variable, the transformed variable was used for the analysis.

Following the data transformation, outliers were readdressed. All z-score values fell between -3.80 and 2.92. Only one potential outlier was identified. Skew (.00) and kurtosis (-.05) values were reconsidered. Using Cook's distance test, no cases were identified that might influence the outcome. Additionally, the data met the assumption of independent errors (DW = 1.98, p = .47).

The assumption of homoscedasticity of residuals was also met, meaning the variation of residuals was constant across the model. As shown in Figure 10 and Figure 11, the residuals remain approximately equal. Additionally, the Breusch-Pagan test was not significant, indicating the null was accepted that error variances were equal (BP(1398) = 20.91, *p* = .40). It was concluded the assumption of homoscedasticity was met.

Figure 10

Studentized Residuals for Fourth Grade ELA

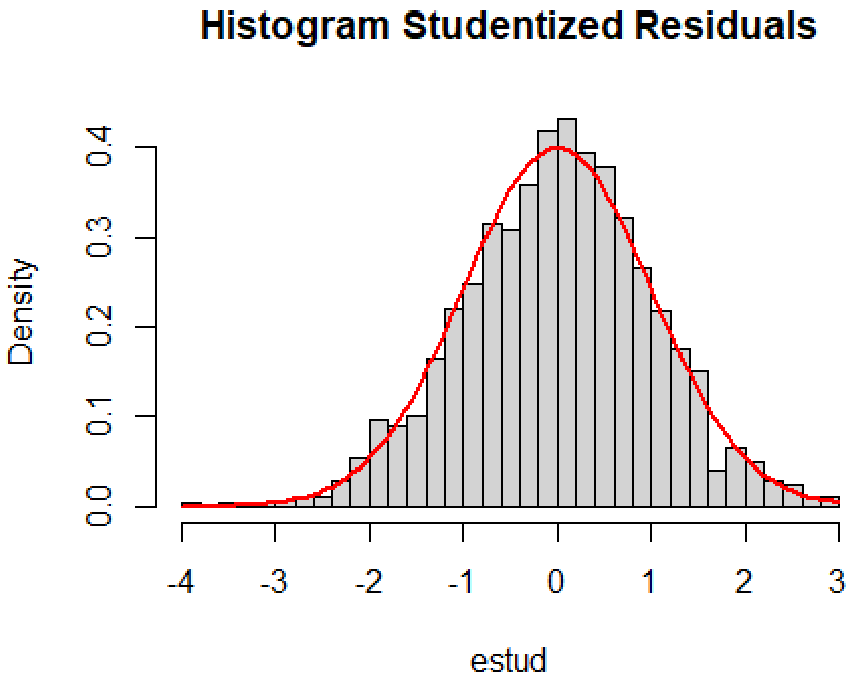
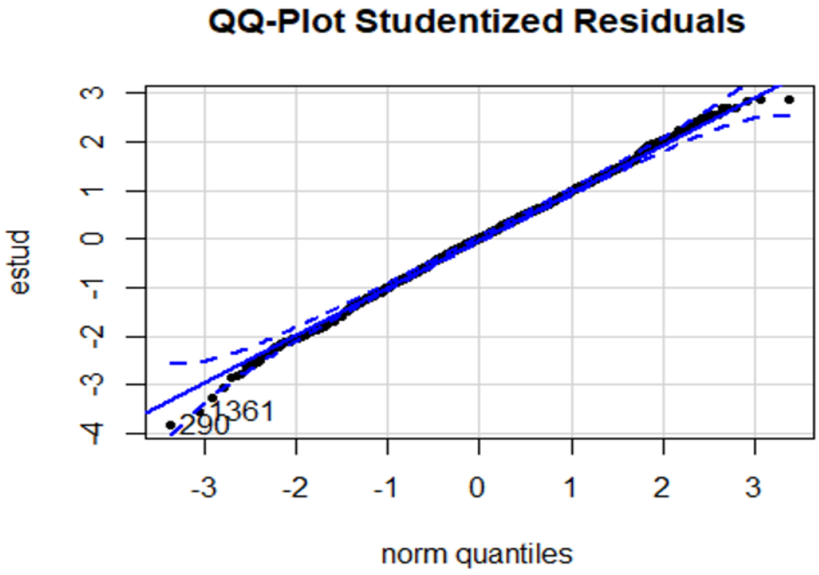


Figure 11

Q-Q Normality Plot of Studentized Residuals for Fourth Grade ELA



The assumption of multicollinearity assumes variables are not highly correlated. This assumption was first tested using VIF factors. VIF factors with values close to 1.0 indicate a low correlation among the predictors. The values in the dataset fell between 1.28 and 1.65. In Table 18, polychoric and polyserial correlations were provided. No values exceeded .90, indicating no multicollinearity.

As noted, Standard 1 and 3 ($r(1398) = .70$, $p < .05$), Standard 3 and 4, ($r(1398) = .70$, $p < .05$), Standard 3 and 8 ($r(1398) = .76$, $p < .05$), Standard 5 and 6 ($r(1398) = .76$, $p < .05$), and Standard 9 and 10 ($r(1398) = .73$ $p < .05$) had a moderate to strong polychoric correlation. The weakest correlations were found between Standard 5 and 9 ($r(1398) = .46$, $p < .05$) and Standard 8 and 9 ($r(1398) = .46$, $p < .05$) which had weak to moderate correlations. All polyserial correlations fell below .40. Standard 8 had a weak to moderate correlation to the MSS ($r(1398) = .40$, $p < .05$). Standard 6 had the weakest correlation to the MSS ($r(1398) = .22$, $p < .05$).

Table 18

Polychoric and Polyserial Correlations among Mean Scale Score (MSS) and TKES Standards in Fourth Grade ELA

	MSS	PS1	PS2	PS3	PS4	PS5	PS6	PS7	PS8	PS9	PS10
MSS											
PS 1	.29										
PS 2	.30	.61									
PS 3	.31	.70	.64								
PS 4	.25	.57	.62	.70							
PS 5	.25	.60	.54	.54	.52						
PS 6	.22	.55	.61	.56	.65	.76					
PS 7	.33	.61	.63	.68	.66	.46	.49				
PS 8	.40	.66	.65	.76	.65	.60	.60	.64			
PS 9	.24	.55	.53	.49	.48	.46	.52	.60	.46		
PS 10	.29	.56	.53	.53	.52	.51	.52	.66	.52	.73	

Note. MSS (Mean Scale Score), PS 1 (Professional Standard 1), PS 2 (Professional Standard 2), PS 3 (Professional Standard 3), PS 4 (Professional Standard 4), PS 5 (Professional Standard 5), PS 6 (Professional Standard 6), PS 7 (Professional Standard 7), PS 8 (Professional Standard 8), PS 9 (Professional Standard 9), PS 10 (Professional Standard 10).

Table 19 shows the results from the regression analysis. Although the model was considered significant, the R^2_{adj} indicated that only 11.9% of the variance was explained by the independent variables, $R = .363$. $R^2 = .132$, $R^2_{adj} = .119$, F(20, 1379) = 10.49, $p < .001$. Out of the 10 standards, 3 standards were significant, Standard 3, Standard 7, and Standard 8. Standard 3 at the exemplary level (B = .067, $t = 2.08$, $p < .04$), Standard 7 at the exemplary level (B = .107, $t = 3.40$, $p < .001$), and Standard 8 at the exemplary level (B = .145, $t = 4.60$, $p < .001$) are expected to have a positive effect on scores. However, Standard 8 at the needs development level (B = -.084, $t = -2.89$, $p < .01$) is expected to have a negative impact on scores.

Table 19

Coefficient Table for Fourth Grade ELA Mean Scale Score (MSS)

	B	*SE of B*	*β*	*t*	*p*
PS 1 ND	-0.002	0.004216	-0.015	-0.50	.61
PS 1 Exemplary	0.001	0.000880	0.025	0.80	.43
PS 2 ND	-0.001	0.002427	-0.017	-0.59	.55
PS 2 Exemplary	0.001	0.000972	0.044	1.44	.15
PS 3 ND	0.003	0.002824	0.040	1.24	.23
PS 3 Exemplary	0.002	0.000921	0.067	2.08	.04
PS 4 ND	-0.005	0.003122	-0.049	-1.61	.11
PS 4 Exemplary	-0.001	0.001097	-0.032	-1.08	.28
PS 5 ND	0.005	0.004587	0.030	1.07	.28
PS 5 Exemplary	0.001	0.001400	0.031	1.06	.29
PS 6 ND	-0.004	0.003339	-0.035	-1.21	.23
PS 6 Exemplary	-0.002	0.001257	-0.040	-1.32	.19
PS 7 ND	-0.003	0.003049	-0.029	-0.94	.37
PS 7 Exemplary	0.003	0.000794	0.107	3.40	<.001
PS 8 ND	-0.006	0.002066	-0.084	-2.89	<.01
PS 8 Exemplary	0.005	0.001065	0.145	4.60	<.001
PS 9 ND	0.002	0.002651	0.025	0.93	.35
PS 9 Exemplary	0.001	0.000752	0.037	1.24	.22
PS 10 ND	-0.003	0.003403	-0.025	-0.87	.40
PS 10 Exemplary	0.002	0.000924	0.058	1.93	.05

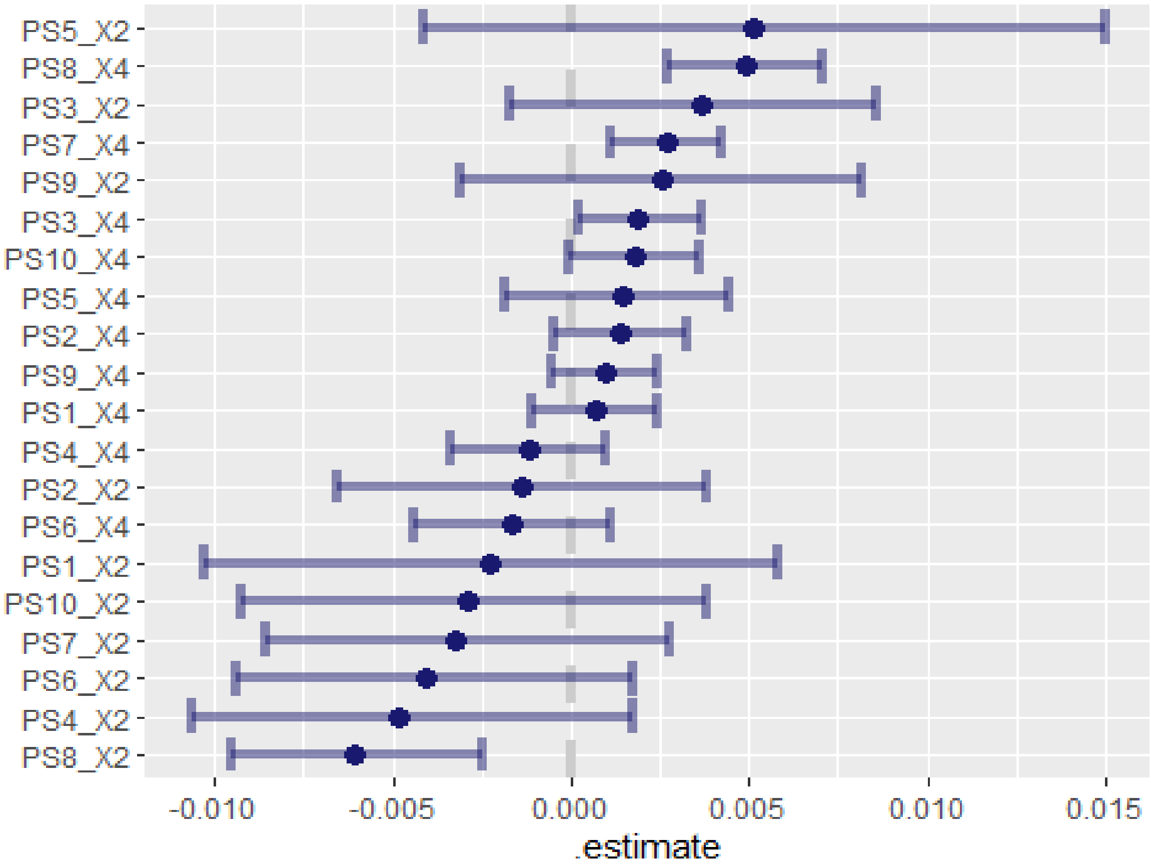
Note. PS 1 ND (Professional Standard 1 Needs Development Level), PS 1 Exemplary (Professional Standard 1 Exemplary Level), PS 2 ND (Professional Standard 2 Needs Development Level), PS 2 Exemplary (Professional Standard 2 Exemplary Level), PS 3 ND (Professional Standard 3 Needs Development Level), PS 3 Exemplary (Professional Standard 3

Exemplary Level), PS 4 ND (Professional Standard 4 Needs Development Level), PS 4 Exemplary (Professional Standard 4 Exemplary Level), PS 5 ND (Professional Standard 5 Needs Development Level), PS 5 Exemplary (Professional Standard 5 Exemplary Level), PS 6 ND (Professional Standard 6 Needs Development Level), PS 6 Exemplary (Professional Standard 6 Exemplary Level), PS 7 ND (Professional Standard 7 Needs Development Level), PS 7 Exemplary (Professional Standard 7 Exemplary Level), PS 8 ND (Professional Standard 8 Needs Development Level), PS 8 Exemplary (Professional Standard 8 Exemplary Level), PS 9 ND (Professional Standard 9 Needs Development Level), PS 9 Exemplary (Professional Standard 9 Exemplary Level), PS 10 ND (Professional Standard 10 Needs Development Level), PS 10 Exemplary (Professional Standard 10 Exemplary Level).

Figure 12 provides a graphic representation of the 1,000 bootstrapping samples of the dataset, rather than a single run of the data as represented in Table 19. The data replication provided a more accurate and robust representation of the output. Because of the increased representation of the data, there is a higher level of confidence in the outcome than the multiple regression model provided. Because the bootstrapping sample confirmed the standards found significant using the one-sample multiple regression, Standard 3 at the exemplary level, Standard 7 at the exemplary level, Standard 8 at the ineffective/needs development and exemplary levels, there is a high level of confidence that these standards have a significant impact on scores. Through the bootstrapping process, Standard 10 at the exemplary level was also found to impact scores positively.

Figure 12

Bootstrapping Sample: Fourth Grade ELA



Fourth Grade Math

b. Are summative scores on TKES Standards significant predictors of the teacher's mean scaled score on the fourth grade Math portion of the Georgia Milestones Assessment System (GMAS)?

The goal of this sub-question was to determine if summative scores on TKES Standards are significant predictors of the teachers' mean scale score on the fourth grade Math portion of the Georgia Milestones Assessment System. Statistical considerations were missing data and outliers. Again, the independent variable was transformed to have three levels, including level 2 (needs development), level 3 (proficient), and level 4 (distinguished). Level 3 was the reference level, while the summary estimates table included levels 2 and 4. Multiple regression was used for this analysis.

There was no missing data in the dataset. There were eight values identified as potential outliers. Z-score values fell between -2.46 and 3.74. Skew (.51) and kurtosis (.23) values were calculated.

Assumptions for multiple regression were examined before completing the analysis. As with the previous dataset, the assumption of normality was not met. The normality assumption was checked using the Shapiro-Wilk test on the raw data. The data were not found to be normally distributed (W(1398) = 0.98, $p < .0001$). The Jarque Bera Test results on the raw data indicated a lack of normality (χ^2 (1398) = 62.86, $p < .0001$). Again, the dataset was transformed using Box-Cox transformation. The Shapiro-Wilk results indicated that the transformed data were normally distributed (W(1398)=1.00, $p = .24$). The Jarque Bera Test also indicated normality on the transformed data (χ^2 (1398) = 3.70, $p = .16$). As seen in Figure 13, the histogram for the transformed data indicates general normality. Because the data transformation helped achieve a normal distribution, the transformed variable was used for the analysis.

Outliers were then readdressed. Z-score values fell between -3.13 and 2.90. Only two potential outliers were identified. Skew (.01) and kurtosis (-.24) values were reconsidered. No cases were identified by Cook's distance test that might influence the outcome. The data met the assumption of independent errors (DW = 1.91, $p = .10$).

The normality plot was examined to determine if residuals were constant. As shown in Figure 13 and Figure 14, the residuals remained approximately equal. The Breusch-Pagan test was not significant, and it was determined that error variances were equal (BP(1398) = 20.26, $p = .44$). It was concluded the assumption of homoscedasticity was met.

Figure 13

Studentized Residuals for Fourth Grade Math

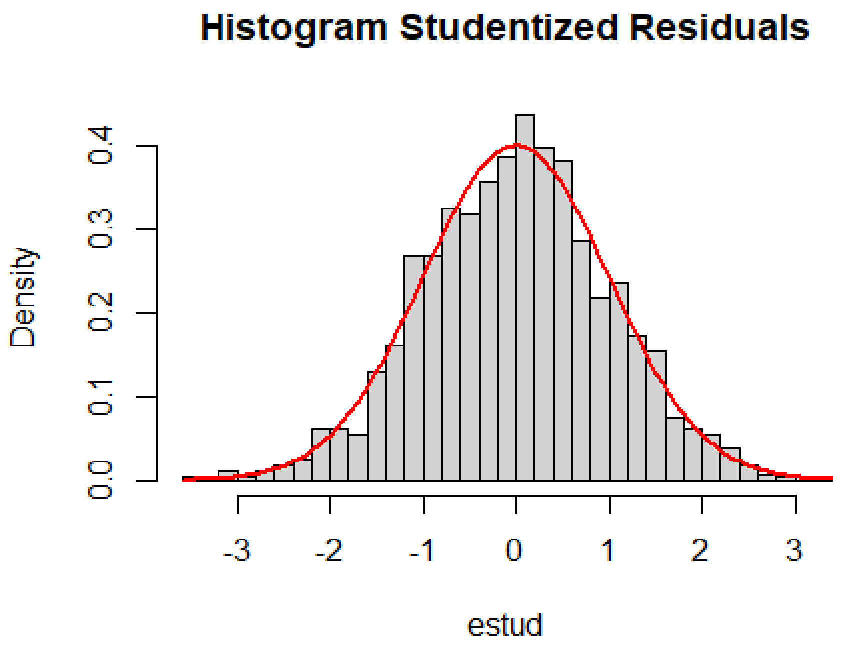
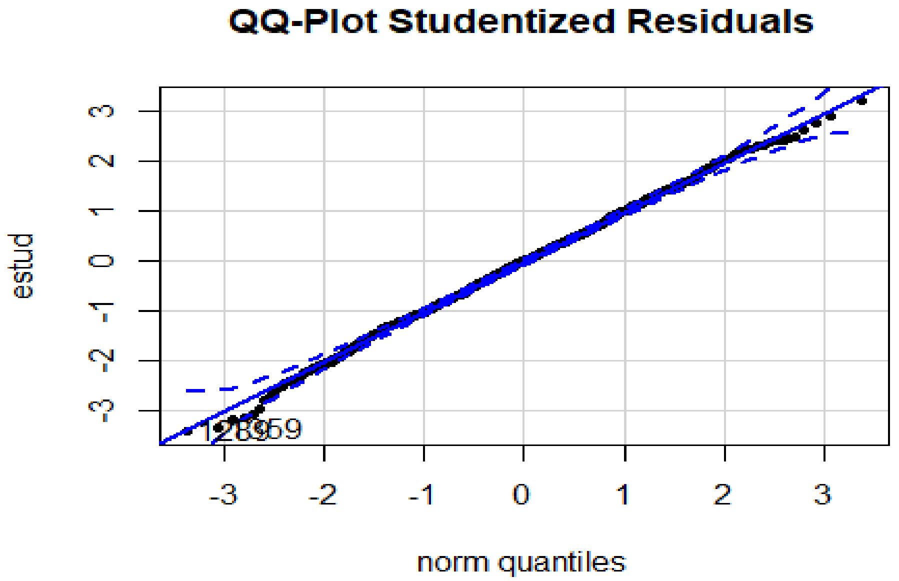


Figure 14

Q-Q Normality Plot of Studentized Residuals for Fourth Grade Math



Multicollinearity was first tested using VIF factors. The values in the dataset fell between 1.31 and 1.96, indicating a low correlation. Table 20 shows the polychoric and polyserial correlations, and no values exceeded .90, again suggesting no multicollinearity.

As noted, Standard 5 and 6 (*r*(1398) = .55, *p* < .05) and Standard 3 and 8, (*r*(1398) = .55, *p* < .05) had a weak to moderate polychoric correlation. The weakest correlation was found between Standard 4 and 9 (*r*(1398) = .29, *p* < .05), Standard 5 and 9 (*r*(1398) = .28, *p* < .05), Standard 6 and 7 (*r*(1398) = .28, *p* < .05), Standard 6 and 9 (*r*(1398) = .28, *p* < .05), Standard 6 and 10 (*r*(1398) = .26, *p* < .05), and Standard 8 and 10 (*r*(1398) = .29, *p* < .05) which had weak correlations. All polyserial correlations fell below .28. Standard 8 had a weak correlation to the MSS (*r*(1398) = .28, *p* < .05). Standard 6 had the weakest correlation to the MSS (*r*(1398) = .18, *p* < .05).

Table 20

Polychoric and Polyserial Correlations among Mean Scale Score (MSS) and TKES Standards in Fourth Grade ELA

	MSS	PS1	PS2	PS3	PS4	PS5	PS6	PS7	PS8	PS9	PS10
MSS											
PS 1	.27										
PS 2	.27	.40									
PS 3	.23	.45	.42								
PS 4	.21	.37	.38	.43							
PS 5	.20	.34	.37	.31	.34						
PS 6	.18	.35	.39	.32	.40	.55					
PS 7	.24	.40	.41	.47	.37	.26	.28				
PS 8	.28	.44	.45	.53	.45	.37	.32	.41			
PS 9	.22	.38	.35	.30	.29	.28	.28	.40	.27		
PS 10	.20	.37	.36	.33	.32	.32	.26	.43	.29	.44	

Note. MSS (Mean Scale Score), PS 1 (Professional Standard 1), PS 2 (Professional Standard 2), PS 3 (Professional Standard 3), PS 4 (Professional Standard 4), PS 5 (Professional Standard 5), PS 6 (Professional Standard 6), PS 7 (Professional Standard 7), PS 8 (Professional Standard 8), PS 9 (Professional Standard 9), PS 10 (Professional Standard 10).

The regression analysis results are noted in Table 21. Although the model was considered significant, the $R^2{}_{adj}$ indicated that only 12.1% of the variance was explained by the independent variables, $R = .365$. $R^2 = .133$, $R^2_{adj} = .121$, F(20, 1379) = 10.6, $p < .001$. Out of the 10 standards, 3 standards were significant, Standard 1, Standard 2, and Standard 8. Standard 1 at the exemplary level (B = .085, $t = 2.74$, $p < .01$), Standard 2 at the exemplary level (B = .090, $t = 2.90$, $p < .01$), and Standard 8 at the exemplary level (B = .092, $t = 2.87$, $p < .01$) are expected to have a positive effect on scores.

Table 21

Coefficient Table for Fourth Grade Math Mean Scale Score (MSS)

	B	*SE of B*	*β*	*t*	*p*
PS1 ND	-3.87e-08	1.99e-08	-0.064	-1.95	.05
PS1 Exemplary	9.24e-09	3.38e-09	0.085	2.74	.01
PS2 ND	-5.44e-09	1.04e-08	-0.015	-0.53	.60
PS2 Exemplary	1.09e-08	3.77e-09	0.090	2.90	<.01
PS3 ND	7.58e-09	1.11e-08	0.022	0.68	.49
PS3 Exemplary	3.93e-09	3.54e-09	0.036	1.11	.27
PS4 ND	-10.00e-09	1.11e-08	-0.027	-0.90	.37
PS4 Exemplary	4.43e-11	4.20e-09	<0.000	0.01	.99
PS5 ND	2.78e-08	1.92e-08	0.051	1.45	.15
PS5 Exemplary	9.25e-09	5.35e-09	0.054	1.73	.08
PS6 ND	-9.21e-09	1.60e-08	-0.018	-0.58	.57
PS6 Exemplary	-1.50e-09	4.83e-09	-0.010	-0.31	.76
PS7 ND	-1.85e-08	1.17e-08	-0.046	-1.59	.11
PS7 Exemplary	5.94e-09	3.12e-09	0.060	1.91	.06
PS8 ND	-1.76e-08	9.34e-09	-0.058	-1.89	.06
PS8 Exemplary	1.16e-08	4.04e-09	0.092	2.87	<.01
PS9 ND	-1.04e-08	1.01e-08	-0.027	-1.03	.30
PS9 Exemplary	4.47e-09	2.94e-09	0.045	1.52	.13
PS10 ND	1.43e-08	1.27e-08	0.030	1.12	.26
PS10 Exemplary	3.21e-09	3.61e-09	0.027	0.89	.37

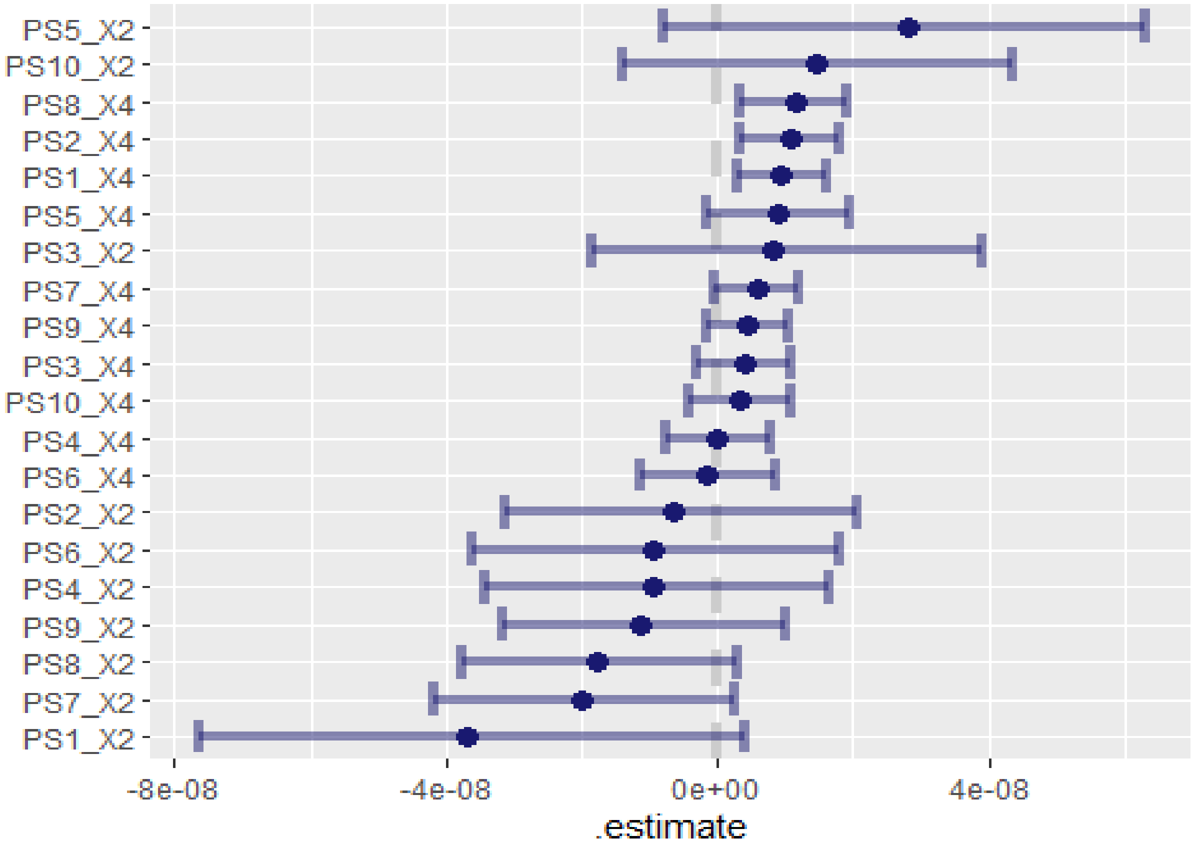
Note. PS 1 ND (Professional Standard 1 Needs Development Level), PS 1 Exemplary (Professional Standard 1 Exemplary Level), PS 2 ND (Professional Standard 2 Needs Development Level), PS 2 Exemplary (Professional Standard 2 Exemplary Level), PS 3 ND (Professional Standard 3 Needs Development Level), PS 3 Exemplary (Professional Standard 3

Exemplary Level), PS 4 ND (Professional Standard 4 Needs Development Level), PS 4 Exemplary (Professional Standard 4 Exemplary Level), PS 5 ND (Professional Standard 5 Needs Development Level), PS 5 Exemplary (Professional Standard 5 Exemplary Level), PS 6 ND (Professional Standard 6 Needs Development Level), PS 6 Exemplary (Professional Standard 6 Exemplary Level), PS 7 ND (Professional Standard 7 Needs Development Level), PS 7 Exemplary (Professional Standard 7 Exemplary Level), PS 8 ND (Professional Standard 8 Needs Development Level), PS 8 Exemplary (Professional Standard 8 Exemplary Level), PS 9 ND (Professional Standard 9 Needs Development Level), PS 9 Exemplary (Professional Standard 9 Exemplary Level), PS 10 ND (Professional Standard 10 Needs Development Level), PS 10 Exemplary (Professional Standard 10 Exemplary Level).

Seen below, in Figure 15, represents 1,000 bootstrapping samples of the dataset. Because of the increased representation of the data, there is a high level of confidence in the outcome. The bootstrapping sample confirmed the standards found significant using the one-sample multiple regression, Standard 1 at the exemplary level, Standard 2 at the exemplary level, and Standard 8 at the exemplary levels. Through bootstrapping, Standard 7 at the exemplary level was also found to impact scores positively.

Figure 15

Bootstrapping Sample: Fourth Grade Math



Fifth Grade ELA

c. Are summative scores on TKES Standards significant predictors of the teacher's mean scaled score on the fifth grade English/Language Arts portion of the Georgia Milestones Assessment System (GMAS)?

Multiple regression analysis was used to determine if summative scores on TKES Standards are significant predictors of the teachers' mean scale score on the fifth grade English Language Arts portion of the Georgia Milestones Assessment System. As with the previous sub-questions for RQ2, the independent variable was transformed to have three levels for each standard, including level 2 (needs development), level 3 (proficient), and level 4 (distinguished).

Level 3 was made the reference level. Levels 2 and 4 values are presented in the summary estimates table for the fitted model.

Statistical considerations were missing data and outliers. There was no missing data in the dataset. There were 5 values identified as potential outliers. Z-score values fell between -3.01 and 3.45. Skew (.32) and kurtosis (.07) values were calculated. TKES Scores were treated as ordinal level variables, and GMAS MSS values were treated as interval level variables.

Assumptions must be met before completing the analysis. The assumption of normality was not met. Using the Shapiro-Wilk test on the raw data, the data were not found to be normally distributed (W(1398) = 0.99, $p < .0001$). The Jarque Bera Test on the raw data also indicated a lack of normality (χ^2 (1398) = 24.24, $p < .0001$). The dataset was transformed using Box-Cox transformation, and the normality was analyzed using the Shapiro-Wilk. The transformed data were normally distributed (W(1398)=1.00, $p = .31$). The Jarque Bera Test also indicated normality on the transformed data (χ^2 (1398) = .01, $p = .99$). The histogram in Figure 16 indicates general normality for the transformed data. The transformed variable was used for the analysis because of its more normal distribution than the raw dataset.

Outliers were readdressed following the transformation. All z-score values fell between -3.56 and 2.98. Only 2 potential outliers were identified. Skew (.00) and kurtosis (-.01) values were reconsidered. No cases were identified by Cook's distance test that might influence the outcome. The data met the assumption of independent errors (DW = 2.00, $p = .94$).

The assumption of homoscedasticity of residuals was also examined. As shown in Figure 16 and Figure 17, the residuals remained approximately equal. The Breusch-Pagan test was not significant, and the null was accepted that error variances were equal (BP(1398) = 15.34, $p = .76$). It was concluded the assumption of homoscedasticity was met.

Figure 16

Studentized Residuals for Fifth Grade ELA

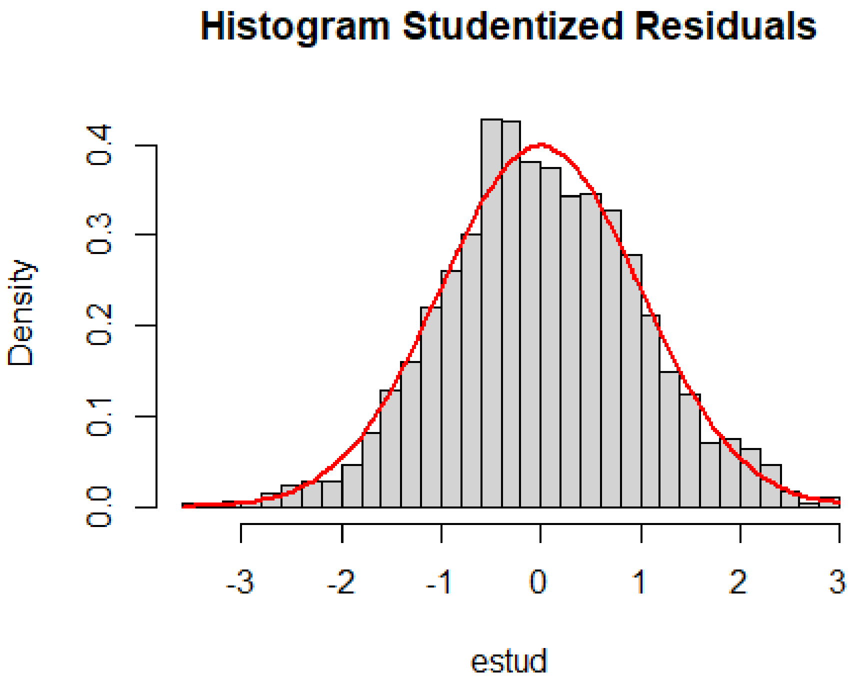
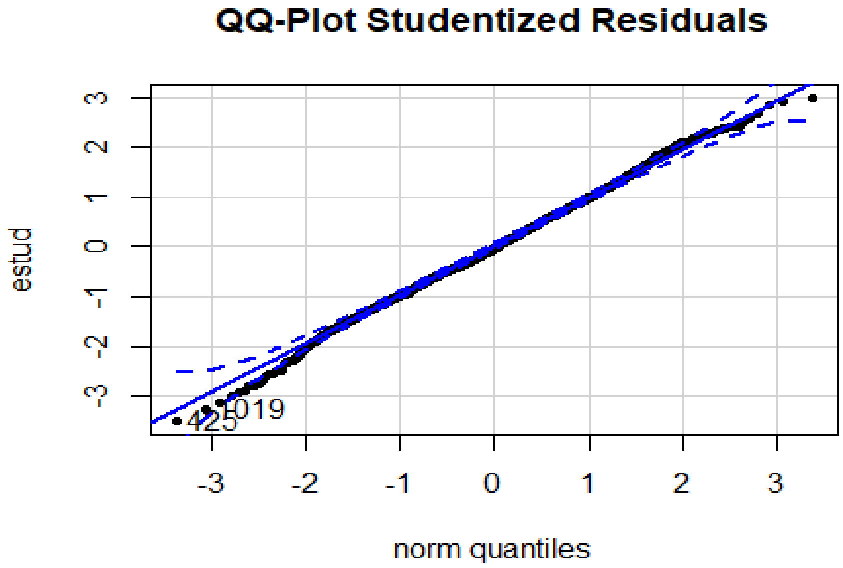


Figure 17

Q-Q Normality Plot of Studentized Residuals for Fifth Grade ELA



The assumption of multicollinearity was first tested using VIF factors. VIF factors with values fell close to 1.0, between 1.26 and 1.96, indicating a low correlation. In Table 22, polychoric and polyserial correlations were provided. No values exceeded .90, indicating no multicollinearity, and it was determined the assumption of multicollinearity was met.

As noted, Standard 3 and 7 ($r(1398) = .70$, $p < .05$), Standard 3 and 8, ($r(1398) = .75$, $p < .05$), Standard 3 and 8 ($r(1398) = .76$, $p < .05$), and Standard 5 and 6 ($r(1398) = .78$, $p < .05$), had a moderate to strong polychoric correlation. The weakest correlations were found between Standard 3 and 10 ($r(1398) = .48$, $p < .05$), Standard 4 and 9 ($r(1398) = .43$, $p < .05$), Standard 4 and 10 ($r(1398) = .40$, $p < .05$), Standard 5 and 9 ($r(1398) = .45$, $p < .05$), Standard 6 and 7 ($r(1398) = .47$, $p < .05$), Standard 6 and 9 ($r(1398) = .41$, $p < .05$), and Standard 8 and 10 ($r(1398) = .46$, $p < .05$), which had weak to moderate correlations. All polyserial correlations fell below .35. Standard 8 had a weak to moderate correlation to the MSS ($r(1398) = .35$, $p < .05$). Standard 10 had the weakest correlation to the MSS ($r(1398) = .21$, $p < .05$).

Table 22

Polychoric and Polyserial Correlations among Mean Scale Score (MSS) and TKES Standards in Fifth Grade ELA

	MSS	PS1	PS2	PS3	PS4	PS5	PS6	PS7	PS8	PS9	PS10
MSS											
PS 1	.26										
PS 2	.27	.61									
PS 3	.30	.64	.66								
PS 4	.23	.57	.65	.67							
PS 5	.25	.66	.66	.60	.61						
PS 6	.22	.62	.60	.60	.65	.78					
PS 7	.29	.58	.55	.70	.55	.55	.47				
PS 8	.35	.66	.65	.75	.65	.60	.60	.60			
PS 9	.32	.52	.53	.51	.43	.45	.41	.61	.50		
PS 10	.21	.57	.53	.48	.40	.54	.51	.58	.46	.67	

Note. MSS (Mean Scale Score), PS 1 (Professional Standard 1), PS 2 (Professional Standard 2), PS 3 (Professional Standard 3), PS 4 (Professional Standard 4), PS 5 (Professional Standard 5), PS 6 (Professional Standard 6), PS 7 (Professional Standard 7), PS 8 (Professional Standard 8), PS 9 (Professional Standard 9), PS 10 (Professional Standard 10).

Table 23 shows the results from the regression analysis. The model was considered significant. However, the R^2_{adj} indicated that only 11.4% of the variance was explained by the independent variables, $R = .356$. $R^2 = .127$, $R^2_{adj} = .114$, F(20, 1379) = 10.01, $p < .001$. Out of the 10 standards, 3 standards were significant, Standard 7, Standard 8, and Standard 9. Standard 7 at the exemplary level (B = .088, $t = 2.83$, $p < .01$), Standard 8 at the exemplary level (B = .128, $t = 2.18$, $p < .001$), and Standard 9 at the exemplary level (B = .146, $t = 4.92$, $p < .001$) are expected to have a positive effect on scores.

Table 23

Coefficient Table for Fifth Grade ELA Mean Scale Score (MSS)

	B	*SE of B*	*β*	*t*	*p*
PS1 ND	-7.95e-05	2.92e-04	-0.009	-0.27	.79
PS1 Exemplary	3.50e-05	4.45e-05	0.024	0.79	.43
PS2 ND	4.54e-05	1.42e-04	0.010	0.32	.75
PS2 Exemplary	8.46e-05	5.23e-05	0.049	1.62	.11
PS3 ND	-3.15e-05	1.71e-04	-0.006	-0.18	.85
PS3 Exemplary	7.64e-05	4.80e-05	0.051	1.59	.11
PS4 ND	-1.56e-04	1.58e-04	-0.034	-0.99	.32
PS4 Exemplary	-3.18e-05	5.90e-05	-0.016	-0.54	.59
PS5 ND	-1.65e-06	2.60e-04	-0.001e-01	-0.01	.99
PS5 Exemplary	4.84e-05	7.64e-05	0.019	0.63	.53
PS6 ND	-1.74e-04	1.69e-04	-0.031	-1.03	.30
PS6 Exemplary	-2.58e-05	6.36e-05	-0.012	-0.41	.68
PS7 ND	2.24e-04	1.34e-04	0.047	1.67	.09
PS7 Exemplary	1.19e-04	4.18e-05	0.088	2.83	<.01
PS8 ND	-1.43e-04	1.31e-04	-0.036	-1.09	.27
PS8 Exemplary	2.21e-04	5.28e-05	0.128	4.18	<.001
PS9 ND	-3.11e-04	1.81e-04	-0.051	-1.72	.09
PS9 Exemplary	1.92e-04	3.91e-05	0.146	4.92	<.001
PS10 ND	-2.18e-04	2.02e-04	-0.031	-1.08	.28
PS10 Exemplary	-3.84e-05	4.79e-05	-0.024	-0.80	.42

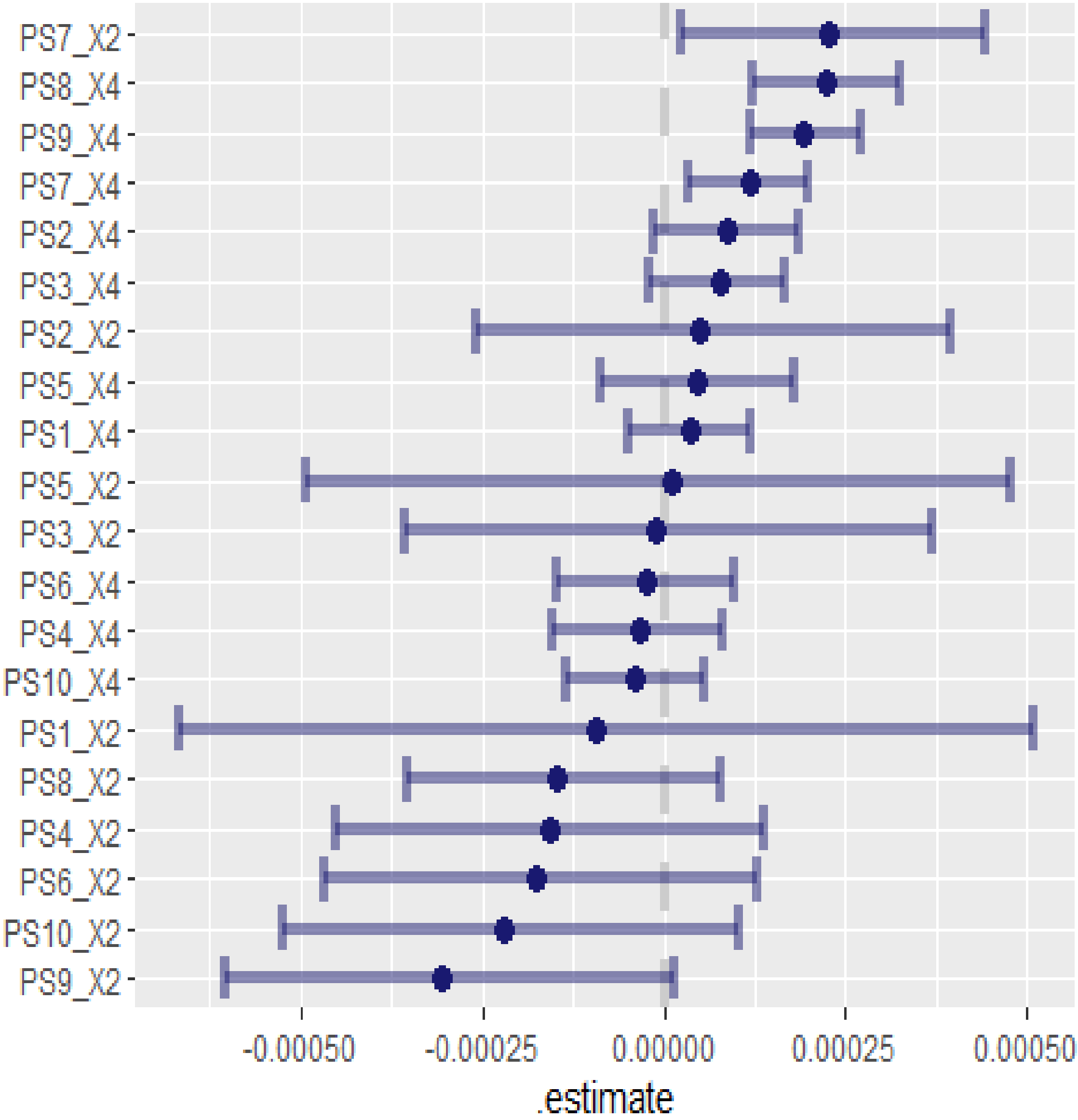
Note. PS 1 ND (Professional Standard 1 Needs Development Level), PS 1 Exemplary (Professional Standard 1 Exemplary Level), PS 2 ND (Professional Standard 2 Needs Development Level), PS 2 Exemplary (Professional Standard 2 Exemplary Level), PS 3 ND (Professional Standard 3 Needs Development Level), PS 3 Exemplary (Professional Standard 3

Exemplary Level), PS 4 ND (Professional Standard 4 Needs Development Level), PS 4 Exemplary (Professional Standard 4 Exemplary Level), PS 5 ND (Professional Standard 5 Needs Development Level), PS 5 Exemplary (Professional Standard 5 Exemplary Level), PS 6 ND (Professional Standard 6 Needs Development Level), PS 6 Exemplary (Professional Standard 6 Exemplary Level), PS 7 ND (Professional Standard 7 Needs Development Level), PS 7 Exemplary (Professional Standard 7 Exemplary Level), PS 8 ND (Professional Standard 8 Needs Development Level), PS 8 Exemplary (Professional Standard 8 Exemplary Level), PS 9 ND (Professional Standard 9 Needs Development Level), PS 9 Exemplary (Professional Standard 9 Exemplary Level), PS 10 ND (Professional Standard 10 Needs Development Level), PS 10 Exemplary (Professional Standard 10 Exemplary Level).

Figure 18 represents the 1,000 bootstrapping samples conducted on the dataset, rather than a single run of the data as represented in multiple regression analysis. Because of the increased representation of the data, there is a higher level of confidence in the outcome. The bootstrapping sample confirmed the standards found significant using the one-sample multiple regression, Standard 7 at the exemplary level, Standard 8 at the exemplary level, and Standard 9 at the exemplary level. Standard 7 at the ineffective/needs development level was also identified through the bootstrapping process as having a positive impact on scores. There is a high level of confidence that these standards significantly impact scores.

Figure 18

Bootstrapping Sample: Fifth Grade ELA



Fifth Grade Math

d. Are summative scores on TKES Standards significant predictors of the teacher's mean scaled score on the fifth grade Math portion of the Georgia Milestones Assessment System (GMAS)?

The goal of this sub-question was to determine if summative scores on TKES Standards are significant predictors of the teachers' mean scale score on the fifth grade Math portion of the GMAS. The independent variable was transformed to have three levels for each standard, including level 2 (needs development), level 3 (proficient), and level 4 (distinguished). Level 3 was made the reference level. Levels 2 and 4 values are presented in the summary estimates table for the fitted model. Multiple regression analysis was used as the analysis for this sub-question.

Statistical considerations were missing data and outliers. There was no missing data in the dataset. There were 17 values identified as potential outliers. Z-score values fell between -2.52 and 4.75. Skew (.81) and kurtosis (1.06) values were calculated.

Assumptions must be addressed before completing the analysis. The assumption of normality was not met. Results on the Shapiro-Wilk test on the raw data found data were not normally distributed (W(1398) = 0.96, $p < .0001$). Results from the Jarque Bera Test on the raw data also indicated a lack of normality (χ^2 (1398) = 257.88, $p < .0001$). Box-Cox transformation was used to transform the dataset. The distribution of the resulting variable was checked using the Shapiro-Wilk test. The results indicated that the transformed data were normally distributed (W(1398)=1.00, $p = .26$). The Jarque Bera Test also indicated normality on the transformed data (χ^2 (1398) = 0.37, $p = .83$). Figure 19, which represents the histogram for the transformed data, also indicates general normality. The data transformation helped achieve a normal distribution for the dependent variable and was used for the analysis.

Following the data transformation, outliers were readdressed. All z-score values fell between -3.65 and 3.10. Four potential outliers were identified. Skew (.00) and kurtosis (.08) values were reconsidered. Using Cook's distance test, no cases were identified that might influence the outcome. The data met the assumption of independent errors (DW = 2.04, *p* = .48).

The assumption of homoscedasticity of residuals was also met. In Figure 19 and Figure 20, the residuals remain approximately equal. The Breusch-Pagan test was not significant, indicating the null was accepted, and error variances were equal (BP(1398) = 21.65, *p* = .36).

Figure 19

Studentized Residuals for Fifth Grade Math

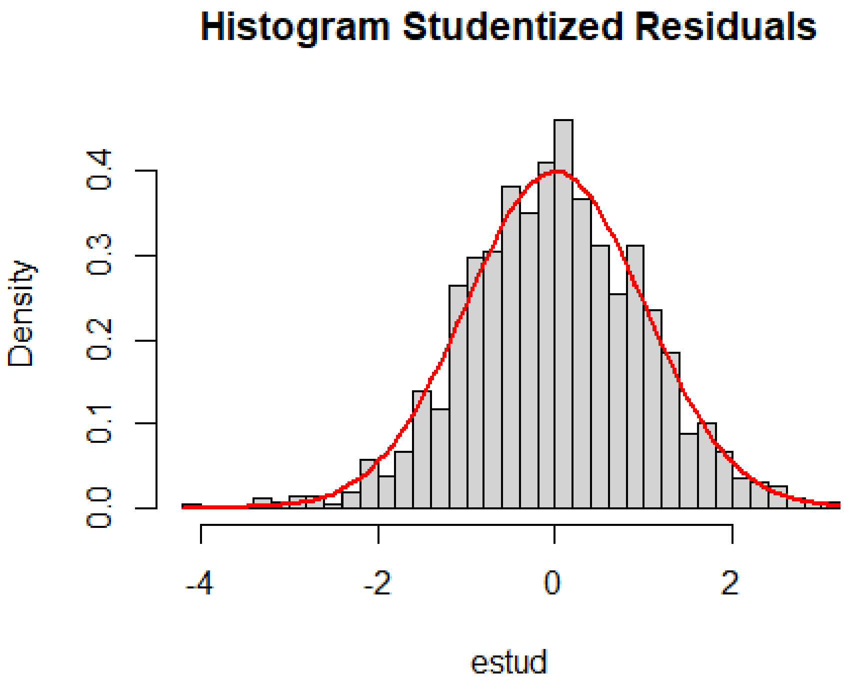
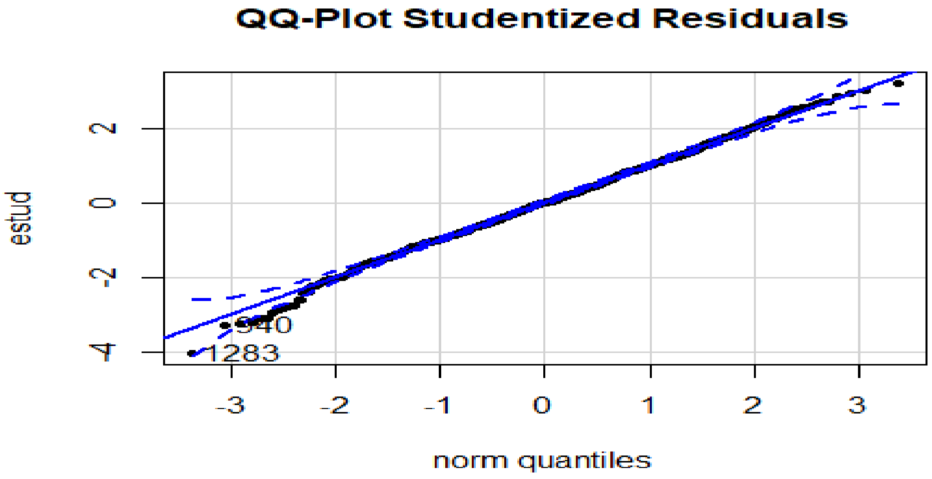


Figure 20

Q-Q Normality Plot of Studentized Residuals for Fifth Grade Math



The assumption of multicollinearity was also met. VIF factors were examined to address multicollinearity. The values in the dataset fell between 1.32 and 1.96, and because the values were close to zero, the assumption of multicollinearity was not violated. In Table 24, polychoric and polyserial correlations were provided. No values exceeded .90, and it was determined there was no multicollinearity.

As noted, Standard 3 and 8 (*r*(1398) = .81, *p* < .05) and Standard 5 and 6, (*r*(1398) = .82, *p* < .05 had a strong polychoric correlation. The weakest correlation was found between Standard 3 and 10 (*r*(1398) = .47, *p* < .05), Standard 4 and 9 (*r*(1398) = .48, *p* < .05), Standard 4 and 10 (*r*(1398) = .44, *p* < .05), Standard 5 and 9 (*r*(1398) = .49, *p* < .05), Standard 6 and 7 (*r*(1398) = .41, *p* < .05), Standard 6 and 9 (*r*(1398) = .42, *p* < .05), Standard 6 and 10 (*r*(1398) = .47, *p* < .05), and Standard 8 and 10 (*r*(1398) = .48, *p* < .05) had weak to moderate correlations. All polyserial correlations fell below .38. Standard 8 had a weak to moderate correlation to the MSS

($r(1398) = .38, p < .05$). Standard 10 had the weakest correlation to the MSS ($r(1398) = .22, p < .05$).

Table 24

Polychoric and Polyserial Correlations among Mean Scale Score (MSS) and TKES Standards in Fifth Grade Math

	MSS	PS1	PS2	PS3	PS4	PS5	PS6	PS7	PS8	PS9	PS10
MSS											
PS 1	.33										
PS 2	.27	.66									
PS 3	.31	.72	.68								
PS 4	.24	.55	.66	.65							
PS 5	.31	.62	.66	.66	.65						
PS 6	.30	.62	.62	.63	.65	.82					
PS 7	.31	.62	.58	.68	.56	.51	.41				
PS 8	.38	.69	.69	.81	.68	.64	.63	.63			
PS 9	.27	.55	.52	.53	.48	.49	.42	.60	.51		
PS 10	.22	.58	.59	.47	.44	.51	.47	.58	.48	.69	

Note. MSS (Mean Scale Score), PS 1 (Professional Standard 1), PS 2 (Professional Standard 2), PS 3 (Professional Standard 3), PS 4 (Professional Standard 4), PS 5 (Professional Standard 5), PS 6 (Professional Standard 6), PS 7 (Professional Standard 7), PS 8 (Professional Standard 8), PS 9 (Professional Standard 9), PS 10 (Professional Standard 10).

Table 25 shows the results from the regression analysis. Although the model was considered significant, the $R^2{}_{adj}$ indicated that only 11.9% of the variance was explained by the independent variables, $R = .362$. $R^2 = .131$, $R^2{}_{adj} = .119$, F(20, 1379) = 10.41, $p < .001$. Out of the 10 standards, 4 standards were significant, Standard 1, Standard 7, Standard 8, and Standard 9. Standard 1 at the exemplary level (B = .069, $t = 2.18$, $p < .03$), Standard 7 at the exemplary level (B = .111, $t = 3.60$, $p < .001$), Standard 8 at the exemplary level (B = .132, $t = 4.10$, $p < .001$), and Standard 9 at the exemplary level (B = .077, $t = 2.56$, $p < .01$) are expected to have a positive effect on scores.

Table 25

Coefficient Table for Fifth Grade Math Mean Scale Score (MSS)

	B	*SE of B*	*β*	*t*	*p*
PS1 ND	-1.42e-10	1.18e-10	-0.035	-1.20	.23
PS1 Exemplary	5.61e-11	2.57e-11	0.069	2.18	.03
PS2 ND	4.10e-11	7.30e-11	0.019	0.56	.57
PS2 Exemplary	2.43e-11	3.00e-11	0.025	0.81	.42
PS3 ND	-8.72e-11	8.17e-11	-0.037	-1.07	.29
PS3 Exemplary	1.64e-11	2.84e-11	0.019	0.58	.56
PS4 ND	4.34e-11	8.83e-11	0.017	0.49	.62
PS4 Exemplary	5.97e-12	3.21e-11	0.005	0.19	.85
PS5 ND	-1.33e-12	1.36e-10	<-0.001	-0.01	.99
PS5 Exemplary	2.43e-11	4.00e-11	0.019	0.61	.54
PS6 ND	2.64e-11	8.69e-11	0.009	0.30	.76
PS6 Exemplary	6.27e-11	3.63e-11	0.054	1.73	.08
PS7 ND	4.80e-11	7.35e-11	0.019	0.65	.51
PS7 Exemplary	8.57e-11	2.38e-11	0.111	3.60	<.001
PS8 ND	-1.01e-10	7.22e-11	-0.047	-1.40	.16
PS8 Exemplary	1.24e-10	3.04e-11	0.132	4.10	<.001
PS9 ND	-7.18e-11	8.75e-11	-0.025	-0.82	.41
PS9 Exemplary	5.79e-11	2.27e-11	0.077	2.56	.01
PS10 ND	-4.29e-11	1.12e-10	-0.011	-0.38	.70
PS10 Exemplary	-1.21e-11	2.75e-11	-0.013	-0.44	.66

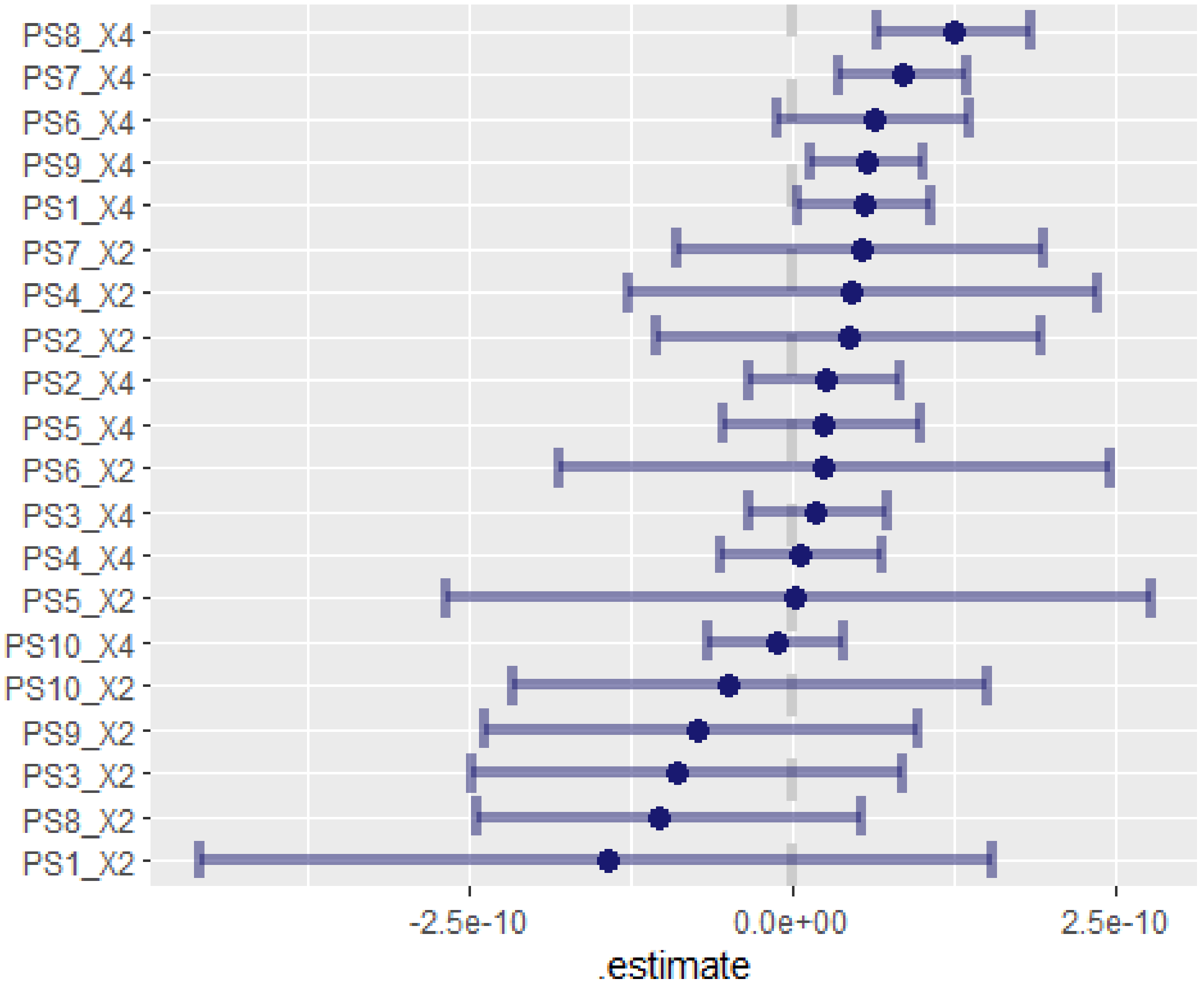
Note. PS 1 ND (Professional Standard 1 Needs Development Level), PS 1 Exemplary (Professional Standard 1 Exemplary Level), PS 2 ND (Professional Standard 2 Needs Development Level), PS 2 Exemplary (Professional Standard 2 Exemplary Level), PS 3 ND (Professional Standard 3 Needs Development Level), PS 3 Exemplary (Professional Standard 3

Exemplary Level), PS 4 ND (Professional Standard 4 Needs Development Level), PS 4 Exemplary (Professional Standard 4 Exemplary Level), PS 5 ND (Professional Standard 5 Needs Development Level), PS 5 Exemplary (Professional Standard 5 Exemplary Level), PS 6 ND (Professional Standard 6 Needs Development Level), PS 6 Exemplary (Professional Standard 6 Exemplary Level), PS 7 ND (Professional Standard 7 Needs Development Level), PS 7 Exemplary (Professional Standard 7 Exemplary Level), PS 8 ND (Professional Standard 8 Needs Development Level), PS 8 Exemplary (Professional Standard 8 Exemplary Level), PS 9 ND (Professional Standard 9 Needs Development Level), PS 9 Exemplary (Professional Standard 9 Exemplary Level), PS 10 ND (Professional Standard 10 Needs Development Level), PS 10 Exemplary (Professional Standard 10 Exemplary Level).

Bootstrapping samples, representing 1000 samples of the dataset, are represented in Table 25. There is a higher confidence level in the outcome than the multiple regression model provided. The bootstrapping sample confirmed the standards found significant using the one-sample multiple regression, Standard 1 at the exemplary level, Standard 7 at the exemplary level, Standard 8 at the needs development and exemplary levels, and Standard 9 at the exemplary level.

Figure 21

Bootstrapping Sample: Fifth Grade Math



Research Question 3

Is there a significant difference in academic setting (departmentalized or self-contained) by level of economically disadvantaged (ED) students on the teacher's mean scale score on the Georgia Milestones Assessment System (GMAS)?

Fourth Grade ELA

a. Is there a significant difference in academic setting (departmentalized or self-contained) by level of economically disadvantaged (ED) students on the teacher's mean scale score on the fourth grade English/Language Arts portion of the Georgia Milestones Assessment System (GMAS)?

The academic setting (departmentalized or self-contained) and levels of the percentage of (ED) students served as the independent variables. The dataset was comprised of 600 departmentalized teachers and 800 self-contained teachers. The percentage of the levels of ED students fell between 0% and 100% and were divided into upper and lower categories based on the median for each grade level and subject. For this sub-question, schools with percentages of students classified as ED between 0%-73.68% were categorized as "ED 1." Schools with levels of ED percentages falling between 73.69% and 100% were classified as "ED 2." There were 699 schools represented in the "ED 1," or lower ED group. There were 701 schools represented in the "ED 2," or group with the higher percentage levels of ED students. The teacher's mean scale score (MSS), or the average of their students' scale scores in their classes, served as the dependent variable.

Descriptive statistics for the MSS in fourth grade ELA are noted in Table 26 below. The MSS for fourth grade ELA was 508.56 (*SD* = 29.23), with scores ranging from 429.67 to 598.24. The MSS for the self-contained settings (SC) was 508.62 (*SD* = 30.98), with scores ranging from

429.67 to 598.24. Similarly, the mean for the departmentalized setting (DEPT) was 508.39 (*SD* = 26.75), with scores ranging from 429.67 to 593.91. The ED 1 mean was 526.55 (*SD* = 23.97), with scores ranging from 447.43 to 598.24. The ED 2 mean was 490.54 (*SD* = 22.06), with scores ranging from 429.67 to 562.92.

The means were also calculated for the ED groups by setting. Teachers in departmentalized ED 1 (DEPT ED 1) had a mean score of 525.77 (*SD* = 20.30) that fell between 482.02 and 593.91. Self-contained ED 1 teachers (SC ED 1) had a mean score of 527.13 (*SD* = 26.33). Scores ranged from 447.73 to 598.24. The MSS for teachers in departmentalized ED 2 (ED 2 DEPT) was 491.58 (*SD* = 20.80), with scores ranging from 429.67 to 562.92. The self-contained ED 2 group of teachers (SC ED 2) had a mean of 489.75 (*SD* = 22.94) with a range of scores between 429.67 and 560.47.

Table 26

Untransformed Descriptive Statistics for Fourth Grade ELA by Academic Setting and ED

Variable	n	*Mdn*	*M*	*SD*	Min	Max	Skew	Kurtosis
Fourth Grade ELA	1400	507.78	508.56	29.23	429.67	598.24	0.20	-0.08
DEPT	600	508.01	508.39	26.75	429.67	593.91	0.08	0.07
SC	800	507.59	508.62	30.98	429.67	598.24	0.26	-0.23
ED 1	699	525.78	526.55	23.97	447.43	598.24	0.20	0.56
ED 2	701	490.51	490.54	22.06	429.67	562.92	0.25	0.48
DEPT ED 1	295	524.08	525.77	20.30	482.02	593.91	0.48	0.17
SC ED 1	404	527.21	527.13	26.33	447.73	598.24	0.07	0.41
DEPT ED 2	305	492.42	491.58	20.80	429.67	562.92	-0.01	0.59
SC ED 2	396	489.11	489.75	22.94	429.67	560.47	0.42	0.43

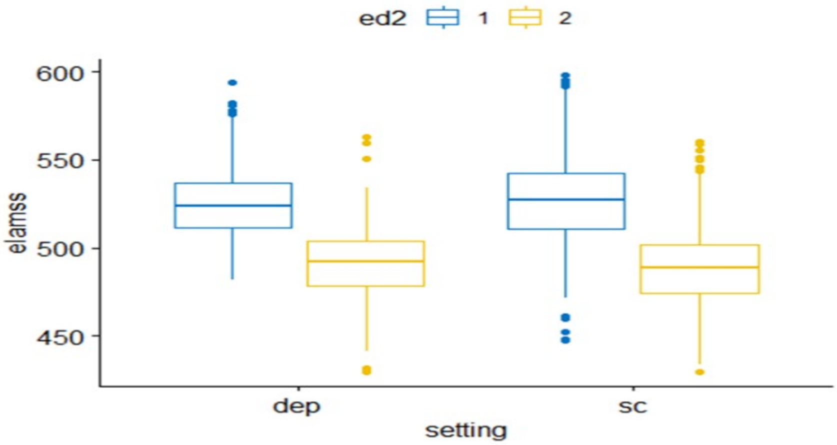
Note. Fourth grade data represents all scores in the dataset. DEPT (departmentalized), SC (self-contained), ED 1 (ED 1 group), ED 2 (ED 2 group), DEPT ED 1 (departmentalized in the ED 1 group), SC ED 1 (self-contained ED 1 group, DEPT ED 2 (departmentalized ED 2 group, SC ED 2 (self-contained ED 2 group).

A factorial analysis of variance (ANOVA) was utilized to determine if there was a significant difference between academic settings (departmentalized or self-contained) by levels of the percentage of ED students on the teacher's MSS on the GMAS. Statistical considerations and assumptions for factorial ANOVA statistical analyses were considered. There was no missing data for this question. The levels of ED students and the academic settings were regarded as nominal data. The dependent variable, the ELA teacher MSS score on the GMAS, was on the interval level of measurement.

Descriptive and exploratory techniques were also used to detect outliers in the dataset. The outliers were further subjected to exploratory analysis using box plots. The box plot confirmed the presence of outliers in the categories of the academic settings, as seen in Figure 22. In this figure, the points outside the box show the number of outliers in each category. As can be observed, all the categories had at least one outlier. Outliers with z-scores less than -3.50 or greater than 3.50 were negated by changing the data point to the closest value, which fell within the normal limits. The resulting Z-score values fell between -2.96 and 3.36.

Figure 22

Graphic Representation of MSS Outliers by Academic Setting and ED Group on Fourth Grade ELA



In the initial checks, assumptions were violated. The Shapiro-Wilk test indicated the data were not normally distributed within groups: DEPT ED 1 (W(1396)=0.98, *p* = .0005), DEPT ED 2 (W(1396)=0.99, *p* = .01), and SC ED 2 (W(1396)=0.98, *p* = .0002). Normality was indicated within SC ED 1 (W(1396)=0.99, *p* = .06). Values on Levene's Test for Homogeneity of Variance (HOV) (*F*(3,1396) = 7.09, *p* < .0001) indicated the assumption of homogeneity of variance was violated.

Due to a lack of HOV and normality, the scale score was transformed using the Yeo-Johnson transformation method to create a more normal distribution. All groups represented had transformed data means between 6.19 and 6.27, as indicated in Table 27. Teachers in DEPT ED 1 had an MSS of 6.26 (*SD* = 0.04), with means falling between 6.18 and 6.40. The mean for SC ED 1 was 6.27 (*SD* = 0.05), with values ranging from 6.10 and 6.39. Teachers in DEPT ED 2 had a mean of 6.20 (*SD* = 0.04) with values between 6.02 and 6.36. The mean for SC ED 2 teachers was 6.19 (*SD* = 0.05), with values ranging from 6.06 to 6.36. As indicated by the skewness and kurtosis values in Table 27, the data may be said to have a more normal univariate distribution than the untransformed data in Table 26.

Table 27

Transformed Descriptive Statistics for Fourth Grade ELA by Academic Setting and ED

Variable	*M*	*SD*	Min	Max	Skew	Kurtosis
DEPT ED 1	6.26	0.04	6.18	6.40	0.43	0.34
SC ED 1	6.27	0.05	6.10	6.39	-0.08	0.40
DEPT ED 2	6.20	0.04	6.02	6.36	-0.22	1.36
SC ED 2	6.19	0.05	6.06	6.36	0.35	0.46

Note. DEPT ED 1 (departmentalized in the ED 1 group), SC ED 1 (self-contained ED 1 group, DEPT ED 2 (departmentalized ED 2 group, SC ED 2 (self-contained ED 2 group).

Resulting Shapiro-Wilk values were as follows: DEPT ED 1 (W(1396)=0.99, *p* = .01), DEPT ED 2 (W(1396)=0.99, *p* = .004), and SC ED 2 (W(1396)=0.99, *p* = .02). Normality was indicated within SC ED 1 teachers (W(1396)=0.99, *p* = .07). Values on Levene's Test for Homogeneity of Variance for the transformed data (*F*(3,1396) = 6.10, *p* = .0004) indicated the assumption of homogeneity of variance was still violated.

Although there was progress toward meeting the assumptions, all subsets of the transformed data were not found to be normal. Because the untransformed data and the transformed data failed to meet the normality and HOV assumptions, an Aligned Rank Transform (ART) for nonparametric factorial ANOVA was used. The ART was run on both the untransformed and transformed data. Following an analysis of the data for both the transformed and untransformed analysis, it was determined the results of significance were in agreement between the two analyses. Because the findings of significance were the same between the two analyses, it was determined to report the findings of the untransformed dataset for ease of understanding and practical application.

Results from the untransformed ART indicated the interaction effect between levels of the percentage of ED students and academic setting was statistically significant (*F*(1,1396) = 4.00, *p* = .046, $\eta_p^2 = 0.003$). The effect size, $\eta_p^2 = 0.003$, was small, accounting for only 0.3% of the variance. The levels of percentage of ED students were statistically significant, *(F*(1,1396) = 985.67, p < .0001, $\eta_p^2 = 0.41$). The effect size, $\eta_p^2 = 0.41$, indicated a large effect size, accounting for 41% of the variance. The academic setting was not statistically significant (*F*(1,1396) = .0004, *p* = .98, $\eta_p^2 < 0.0001$).

Post hoc comparisons of significant results were conducted using the Tukey HSD. Although the overall interaction effect was significant, none of the individual post hoc comparisons were significant. The main effect of the levels of the percentage of ED was significant. Teachers in schools with a higher level of percentage of ED students had an estimated marginal mean (EMM) of 440 (*SE* = 11.8, 95% CI [413, 466]) and teachers in schools with a lower level of percentage of ED students had an EMM of 963 (*SE* = 11.8, 95% CI [937, 990]); (*t*(1396) = 31.40, *p* < .0001).

Fourth Grade Math

b. Is there a significant difference in academic setting (departmentalized or self-contained) by level of economically disadvantaged (ED) students on the teacher's mean scale score on the fourth grade math portion of the Georgia Milestones Assessment System (GMAS)?

In the fourth grade math dataset, schools with percentages of students classified as ED between 0%-74.47% were categorized as "ED 1." Schools with levels of ED percentages falling between 73.69% and 100% were classified as "ED 2." There were 701 schools represented in the

"ED 1," or lower ED group. There were 699 schools represented in the "ED 2," or group with the higher percentage levels of ED students.

Descriptive statistics for the MSS in fourth grade math are noted in Table 28 below. The MSS for fourth grade math was 524.59 (*SD* = 29.95), with scores ranging from 450.53 to 533.05. The MSS for the self-contained setting (SC) was 523.33 (*SD* = 29.68), with scores ranging from 450.53 to 630.52. Similarly, the departmentalized setting (DEPT) mean was 524.40 (*SD* = 30.26), with scores ranging from 457.59 to 633.05. The ED 1 mean was 540.44 (*SD* = 26.41), with scores ranging from 469.68 to 633.05. The ED 2 mean was 507.29 (*SD* = 22.36), with scores falling between 450.53 and 584.31.

The means were also calculated for the ED groups by setting. Teachers in departmentalized ED 1 (DEPT ED 1) had a mean score of 545.34 (*SD* = 25.35) that fell between 489.84 and 633.05. Self-contained ED 1 teachers (SC ED 1) had a mean score of 539.41 (*SD* = 26.89). Scores ranged from 469.68 to 630.52. The MSS for teachers in departmentalized ED 2 (ED 2 DEPT) was 508.67 (*SD* = 22.86), with scores ranging from 457.59 to 584.31. The self-contained ED 2 group of teachers (SC ED 2) had a mean of 506.17 (*SD* = 21.91) with a range of scores between 450.53 and 578.67.

Table 28

Untransformed Descriptive Statistics for Fourth Grade Math by Academic Setting and ED

Variable	n	*Mdn*	*M*	*SD*	Min	Max	Skew	Kurtosis
Fourth Grade Math	1400	522.14	524.59	29.95	450.53	533.05	0.50	0.21
DEPT	600	524.40	526.27	30.26	457.59	633.05	0.46	0.16
SC	800	520.51	523.33	29.68	450.53	630.52	0.54	0.25
ED 1	701	540.44	541.85	26.41	469.68	633.05	0.55	0.51
ED 2	699	504.63	507.29	22.36	450.53	584.31	0.61	0.65
DEPT ED 1	288	543.19	545.34	25.35	489.84	633.05	0.65	0.61
SC ED 1	413	537.81	539.41	26.89	469.68	630.52	0.53	0.44
DEPT ED 2	312	505.49	508.67	22.86	457.59	584.31	0.63	0.71
SC ED 2	387	503.52	506.17	21.91	450.53	578.67	0.58	0.54

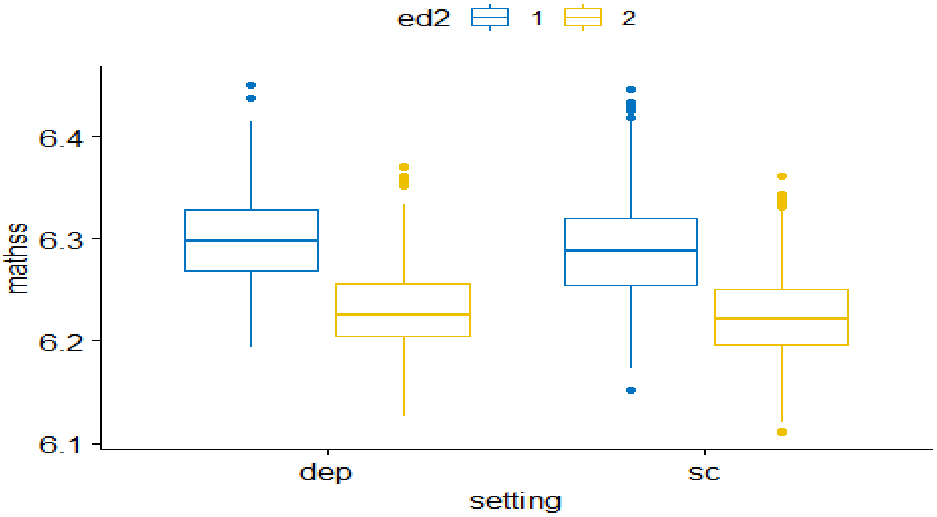
Note. Fourth grade data represents all scores in the dataset. DEPT (departmentalized), SC (self-contained), ED 1 (ED 1 group), ED 2 (ED 2 group), DEPT ED 1 (departmentalized in the ED 1 group), SC ED 1 (self-contained ED 1 group, DEPT ED 2 (departmentalized ED 2 group, SC ED 2 (self-contained ED 2 group).

Statistical considerations and assumptions for factorial ANOVA statistical analyses were considered. There was no missing data for this sub-question. The levels of ED students and the academic settings were regarded as nominal data, and the dependent variable, the math teacher MSS score on the GMAS, was on the interval level of measurement.

The outliers were detected using descriptive and exploratory techniques and subjected to further analysis using box plots. Outliers in the categories of the academic settings were confirmed, as noted in Figure 23. All categories had at least one outlier. Outliers with z-scores less than -3.50 or greater than 3.50 were negated by changing the data point to the closest value. The resulting Z-score values fell between -2.54 and 3.46.

Figure 23

Graphic Representation of MSS Outliers by Academic Setting and ED Group in Fourth Grade Math



In the initial checks, assumptions were violated. The Shapiro-Wilk test indicated the data were not normally distributed within groups: DEPT ED 1 (W(1396)=0.97, $p < .0001$), DEPT ED 2 (W(1396)=0.97, $p < .0001$), SC ED 1 (W(1396)=0.98, $p < .0001$), and SC ED 2 (W(1396)=0.98, $p < .0001$. Values on Levene's Test for Homogeneity of Variance (HOV) (F(3,1396) = 6.09, $p = .0004$) indicated the assumption of homogeneity of variance was violated.

The scale scores were transformed using the Yeo-Johnson transformation due to a lack of normality and HOV. All groups represented had transformed data means between 6.23 and 6.30, as indicated in Table 29. Teachers in DEPT ED 1 had an MSS of 6.30 (*SD* = 0.05), with means falling between 6.19 and 6.45. The mean for SC ED 1 was 6.29 (*SD* = 0.05), with values ranging from 6.15 and 6.45. Teachers in DEPT ED 2 had a mean of 6.23 (*SD* = 0.04) with values between 6.13 and 6.37. The mean for SC ED 2 teachers was 6.23 (*SD* = 0.04), with values ranging from 6.11 to 6.36. As indicated by the skewness and kurtosis values in Table 29, the data may be said to have a more normal univariate distribution than the untransformed data in Table 28.

Table 29

Transformed Descriptive Statistics for Fourth Grade Math by Academic Setting and ED

Variable	*M*	*SD*	Min	Max	Skew	Kurtosis
DEPT ED 1	6.30	0.05	6.19	6.45	0.51	0.36
SC ED 1	6.29	0.05	6.15	6.45	0.37	0.22
DEPT ED 2	6.23	0.04	6.13	6.37	0.47	0.45
SC ED 2	6.23	0.04	6.11	6.36	0.44	0.38

Note. DEPT ED 1 (departmentalized in the ED 1 group), SC ED 1 (self-contained ED 1 group, DEPT ED 2 (departmentalized ED 2 group, SC ED 2 (self-contained ED 2 group).

Resulting Shapiro-Wilk values were as follows: DEPT ED 1 (W(1396)=0.98, *p* = .0009), DEPT ED 2 (W(1396)=0.98, *p* = .0008), SC ED 1 teachers (W(1396)=0.99, *p* = .005). and SC ED 2 (W(1396)=0.98, *p* = .0003. Values on Levene's Test for Homogeneity of Variance for the transformed data (*F*(3,1396) = 3.10, *p* = .03) indicated the assumption of homogeneity of variance was still violated.

There was progress toward meeting the assumptions, but subsets of the transformed data were not found to be normal. An Aligned Rank Transform (ART) for nonparametric factorial ANOVA was used due to the lack of normality of both the untransformed and transformed data. The ART was run on both the untransformed and transformed data. The results of significance were in agreement, and it was determined to report the findings of the untransformed dataset for ease of understanding and practical application.

Results from the untransformed ART indicated the interaction effect between levels of the percentage of ED students and academic setting was not statistically significant (*F*(1,1396) = 1.58, *p* = .21, η_p^2 = 0.001) The levels of percentage of ED students were statistically significant, (*F*(1,1396) = 797.46, p < .0001, η_p^2 = 0.36). The effect size, η_p^2 = 0.36, indicated a large effect size, accounting for 36% of the variance. The academic setting was statistically significant ($F(1,1396) = 5.37$ $p = .02$, $\eta_p^2 < 0.004$). The effect size, $\eta_p^2 < 0.004$, indicated a small effect size, accounting for .4% of the variance.

Post hoc comparisons of significant results were conducted using the Tukey HSD. The main effects of the levels of the percentage of ED and the setting were significant. Teachers in schools with a higher level of percentage of ED students had an EMM of 455 (*SE* = 12.2, 95% CI [428, 482]) and teachers in schools with a lower level of percentage of ED students had an EMM of 946 (*SE* = 12.3, 95% CI [918, 974]); ($t(1396) = 28.28$, $p < .0001$). Self-contained teachers had an EMM of 679 (*SE* = 14.3, 95% CI [647, 711]) and departmentalized teachers had an EMM of 729 (*SE* = 16.5, 95% CI [692, 766]); ($t(1396) = 2.32$, $p =.02$).

Fifth Grade ELA

c. Is there a significant difference in academic setting (departmentalized or self-contained) by level of economically disadvantaged (ED) students on the teacher's mean scale score on the fourth grade English/Language Arts portion of the Georgia Milestones Assessment System (GMAS)?

For this sub-question, schools with percentages of students classified as ED between 0%-75.56% were categorized as "ED 1." Schools with levels of ED percentages falling between 75.57% and 100% were classified as "ED 2." There were 700 schools represented in the "ED 1," or lower ED group. There were 700 schools represented in the "ED 2," or group with the higher percentage levels of ED students.

Descriptive statistics for the MSS in fourth grade ELA are noted in Table 30 below. The MSS for fifth grade ELA was 514.38 (*SD* = 31.94), with scores ranging from 418.11 to 612.59. The MSS for the self-contained setting (SC) was 515.09 (*SD* = 33.81), with scores from 425.73 to 612.59. Similarly, the departmentalized setting (DEPT) mean was 513.44 (*SD* = 29.27), with scores ranging from 418.11 to 612.59. The ED 1 mean was 532.87 (*SD* = 26.98), with scores ranging from 456.83 to 612.59. The ED 2 mean was 495.16 (*SD* = 23.96.), with scores ranging from 418.11 to 574.07.

The means were also calculated for the ED groups by setting. Teachers in departmentalized ED 1 (DEPT ED 1) had a mean score of 531.95 (*SD* = 23.55) that fell between 456.85 and 612.59. Self-contained ED 1 teachers (SC ED 1) had a mean score of 534.80 (*SD* = 29.17). Scores ranged from 456.83 to 612.59. The MSS for teachers in departmentalized ED 2 (ED 2 DEPT) was 495.78 (*SD* = 22.51), with scores ranging from 418.11 to 574.07. The self-contained ED 2 group of teachers (SC ED 2) had a mean of 494.67 (*SD* = 22.05) with scores between 425.73 and 574.07.

Table 30

Untransformed Descriptive Statistics for Fifth Grade ELA by Academic Setting and ED

Variable	n	*Mdn*	*M*	*SD*	Min	Max	Skew	Kurtosis
Fifth Grade ELA	1400	511.50	514.38	31.94	418.11	612.59	0.30	0.02
DEPT	600	512.42	513.44	29.27	418.11	612.59	0.23	0.13
SC	800	511.15	515.09	33.81	425.73	612.59	0.31	-0.13
ED 1	700	532.87	533.61	26.98	456.83	612.59	0.35	0.10
ED 2	700	494.09	495.16	23.96	418.11	574.07	0.27	0.75
DEPT ED 1	293	529.40	531.95	23.55	456.85	612.59	0.39	0.62
SC ED 1	407	535.36	534.80	29.17	456.83	612.59	0.29	-0.24
DEPT ED 2	307	494.71	495.78	22.51	418.11	574.07	0.27	1.04
SC ED 2	393	493.76	494.67	25.05	425.73	574.07	0.28	0.54

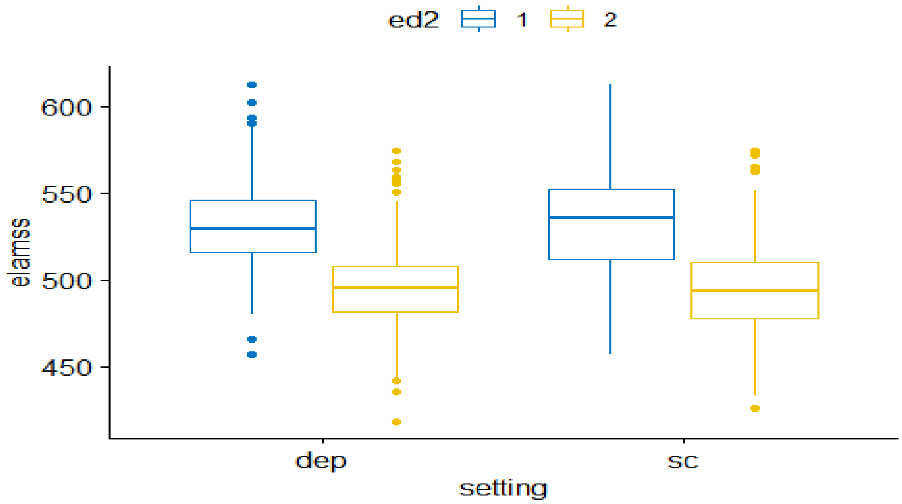
Note. Fifth grade data represents all scores in the dataset. DEPT (departmentalized), SC (self-contained), ED 1 (ED 1 group), ED 2 (ED 2 group), DEPT ED 1 (departmentalized in the ED 1 group), SC ED 1 (self-contained ED 1 group, DEPT ED 2 (departmentalized ED 2 group, SC ED 2 (self-contained ED 2 group).

Statistical considerations and assumptions for factorial ANOVA statistical analyses were considered for this dataset, and there was no missing data. The levels of ED students and the academic settings were again regarded as nominal data, and the dependent variable, the ELA teacher's MSS score on the GMAS, was on the interval level of measurement.

Outliers were identified using descriptive and exploratory analyses. As seen in Figure 24, the box plot confirmed the presence of outliers in each category. Outliers with z-scores less than -3.50 or greater than 3.50 were negated by changing the data point to the closest value. The resulting Z-score values fell between -3.42 and 3.45.

Figure 24

Graphic Representation of MSS Outliers by Academic Setting and ED Group in Fifth Grade ELA



In the initial checks, assumptions were violated. The Shapiro-Wilk test indicated the data were not normally distributed within groups: DEPT ED 1 (W(1396)=0.98, *p* < .002), DEPT ED 2 (W(1396)=0.98, *p* = .003), SC ED 1 (W(1396)=0.99, *p* = .003), and SC ED 2 (W(1396)=0.99, *p* = .008). Values on Levene's Test for Homogeneity of Variance (HOV) (*F*(3,1396) = 11.35, *p* < .0001) indicated the assumption of homogeneity of variance was violated.

The Yeo-Johnson transformation method was used to create a more normal distribution because no groups were normally distributed or met the HOV assumption. Once transformed, all groups represented had data means between 6.20 and 6.28, as indicated in Table 31. Teachers in DEPT ED 1 had an MSS of 6.28 (*SD* = 0.04), with means falling between 6.12 and 6.42. The mean for SC ED 1 was 6.28 (*SD* = 0.05), with values ranging from 6.12 and 6.42. Teachers in DEPT ED 2 had a mean of 6.21 (*SD* = 0.05) with values between 6.04 and 6.35. The mean for SC ED 2 teachers was 6.20 (*SD* = 0.05), with values ranging from 6.05 to 6.35. As indicated by the skewness and kurtosis values in Table 31, the data may be said to have a more normal univariate distribution than the untransformed data in Table 30.

Table 31

Transformed Descriptive Statistics for Fifth Grade ELA by Academic Setting and ED

Variable	*M*	*SD*	Min	Max	Skew	Kurtosis
DEPT ED 1	6.28	0.04	6.12	6.42	0.23	0.58
SC ED 1	6.28	0.05	6.12	6.42	0.15	-0.33
DEPT ED 2	6.21	0.05	6.04	6.35	0.07	1.02
SC ED 2	6.20	0.05	6.05	6.35	0.09	0.41

Note. DEPT ED 1 (departmentalized in the ED 1 group), SC ED 1 (self-contained ED 1 group, DEPT ED 2 (departmentalized ED 2 group, SC ED 2 (self-contained ED 2 group).

Resulting Shapiro-Wilk values were as follows: DEPT ED 1 (W(1396)=0.99, p = .01), DEPT ED 2 (W(1396)=0.99, p = .01), SC ED 1 (W(1396)=0.99, p = .01). and SC ED 2 (W(1396)=0.99, p = .07). Values on Levene's Test for Homogeneity of Variance for the transformed data ($F(3,1396) = 7.83, p < .0001$) indicated the assumption of homogeneity of variance was still violated.

The transformed data were not found to be normal. Because both the untransformed and the transformed data failed to meet the normality and HOV assumptions, an Aligned Rank Transformation (ART) was run on both the untransformed and transformed data. Because the results of significance were in agreement between the analyses, the findings of the untransformed ART were interpreted for ease of understanding and practical application.

Results from the untransformed ART indicated the interaction effect between levels of the percentage of ED students and academic setting was not statistically significant ($F(1,1396) = 2.42, p = .12, \eta_p^2 = 0.002$). The levels of percentage of ED students were statistically significant, ($F(1,1396) = 910.72$, p < .0001, $\eta_p^2 = 0.40$). The effect size, $\eta_p^2 = 0.40$, indicated a large effect size, accounting for 40% of the variance. The academic setting was not statistically significant ($F(1,1396) = 1.39$, $p = .24$, $\eta_p^2 < 0.001$).

Post hoc comparisons of significant results were conducted using the Tukey HSD. The main effect of the levels of the percentage of ED was significant. Teachers in schools with a higher level of percentage of ED students had an estimated marginal mean (EMM) of 445 (*SE* = 12.0, 95% CI [418, 472]) and teachers in schools with a lower level of percentage of ED students had an EMM of 957 (*SE* = 12.0, 95% CI [930, 984]); ($t(1396) = 30.18$, $p < .0001$).

Fifth Grade Math

d. Is there a significant difference in academic setting (departmentalized or self-contained) by level of economically disadvantaged (ED) students on the teacher's mean scale score on the fourth grade English/Language Arts portion of the Georgia Milestones Assessment System (GMAS)?

In the fifth grade math dataset used for this sub-question, schools with percentages of students classified as ED between 0%-76.00% were categorized as "ED 1." Schools with levels of ED percentages falling between 76.01% and 100% were classified as "ED 2." There were 701 schools represented in the "ED 1," or lower ED group. There were 699 schools represented in the "ED 2," or group with the higher percentage levels of ED students.

Descriptive statistics for the MSS in fifth grade math are noted in Table 32 below. The MSS for fifth grade math was 514.79 (*SD* = 32.62), with scores ranging from 432.12 to 640.05. The MSS for the self-contained setting (SC) was 514.13 (*SD* = 34.32), with scores ranging from 432.12 to 640.05. Similarly, the mean for the departmentalized setting (DEPT) was 515.68 (*SD* = 30.20), with scores ranging from 437.44 to 625.00. The ED 1 mean was 533.79 (*SD* = 31.04), with scores ranging from 459.88 to 640.05. The ED 2 mean was 496.69 (*SD* = 22.11), with scores ranging from 432.12 to 575.35.

The means were also calculated for the ED groups by setting. Teachers in departmentalized ED 1 (DEPT ED 1) had a mean score of 534.24 (*SD* = 27.38) that fell between 459.73 and 625.00. Self-contained ED 1 teachers (SC ED 1) had a mean score of 533.48 (*SD* = 33.41). Scores ranged from 458.89 to 640.05. The MSS for teachers in departmentalized ED 2 (ED 2 DEPT) was 499.22 (*SD* = 21.93), with scores ranging from 437.44 to 575.35. The self-contained ED 2 group of teachers (SC ED 2) had a mean of 494.68 (*SD* = 21.17) with scores between 432.12 and 569.60.

Table 32

Untransformed Descriptive Statistics for Fifth Grade Math by Academic Setting and ED

Variable	n	*Mdn*	*M*	*SD*	Min	Max	Skew	Kurtosis
Fifth Grade Math	1400	509.80	514.79	32.62	432.12	640.05	0.77	0.86
DEPT	600	511.51	515.68	30.20	437.44	625.00	0.62	0.72
SC	800	508.03	514.13	34.32	432.12	640.05	0.86	0.87
ED 1	701	529.69	533.79	31.04	459.88	640.05	0.75	0.60
ED 2	699	495.89	496.69	22.11	432.12	575.35	0.43	0.73
DEPT ED 1	282	532.30	534.24	27.38	459.73	625.00	0.74	0.98
SC ED 1	401	529.28	533.48	33.41	458.89	640.05	0.75	0.32
DEPT ED 2	318	499.31	499.22	21.93	437.44	575.35	0.43	0.65
SC ED 2	399	494.74	494.68	21.17	432.12	569.60	0.44	0.80

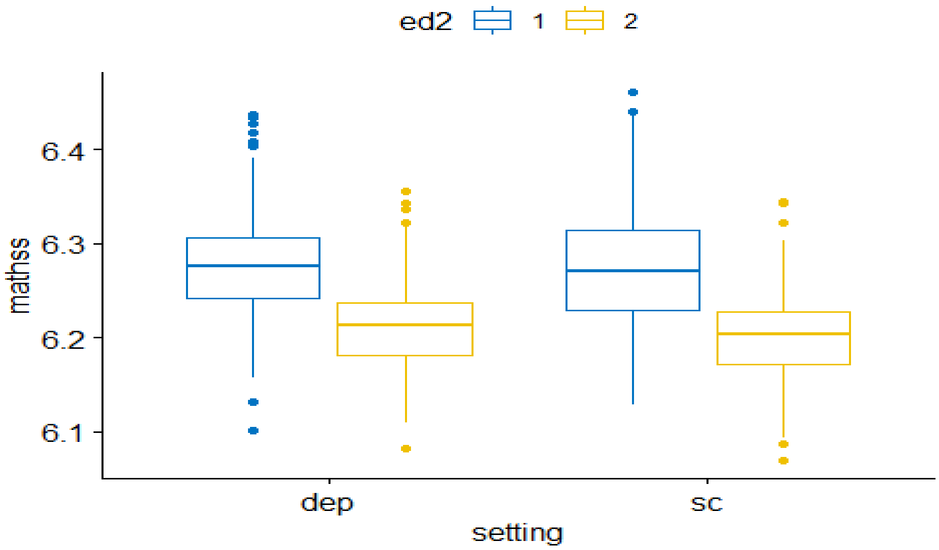
Note. Fifth grade data represents all scores in the dataset. DEPT (departmentalized), SC (self-contained), ED 1 (ED 1 group), ED 2 (ED 2 group), DEPT ED 1 (departmentalized in the ED 1 group), SC ED 1 (self-contained ED 1 group, DEPT ED 2 (departmentalized ED 2 group, SC ED 2 (self-contained ED 2 group).

Statistical considerations and assumptions for factorial ANOVA statistical analyses were considered in the fifth grade math dataset. There was no missing data for this question. The levels of ED students and the academic settings were regarded as nominal data. The dependent variable, the math teacher's MSS score on the GMAS, was on the interval level of measurement.

Outliers were identified using descriptive and exploratory techniques. As seen in the box plot below (Figure 25), outliers were identified. As can be observed, all the categories had at least one outlier. Outliers with z-scores less than -3.50 or greater than 3.50 were negated by changing the data point to the closest value, which fell within the normal limits. The resulting Z-score values fell between -3.13 and 3.49.

Figure 25

Graphic Representation of MSS Outliers by Academic Setting and ED Group in Fifth Grade Math



Initially, assumptions were violated. The Shapiro-Wilk test indicated the data were not normally distributed within groups: DEPT ED 1 (W(1396)=0.97, p < .0001), DEPT ED 2 (W(1396)=0.98, p = .001), SC ED 1 (W(1396)=0.96, p < .0001), and SC ED 2 (W(1396)=0.98, p = .0002). Values on Levene's Test for Homogeneity of Variance (HOV) (F(3,1396) = 23.04, p < .0001) indicated the assumption of homogeneity of variance was violated.

The scale score was transformed using the Yeo-Johnson transformation method to create a more normal distribution. Transformed means fell between 6.20 and 6.28, as indicated in Table 33. Teachers in DEPT ED 1 had an MSS of 6.28 (*SD* = 0.05), with means falling between 6.10 and 6.44. The mean for SC ED 1 was 6.28 (*SD* = 0.06), with values ranging from 6.13 and 6.46. Teachers in DEPT ED 2 had a mean of 6.21 (*SD* = 0.04) with values between 6.08 and 6.35. The mean for SC ED 2 teachers was 6.20 (*SD* = 0.04), with values ranging from 6.07 to 6.34. As indicated by the skewness and kurtosis values in Table 33, the data may be said to have a more normal univariate distribution than the untransformed data in Table 32.

Table 33

Transformed Descriptive Statistics for Fifth Grade Math by Academic Setting and ED

Variable	*M*	*SD*	Min	Max	Skew	Kurtosis
DEPT ED 1	6.28	0.05	6.10	6.44	0.42	0.90
SC ED 1	6.28	0.06	6.13	6.46	0.62	0.06
DEPT ED 2	6.21	0.04	6.08	6.35	0.33	0.48
SC ED 2	6.20	0.04	6.07	6.34	0.29	0.60

Note. DEPT ED 1 (departmentalized in the ED 1 group), SC ED 1 (self-contained ED 1 group, DEPT ED 2 (departmentalized ED 2 group, SC ED 2 (self-contained ED 2 group).

Resulting Shapiro-Wilk values were as follows: DEPT ED 1 (W(1396)=0.98, p = .0002), DEPT ED 2 (W(1396)=0.99, p =.024), SC ED 1 teachers (W(1396)=0.97, p < .0001), and SC ED 2 (W(1396)=0.99, p = .008). Normality was indicated within SC ED 1 teachers (W(1396)=0.97, p < .0001). Values on Levene's Test for Homogeneity of Variance for the transformed data (F(3,1396) = 16.65, p < .0001) indicated the assumption of homogeneity of variance was still violated.

Due to the lack of HOV and normality, the Yeo Johnson transformation was employed. The transformed data were not found to be normal. Because both the untransformed and transformed data failed to meet the normality and HOV assumptions, an Aligned Rank Transformation (ART) was used on both the untransformed and transformed data. The results of significance were in agreement. It was determined to report the findings of the untransformed dataset for understanding and application.

Results from the untransformed ART indicated the interaction effect between levels of the percentage of ED students and academic setting was not statistically significant ($F(1,1396) = 1.03$, $p = .31$, $\eta_p^2 = 0.0007$). The levels of percentage of ED students were statistically significant, ($F(1,1396) = 748.04$, $p < .0001$, $\eta_p^2 = 0.35$). The effect size, $\eta_p^2 = 0.35$, indicated a large effect size, accounting for 35% of the variance. The academic setting was not statistically significant ($F(1,1396) = 2.92$, $p = .09$, $\eta_p^2 = 0.002$).

Post hoc comparisons of significant results were conducted using the Tukey HSD. The main effect of the levels of the percentage of ED was significant. Teachers in schools with a higher level of percentage of ED students had a significantly different mean (EMM = 461, *SE* = 12.4, 95% CI [433, 489]) than teachers in schools with a lower level of percentage of ED (EMM = 963, *SE* = 12.5 95% CI [915, 971]); ($t(1396) = 27.35$, $p < .0001$).

Research Question 4

How many reliable and interpretable components are there among the following variables: Professional Knowledge, Instructional Planning, Instructional Strategies, Differentiated Instruction, Assessment Strategies, Assessment Uses, Positive Learning Environment, Academically Challenging Environment, Professionalism, and Communication?

Fourth Grade ELA

a. How many reliable and interpretable components are there among the following variables: Professional Knowledge, Instructional Planning, Instructional Strategies, Differentiated Instruction, Assessment Strategies, Assessment Uses, Positive Learning Environment, Academically Challenging Environment, Professionalism, and Communication among fourth grade English Language Arts teachers?

Descriptive Statistics

The analysis for this research question utilized TKES Standard score data from 600 fourth grade ELA teachers. Only teachers who exclusively taught ELA were included so that the results would be specific to subjects and grades. As indicated in Table 34, most scores for each standard fell in the proficient range. Interestingly, Standard 5 had the largest percentage of scores (93.17%) in the proficient area. Standards 7 and 9 had the highest percentages in the exemplary area, 31.83% and 36.17%, respectively. Standards 5 and 6 had the smallest percentage of scores in the exemplary level, with larger percentages than other standards falling in the ineffective/needs improvement and proficient ranges. Standard 8 had the highest percentage (3.50%) of teachers scoring in the ineffective/needs improvement range. Standards 2, 4, 5, 6, 8, and 10 had over 80% of teachers scoring in the proficient range.

Table 34

Number and Percentage of Fourth Grade ELA Teachers Scoring at Ineffective/Needs Improvement, Proficient, or Exemplary by Professional Standard

TKES Categories	Ineffective/ Needs Improvement	Proficient	Exemplary
PS 1	6 (1.00%)	457 (76.17%)	137 (22.83%)
PS 2	14 (2.33%)	492 (82.00%)	94 (15.67%)
PS 3	11 (1.83%)	464 (77.33%)	125 (20.83%)
PS 4	8 (1.33%)	531 (88.50%)	61 (10.17%)
PS 5	3 (0.50%)	559 (93.17%)	38 (6.33%)
PS 6	8 (1.33%)	535 (89.17%)	57 (9.50%)
PS 7	10 (1.67%)	399 (66.50%)	191 (31.83%)
PS 8	21 (3.50%)	500 (83.33%)	79 (13.17%)
PS 9	10 (1.67%)	373 (62.17%)	217 (36.17%)
PS 10	8 (1.33%)	494 (82.33%)	98 (16.33%)

Note. PS 1 (Professional Standard 1), PS 2 (Professional Standard 2), PS 3 (Professional Standard 3), PS 4 (Professional Standard 4), PS 5 (Professional Standard 5), PS 6 (Professional Standard 6), PS 7 (Professional Standard 7), PS 8 (Professional Standard 8), PS 9 (Professional Standard 9), PS 10 (Professional Standard 10).

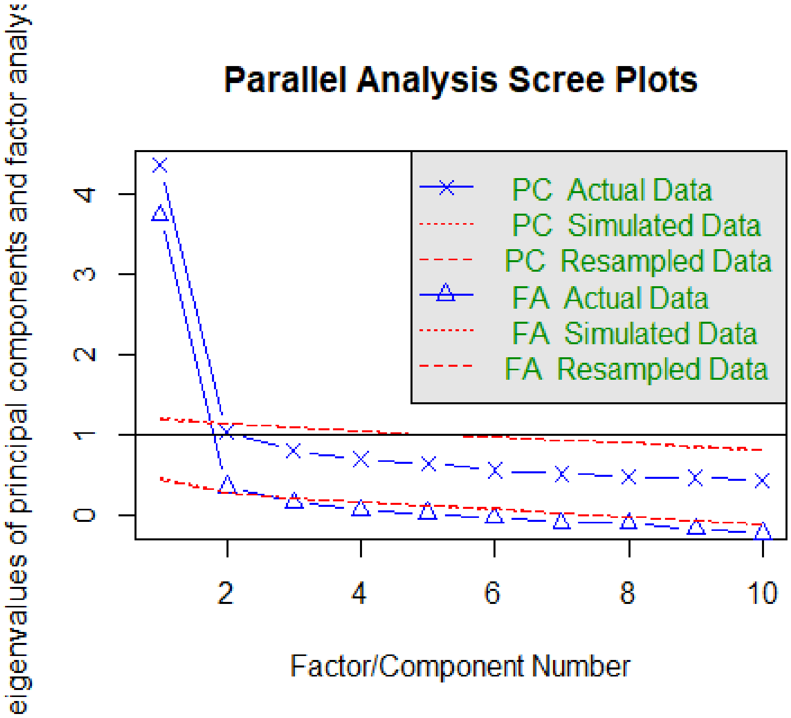
Preliminary Analysis

Bartlett's test of sphericity, used to test the overall significance of the correlations within the matrix, was significant (χ^2 (45) = 1825.87, p <.01), indicating that it was appropriate to use factor analysis on this dataset. The factorability of the ten variables was examined before completing the factor analysis. The Kaiser-Meyer-Olkin (KMO) overall measure of sampling adequacy (MSA) was 0.90. The Eigenvalues fell between .86 and .92, with Standards 1, 2, 3, 4, 7, and 8 having values above .90. The Eigenvalues for Standards 5, 6, 9, and 10 fell between .86 and .89. The determinate of the correlation matrix, or volume of space occupied by the group of data points, was .0002, which was an acceptable value because it is greater than the recommended .00001.

Initially, based on preliminary statistics, a three-factor model was planned for the analysis. According to Kaiser's Criterion, the use of three factors was appropriate as indicated by Eigenvalues greater than 1.0 for three components. These values also met Joliffe's Criterion with values greater than .70. As noted in Figure 26, Catell's Scree Plot also indicated the use of three factors as evidenced by the leveling off of the Eigenvalues. Other indices were also used to determine the appropriate number of factors. Velicer's Map indicated the use of one factor. BIC achieved a minimum with three factors present. The sample size adjusted BIC achieves a minimum of -28.88 with three factors. Parallel analysis indicated two factors present. Based on the preponderance of the evidence, three factors were appropriate. The root mean square residual value for the three-factor model of .03 aided in concluding how well the model fits. However, due to the Heywood Factor, the three-factor decision was deemed inappropriate for the study. Hence, it was concluded that two-factor loading was necessary for the factor analysis. The root mean square residual value for the two-factor model was .04. The degrees of freedom were 45, and the objective function was .93. Based on these values, it was again confirmed that using a two-factor model for factor analysis was appropriate.

Figure 26

Scree Plot for Fourth Grade ELA



Primary Analysis

Table 35 provides the summary statistics of the factor for the two-factor model. As represented in the following table, TKES Standards 1, 2, 3, 4, 5, 6, and 8 loaded on the first component. Standards 7, 9, and 10 loaded on the second component. All loadings on factor 1 exceeded .70 except Standard 2 (.58) and Standard 8 (.68). On Factor 2, Standards 9 and 10's values exceeded .80 (.84 and .86, respectively). Standard 7's value was the lowest of any standard on either factor (0.55).

Table 35

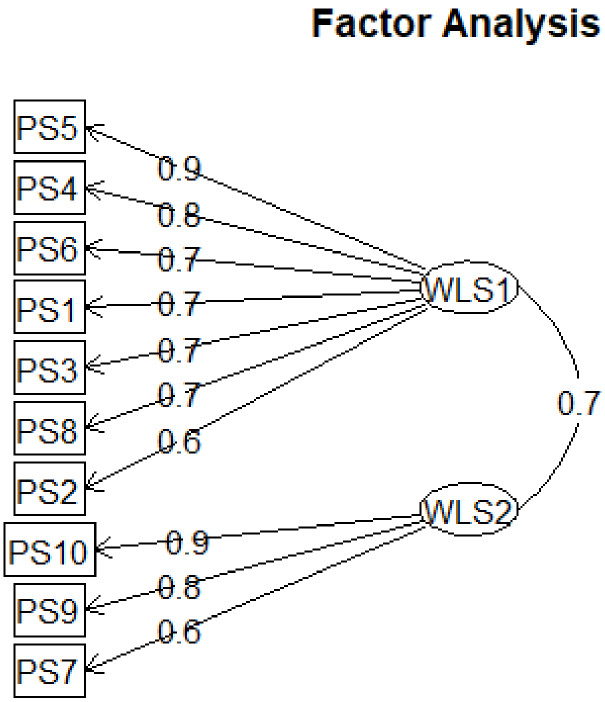
Summary of Standard Loadings for Two-factor Component for Fourth Grade ELA

	WLS1	WLS2	h2	U2	Com	Component
PS 1	**0.72**	0.09	0.61	0.39	1.00	1
PS 2	**0.58**	0.23	0.58	0.42	1.30	1
PS 3	**0.71**	0.23	0.78	0.22	1.20	1
PS 4	**0.77**	0.11	0.73	0.27	1.00	1
PS 5	**0.94**	-0.26	0.62	0.38	1.10	1
PS 6	**0.74**	0.10	0.67	0.33	1.00	1
PS 7	0.33	**0.55**	0.67	0.33	1.60	2
PS 8	**0.68**	0.21	0.71	0.29	1.20	1
PS 9	0.05	**0.84**	0.76	0.24	1.00	2
PS 10	0.05	**0.86**	0.81	0.19	1.00	2

Note: *WLS1* and *WLS2* are the factors on which the standards load. *h2* is the variance in the item explained by the retained factors. The *u2* column provides the residual variance or uniqueness. The *Com* column provides Hoffman's Index of Complexity for each standard. PS 1 (Professional Standard 1), PS 2 (Professional Standard 2), PS 3 (Professional Standard 3), PS 4 (Professional Standard 4), PS 5 (Professional Standard 5), PS 6 (Professional Standard 6), PS 7 (Professional Standard 7), PS 8 (Professional Standard 8), PS 9 (Professional Standard 9), PS 10 (Professional Standard 10).

Figure 27

Two Interpretable Components for Fourth Grade ELA



Through the factor analysis, two latent variables were created. The first variable included the following standards: Standards 1 (Professional Knowledge), 2 (Instructional Planning), 3 (Instructional Strategies), 4 (Differentiation), 5 (Assessment Strategies), 6 (Assessment Uses), and 8 (Academically Challenging Environment). The second variable included Standards 7 (Positive Learning Environment), 9 (Professionalism), and 10 (Communication). The standards included in the first variable are centered around content and instruction. For example, indicators within these standards include the teacher's knowledge of the subject matter, planning and delivery of research-based strategies based on differentiated learning needs, providing summative and formative instruction and redelivering content as necessary, and maintaining rigorous instruction. The standards included in the second variable involve relationships, lifelong learning, and communication. These standards require equitable learning opportunities, commitment to the students and profession, constant aspiration to learn and grow professionally, and positive and clear communication with students, staff members, and stakeholders.

Fourth Grade Math

b. How many reliable and interpretable components are there among the following variables: Professional Knowledge, Instructional Planning, Instructional Strategies, Differentiated Instruction, Assessment Strategies, Assessment Uses, Positive Learning Environment, Academically Challenging Environment, Professionalism, and Communication among fourth grade math teachers?

Descriptive Statistics

The analysis for this research question utilized TKES Standard score data from 600 fourth grade math teachers. Only teachers who exclusively taught math were included so that the results would be specific to subjects and grades. As indicated in Table 36, most scores for each standard fell in the proficient range. Like fourth grade ELA, Standard 5 had the largest percentage of scores (88.17%) in the proficient area. Again, as with fourth grade ELA, Standards 7 and 9 had the highest percentages in the exemplary area, 36.00% and 37.50%, respectively. Standards 4, 5, and 6 had the smallest percentage of scores in the exemplary level, with larger percentages than other standards falling in the ineffective/needs improvement and proficient ranges. Standard 8, again as fourth grade ELA, had the highest percentage (2.50%) of teachers scoring in the ineffective/needs improvement range. Standards 4, 5, and 6 had over 80% of teachers scoring proficiently.

Table 36

Number and Percentage of Fourth Grade Math Teachers Scoring at Ineffective/Needs Improvement, Proficient, or Exemplary by Professional Standard

TKES Categories	Ineffective/ Needs Improvement	Proficient	Exemplary
PS 1	5 (0.83%)	418 (69.67%)	177 (29.50%)
PS 2	11 (1.83%)	467 (77.83%)	122 (20.33%)
PS 3	11 (1.83%)	434 (72.33%)	155 (25.83%)
PS 4	12 (2.00%)	501 (83.50%)	87 (14.50%)
PS 5	6 (1.00%)	529 (88.17%)	65 (10.83%)
PS 6	4 (0.67%)	510 (85.00%)	86 (14.33%)
PS 7	9 (1.50%)	375 (62.50%)	216 (36.00%)
PS 8	15 (2.50%)	466 (77.67%)	119 (19.83%)
PS 9	11 (1.83%)	364 (60.67%)	225 (37.50%)
PS 10	8 (1.33%)	473 (78.83%)	119 (19.83%)

Note. PS 1 (Professional Standard 1), PS 2 (Professional Standard 2), PS 3 (Professional Standard 3), PS 4 (Professional Standard 4), PS 5 (Professional Standard 5), PS 6 (Professional Standard 6), PS 7 (Professional Standard 7), PS 8 (Professional Standard 8), PS 9 (Professional Standard 9), PS 10 (Professional Standard 10).

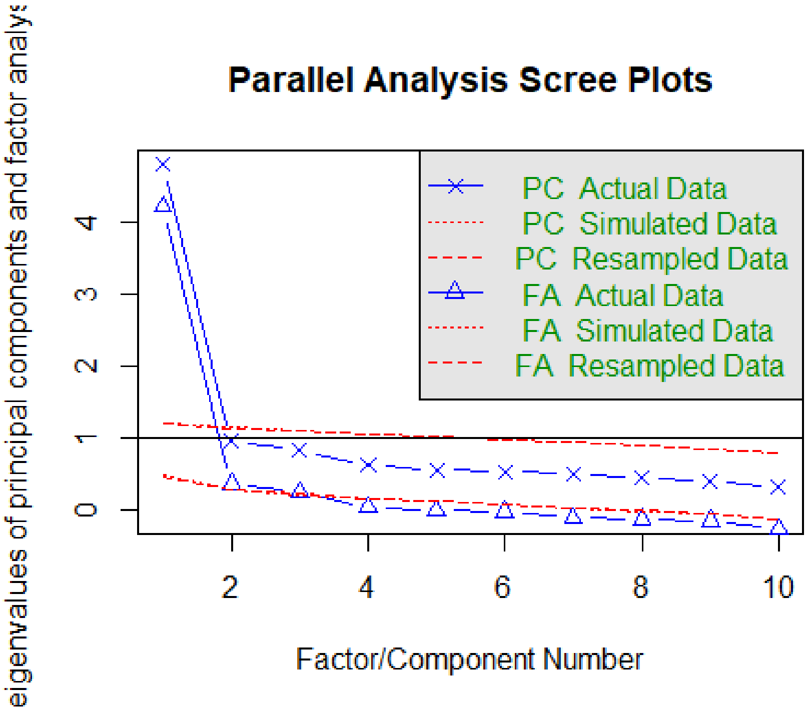
Preliminary Analysis

Bartlett's test of sphericity, used to test the overall significance of the correlations within the matrix, was significant (χ^2 (45) = 2276.16, p <.01), indicating that it was appropriate to use factor analysis on this dataset. The factorability of the ten variables was examined before completing the factor analysis. The Kaiser-Meyer-Olkin (KMO) overall measure of sampling adequacy (MSA) was 0.90. The Eigenvalues fell between .84 and .95, with Standards 1, 2, 3, 4, 7, 9, and 10 having values above .90. The Eigenvalues for Standards 5, 6, and 8 fell between .84 and .87. The determinate of the correlation matrix was .00003, again greater than the recommended .00001.

Based on preliminary statistics, a three-factor model was planned for the analysis. According to Kaiser's Criterion, the use of three factors was appropriate as indicated by Eigenvalues greater than 1.0 forthree components. These values also met Joliffe's Criterion with values greater than .70. As noted in Figure 28, Catell's Scree Plot also indicated the use of three factors as evidenced by the leveling off of the Eigenvalues. Other indices were also used to determine the appropriate number of factors. Velicer's Map indicated the use of one factor. BIC achieved a minimum of 81.14 with three factors present. The sample size adjusted BIC achieves a minimum of -25.17 with four factors. Parallel analysis indicated three factors present. The root mean square residual value for the three-factor model of .02 aided in concluding how well the model fits. The degrees of freedom were 18, and the objective function was .59. Based on the preponderance of the evidence, three factors were appropriate.

Figure 28

Scree Plot for Fourth Grade Math



Primary Analysis

Table 37 provides the summary statistics of the factor for the three-factor model. As represented in this table, TKES Standards 1, 5, and 6 loaded on the first component. Standards 3, 4, and 8 loaded on the second component. Standards 2, 7, 9, and 10 loaded on the third component. Loadings on factor 1 that exceeded .70 include Standard 5 (.74) and Standard 6 (1.02). Standard 1's had the lowest value on factor 1 (.34). For factor 2, Standard 8 had the highest value (1.01), and Standard 4 had the lowest value (.51). Standard 10 had the highest loading value on factor 3 (.93). Standard 2 had the lowest loading value on factor 3 (.35).

Table 37

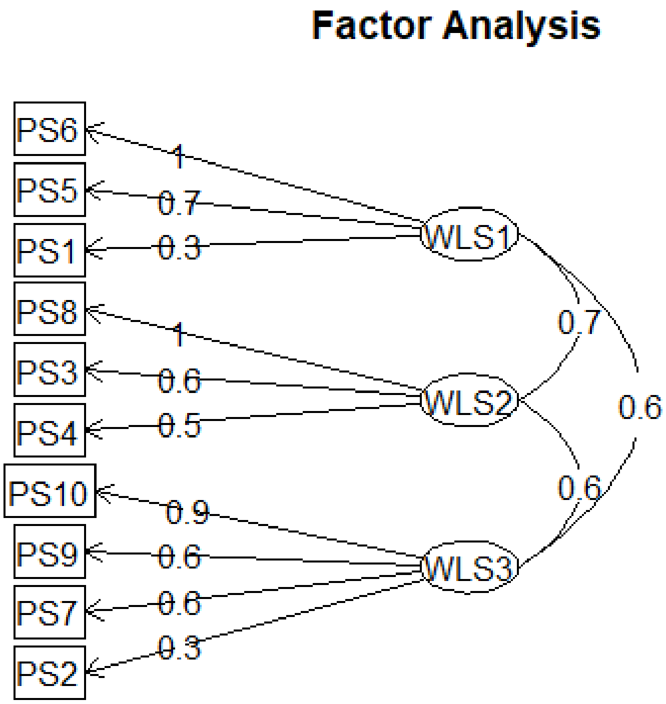
Summary of Standard Loadings for Three-factor Component for Fourth Grade Math

	WLS1	WLS2	WLS3	h2	U2	Com	Component
PS 1	**0.34**	0.32	0.25	0.64	0.36	2.80	1
PS 2	0.33	0.28	**0.35**	0.71	0.29	2.90	3
PS 3	0.05	**0.64**	0.26	0.75	0.25	1.30	2
PS 4	0.39	**0.51**	0.07	0.76	0.24	1.90	2
PS 5	**0.74**	0.12	0.12	0.83	0.17	1.10	1
PS 6	**1.02**	-0.01	-0.03	0.99	0.01	1.00	1
PS 7	-0.03	0.38	**0.56**	0.69	0.31	1.80	3
PS 8	0.02	**1.01**	-0.05	0.99	0.01	1.00	2
PS 9	0.33	-0.09	**0.61**	0.64	0.36	1.60	3
PS 10	-0.01	0.00	**0.93**	0.85	0.15	1.00	3

Note: *WLS1* and *WLS2* are the factors on which the standards load. *h2* is the variance in the item explained by the retained factors. The *u2* column provides the residual variance or uniqueness. The *Com* column provides Hoffman's Index of Complexity for each standard. (Professional Standard 1), PS 2 (Professional Standard 2), PS 3 (Professional Standard 3), PS 4 (Professional Standard 4), PS 5 (Professional Standard 5), PS 6 (Professional Standard 6), PS 7 (Professional Standard 7), PS 8 (Professional Standard 8), PS 9 (Professional Standard 9), PS 10 (Professional Standard 10).

Figure 29

Three Interpretable Components for Fourth Grade Math



Through the factor analysis, 3 latent variables were created. The first variable included the following standards: Standards 1 (Professional Knowledge), 5 (Assessment Strategies), and 6 (Assessment Uses). The second variable included Standards 3 (Instructional Strategies), 4 (Differentiation), and 8 (Academically Challenging Environment). The third variable included Standards 2 (Instructional Planning), 7 (Positive Learning Environment, 9 (Professionalism), and 10 (Communication). The standards contained in the first variable are based on professional knowledge, assessment strategies, and assessment uses with a focus on student goal setting and acquisition of goals. Indicators for these standards include assessing learning and using the assessment data obtained to deliver content, set learning goals, and monitor progress. The second variable consists of the effective delivery of differentiated content. Indicators in these standards include using research-based strategies to provide content based on specific learning needs to provide an individualized, rigorous learning environment. The third variable includes preparation for learning, providing a positive, professional environment with clear communication. Indicators within the third variable encompass an overall commitment to the profession through quality lesson planning, life-long learning, and continuous communication with students and other stakeholders.

Fifth Grade ELA

c. How many reliable and interpretable components are there among the following variables: Professional Knowledge, Instructional Planning, Instructional Strategies, Differentiated Instruction, Assessment Strategies, Assessment Uses, Positive Learning Environment, Academically Challenging Environment, Professionalism, and Communication among fifth grade English Language Arts teachers?

Descriptive Statistics

The analysis for this research question utilized TKES Standard score data from 600 fifth grade ELA teachers. Only teachers who exclusively taught ELA were included so that the results would be specific to subjects and grades. As indicated in Table 38, most scores for each standard fell in the proficient range. Again, Standard 5 had the largest percentage of scores (92.50%) in the proficient area. Similar to the other subjects and grades, Standards 7 and 9 had the highest percentages in the exemplary area, 36.83% and 43.50%, respectively. Standards 4 and 5 had the smallest percentage of scores in the exemplary level. Standards 2, 7, and 8 had the highest percentage (1.67%) of teachers scoring in the ineffective/needs improvement range. Standards 4, 5, 6, and 8 had over 80% of teachers scoring in the proficient range.

Table 38

Number and Percentage of Fifth Grade ELA Teachers Scoring at Ineffective/Needs Improvement, Proficient, or Exemplary by Professional Standard

TKES Categories	Ineffective/ Needs Improvement	Proficient	Exemplary
PS 1	2 (0.33%)	423 (70.50%)	175 (29.17%)
PS 2	7 (1.67%)	479 (79.83%)	114 (19.00%)
PS 3	2 (0.33%)	432 (72.00%)	166 (27.67%)
PS 4	9 (1.50%)	520 (86.67%)	71 (11.83%)
PS 5	2 (0.33%)	555 (92.50%)	43 (7.17%)
PS 6	6 (1.00%)	519 (86.50%)	75 (12.50%)
PS 7	7 (1.67%)	372 (62.00%)	221 (36.83%)
PS 8	10 (1.67%)	480 (80.00%)	110 (18.33%)
PS 9	5 (0.83%)	334 (55.67%)	261 (43.50%)
PS 10	3 (0.50%)	470 (78.33%)	127 (21.17%)

Note. PS 1 (Professional Standard 1), PS 2 (Professional Standard 2), PS 3 (Professional Standard 3), PS 4 (Professional Standard 4), PS 5 (Professional Standard 5), PS 6 (Professional Standard 6), PS 7 (Professional Standard 7), PS 8 (Professional Standard 8), PS 9 (Professional Standard 9), PS 10 (Professional Standard 10).

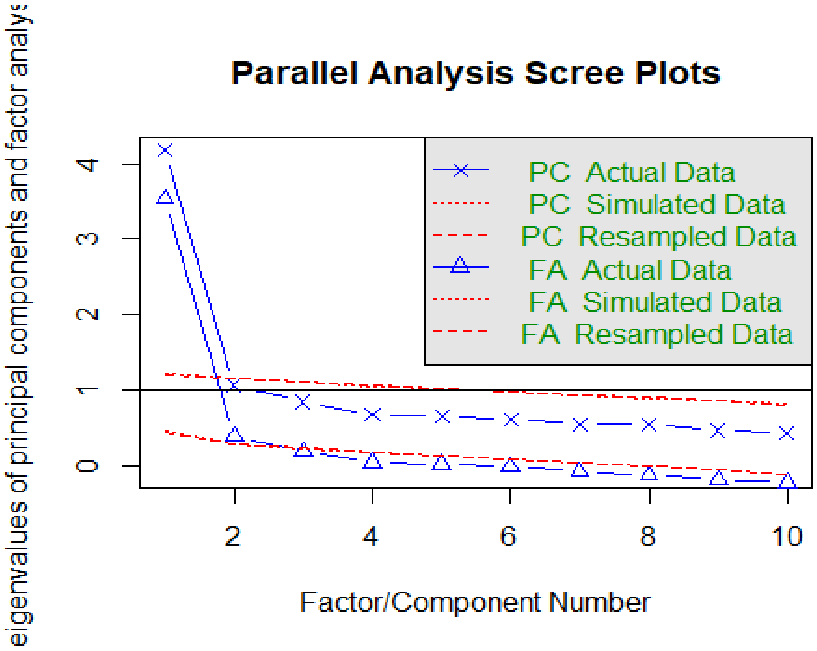
Preliminary Analysis

Bartlett's test of sphericity, used to test the overall significance of the correlations within the matrix, was significant (χ^2 (45) = 1677.26, p <.0001), indicating that it was appropriate to use factor analysis on this dataset. The factorability of the ten variables was examined before completing the factor analysis. The Kaiser-Meyer-Olkin (KMO) overall measure of sampling adequacy (MSA) was 0.89. The Eigenvalues fell between .86 and .93, with Standards 1, 2, 4, and 10 having values above .90. The Eigenvalues for Standards 3, 5, 6, 7, 8, and 9 fell between .86 and .89. The determinate of the correlation matrix, or volume of space occupied by the group of data points, was .0004, which was an acceptable value because it is greater than the recommended .00001.

A three-factor model was planned for the analysis based on initial findings. According to Kaiser's Criterion, the use of three factors was appropriate as indicated by Eigenvalues greater than 1.0 for three components. These values also met Joliffe's Criterion with values greater than .70. As noted in Figure 30, Catell's Scree Plot also indicated the use of three factors as evidenced by the leveling off of the Eigenvalues. Other indices were also used to determine the appropriate number of factors. Velicer's Map indicated the use of one factor. BIC achieved a minimum with three factors present. The sample size adjusted BIC achieves a minimum of -20.72 with four factors. Parallel analysis indicated two factors present. Based on the preponderance of the evidence, three factors were appropriate. The root mean square residual value for the three-factor model of .03 aided in concluding how well the model fits. The degrees of freedom were 18, and the objective function was .83. Based on the preponderance of the evidence, a three-factor model was chosen.

Figure 30

Scree Plot for Fifth Grade ELA



Primary Analysis

Table 39 provides the summary statistics of the factor for the three-factor model. TKES Standards 1, 2, 4, 5, and 6 loaded on the first component. Standards 9 and 10 loaded on the second component, and Standards 3, 7, and 8 loaded on the third component. All loadings on factor 1 exceeded .50 except Standard 2 (.42). Standard five had the most significant loading value (.96). On factor 2, both values exceeded .70. Standard 3 had the highest loading value (.84) on factor 3, with Standard 7 and 8's values falling below .60 (.48 and .58, respectively.)

Table 39

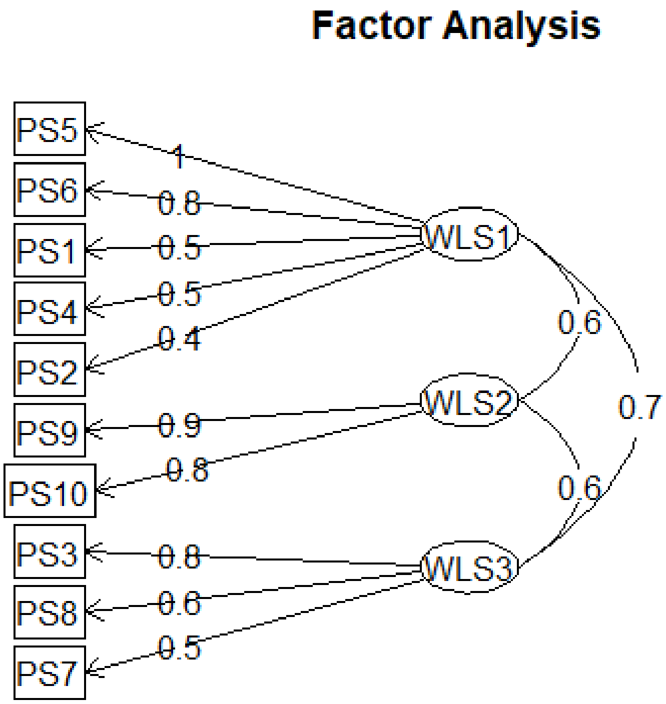
Summary of Standard Loadings for Three-factor Component for Fifth Grade ELA

	WLS1	WLS2	WLS3	h2	U2	Com	Component
PS 1	**0.54**	0.09	0.28	0.67	0.33	1.60	1
PS 2	**0.42**	0.18	0.30	0.63	0.37	2.20	1
PS 3	0.00	0.10	**0.84**	0.82	0.18	1.00	3
PS 4	**0.52**	-0.03	0.35	0.61	0.39	1.80	1
PS 5	**0.96**	0.04	-0.07	0.89	0.11	1.00	1
PS 6	**0.77**	0.07	0.05	0.72	0.28	1.00	1
PS 7	0.00	0.42	**0.48**	0.66	0.34	2.00	3
PS 8	0.35	-0.06	**0.58**	0.68	0.32	1.70	3
PS 9	-0.05	**0.86**	0.04	0.73	0.27	1.00	2
PS 10	0.16	**0.78**	-0.04	0.74	0.26	1.10	2

Note: *WLS1* and *WLS2* are the factors on which the standards load. *h2* is the variance in the item explained by the retained factors. The *u2* column provides the residual variance or uniqueness. The *Com* column provides Hoffman's Index of Complexity for each standard. PS 1 (Professional Standard 1), PS 2 (Professional Standard 2), PS 3 (Professional Standard 3), PS 4 (Professional Standard 4), PS 5 (Professional Standard 5), PS 6 (Professional Standard 6), PS 7 (Professional Standard 7), PS 8 (Professional Standard 8), PS 9 (Professional Standard 9), PS 10 (Professional Standard 10).

Figure 31

Three Interpretable Components for Fifth Grade ELA



Through the factor analysis, 3 latent variables were created. The standards included in the first variable are Standard 1 (Professional Knowledge), Standard 2 (Instructional Planning), Standard 4 (Differentiated Instruction), Standard 5 (Assessment Planning), and Standard 6 (Assessment Uses). The second variable includes Standards 9 (Professionalism) and 10 (Communication). The third variable includes Standards 3 (Instructional Strategies), 8 (Challenging Learning Environment), and 7 (Positive Learning Environment). The standards in the first variable focus on knowing and understanding student progress through the planning and assessment of student needs. These standards indicate the teacher's knowledge of the subject matter, planning individualized instruction, providing summative and formative instruction, and redelivering content as necessary to meet individual learning needs. The standards included in the second variable are professionalism and communication. To master these standards included in the second variable, teachers must maintain a commitment to the profession and serve their colleagues and communicate effectively to support mutual students. The third variable created requires teachers to deliver content using research-based strategies in a positive, nurturing environment that challenges all students at their levels.

Fifth Grade Math

d. How many reliable and interpretable components are there among the following variables: Professional Knowledge, Instructional Planning, Instructional Strategies, Differentiated Instruction, Assessment Strategies, Assessment Uses, Positive Learning Environment, Academically Challenging Environment, Professionalism, and Communication among fifth grade math teachers?

Descriptive Statistics

The analysis for this research question utilized TKES Standard score data from 600 fifth grade Math teachers. Only teachers who exclusively taught Math were included so that the results would be specific to subjects and grades. As indicated in Table 40, most scores for each standard fell in the proficient range. Standard 5 had the largest percentage of scores (86.67%) in the proficient area. Standards 7 and 9 had the highest percentages in the exemplary area, 36.67% and 40.00%, respectively. Standards 4, 5, and 6 had the smallest percentage of scores in the exemplary level, with larger percentages than other standards falling in the ineffective/needs improvement and proficient ranges. Standard 2 had the highest percentage (2.83%) of teachers scoring in the ineffective/needs improvement range. Standards 4, 5, and 6 had over 80% of teachers scoring in the proficient range.

Table 40

Number and Percentage of Fifth Grade Math Teachers Scoring at Ineffective/Needs Improvement, Proficient, or Exemplary by Professional Standard

TKES Categories	Ineffective/ Needs Improvement	Proficient	Exemplary
PS 1	7 (1.17%)	384 (64.00%)	209 (34.83%)
PS 2	17 (2.83%)	462 (77.00%)	121 (20.17%)
PS 3	15 (2.50%)	416 (69.33%)	169 (28.17%)
PS 4	10 (1.67%)	502 (83.67%)	88 (14.67%)
PS 5	4 (0.67%)	520 (86.67%)	76 (12.67%)
PS 6	11 (1.83%)	502 (83.67%)	87 (14.50%)
PS 7	11 (1.83%)	369 (61.50%)	220 (36.67%)
PS 8	14 (2.33%)	447 (74.50%)	139 (23.17%)
PS 9	13 (2.17%)	347 (57.83%)	240 (40.00%)
PS 10	4 (0.67%)	468 (78.00%)	128 (21.33%)

Note. PS 1 (Professional Standard 1), PS 2 (Professional Standard 2), PS 3 (Professional Standard 3), PS 4 (Professional Standard 4), PS 5 (Professional Standard 5), PS 6 (Professional Standard 6), PS 7 (Professional Standard 7), PS 8 (Professional Standard 8), PS 9 (Professional Standard 9), PS 10 (Professional Standard 10).

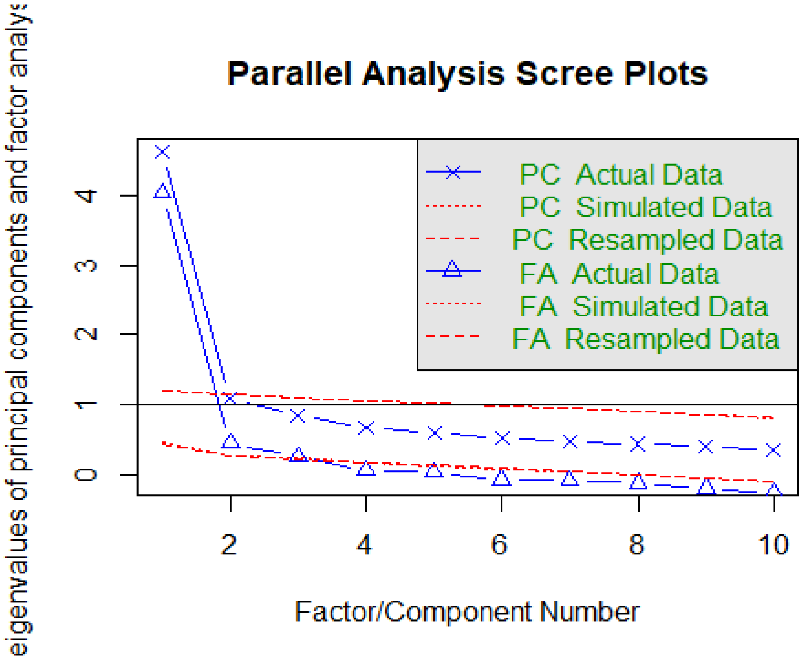
Preliminary Analysis

Bartlett's test of sphericity, used to test the overall significance of the correlations within the matrix, was significant (χ^2 (45) = 2179.53, p <.01), indicating that it was appropriate to use factor analysis on this dataset. The factorability of the ten variables was examined before completing the factor analysis. The Kaiser-Meyer-Olkin (KMO) overall measure of sampling adequacy (MSA) was 0.90. The Eigenvalues fell between .84 and .93, with Standards 1, 2, 3, 4, 7, and 8 having values above .90. The Eigenvalues for Standards 5, 6, 9, and 10 fell between .86 and .89. The determinate of the correlation matrix, or volume of space occupied by the group of data points, was .00009, which was an acceptable value because it is greater than the recommended .00001.

A three-factor model was planned for the analysis according to preliminary statistics. According to Kaiser's Criterion, the use of three factors was appropriate as indicated by Eigenvalues greater than 1.0 for three components. These values also met Joliffe's Criterion with values greater than .70. As noted in Figure 32, Catell's Scree Plot also indicated the use of three factors as evidenced by the leveling off of the Eigenvalues. Other indices were also used to determine the appropriate number of factors. Velicer's Map indicated the use of one factor. BIC achieved a minimum with three factors present. The sample size adjusted BIC achieves a minimum of -20.15 with four factors. Parallel analysis indicated three factors present. Based on the preponderance of the evidence, three factors were appropriate. The root mean square residual value for the three-factor model of .02 aided in concluding how well the model fits. The degrees of freedom were 18, and the objective function was .48. Based on these values, it was confirmed that using a three-factor model for factor analysis was appropriate.

Figure 32

Scree Plot for Fifth Grade Math



Primary Analysis

Table 41 provides the summary statistics of the factor for the three-factor model. TKES Standards 1, 2, 3, 7, and 8 loaded on the first component. Standards 9 and 10 loaded on the second component. Standards 4, 5, and 6 loaded on the third component. All loadings on factor 1 exceeded .70 except Standard 2 (.44). Both loadings on factor 2 exceeded .65. On factor 3, all loadings exceeded .70 except Standard 4 (.46).

Table 41

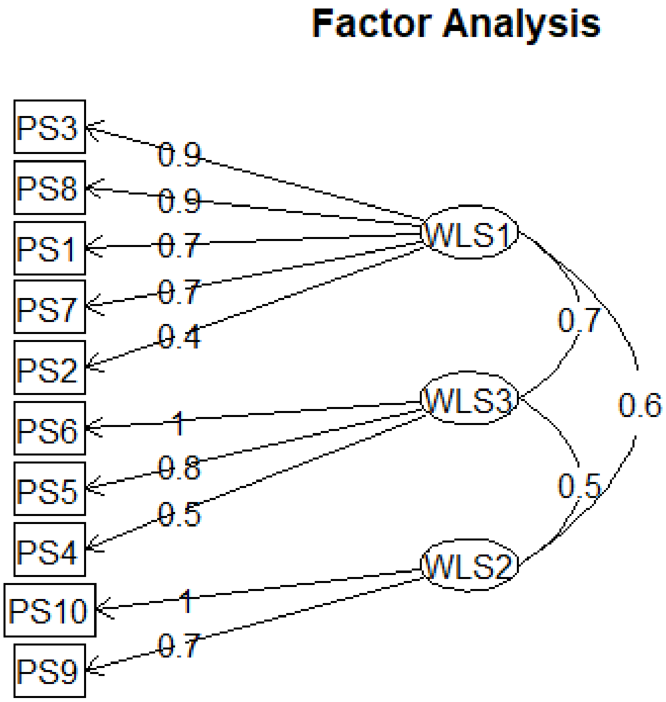
Summary of Standard Loadings for Three-factor Component for Fifth Grade Math

	WLS1	WLS2	WLS3	h2	U2	Com	Component
PS 1	**0.75**	0.10	0.06	0.74	0.27	1.00	1
PS 2	**0.44**	0.26	0.25	0.70	0.30	2.30	1
PS 3	**0.93**	-0.07	0.04	0.85	0.15	1.00	1
PS 4	0.22	0.23	**0.46**	0.61	0.39	1.90	3
PS 5	0.14	0.06	**0.77**	0.82	0.18	1.10	3
PS 6	-0.03	0.00	**1.00**	0.95	0.05	1.00	3
PS 7	**0.79**	0.29	-0.18	0.64	0.36	1.50	1
PS 8	**0.88**	-0.07	0.10	0.83	0.17	1.00	1
PS 9	0.14	**0.65**	0.08	0.64	0.36	1.10	2
PS 10	-0.03	**0.97**	0.04	0.93	0.07	1.00	2

Note: *WLS1* and *WLS2* are the factors on which the standards load. *h2* is the variance in the item explained by the retained factors. The *u2* column provides the residual variance or uniqueness. The *Com* column provides Hoffman's Index of Complexity for each standard. PS1 (Professional Standard 1), PS 2 (Professional Standard 2), PS 3 (Professional Standard 3), PS 4 (Professional Standard 4), PS 5 (Professional Standard 5), PS 6 (Professional Standard 6), PS 7 (Professional Standard 7), PS 8 (Professional Standard 8), PS 9 (Professional Standard 9), PS 10 (Professional Standard 10).

Figure 33

Three Interpretable Components for Fifth Grade Math



Through the factor analysis, three latent variables were created. The first variable included the following standards: Standards 1 (Professional Knowledge), 2 (Instructional Planning), 3 (Instructional Strategies), Standard 7 (Positive Learning Environment), and 8 (Academically Challenging Environment). The second variable included Standards 9 (Professionalism) and 10 (Communication). The third variable included 4 (Differentiation), 5 (Assessment Strategies), and 6 (Assessment Uses). The standards included in the first variable are based on professional knowledge, lesson development, rigorous content delivery, and interaction with the student. The standards included in the second variable focus on an overall commitment to education. These standards highlight the teacher's dedication to the profession, coupled with positive and clear communication with students, other staff members, and stakeholders. The third variable includes understanding academic progress through assessment planning and providing instruction at the level needed.

Summary

The results reported in this chapter were presented in alignment with a non-experimental, quantitative design. The study determined the extent to which TKES scores are predictive of student growth and student achievement as measured by the GMAS. Additionally, the study examined differences in student achievement based on academic setting (self-contained or departmentalized) and levels of the percentage of economically disadvantaged (ED). The reliability and interpretability of the TKES components was also examined. There was no manipulation of data.

The independent variables included the teachers' summative scores on each of the 10 TKES Standards (the TAPS portion of the TKES evaluation), levels of the percentage of students identified as ED, and academic setting (self-contained or departmentalized). The TKES Standards are as follows: Professional Knowledge, Instructional Planning, Instructional Strategies, Differentiated Instruction, Assessment Strategies, Assessment Uses, Positive Learning Environment, Academically Challenging Environment, Academically Challenging Environment, Professionalism, and Communication. The levels of the percent of ED were reported as a percentage between 0 and 100. The levels were divided into two levels based on the mean and were considered ordinal data. The academic setting, regarded as nominal data, was categorized as self-contained or departmentalized. Self-contained teachers provide instruction on all subjects to the same children without class changes. In a departmentalized setting, teachers only provide instruction on specific subjects to students.

The dependent variables for the first research question were levels of student growth percentiles. The levels of growth percentiles were measured as ordinal data. The levels of student growth were low growth (1-34), typical growth (35-65), and high growth (66-99). The GaDOE calculates growth percentiles for fourth and fifth grade educators based on their students' growth from the previous year's score on the GMAS (A Guide to the Georgia Student Growth Model, 2018). The teacher's mean scale scores (MSS) on the GMAS served as the dependent variable in question 2 and question 3. The teacher's MSS is an average of the students' scale scores. The MSS served as the dependent variable. These data were considered to be on the interval level of measurement.

For research question 1, an ordinal regression model was used to determine the predictive power of TKES scores on SGP's. A multiple regression analysis was used in question 2 to determine the predictive power of TKES scores on the GMAS MSS. Research question 3 employed a factorial ANOVA to determine if there was a significant difference in the academic setting (departmentalized or self-contained) by the level of ED students on the GMAS MSS. A factor analysis was used in question 4 to determine the number of reliable components among the TKES Standards.

Primarily, the results of this study revealed inconsistencies between grade levels and subjects in the ability of TKES Standards to predict student growth. While no teacher scores on TKES Standards were predictive of student growth in fifth grade ELA or math, TKES scores had predictive power on student growth in fourth grade ELA and math. In fourth grade ELA, Standard 4 at the proficient and exemplary level and Standard 8 at the exemplary level were found to have a positive impact on student growth. Standard 9 at the proficient level and Standard 3 at the proficient and exemplary level were found to have a negative impact on student growth. In fourth grade math, Standards 1 and 3 at the exemplary level were found to have a positive impact on student growth, and no standards were noted as negatively impacting SGP's.

Regarding TKES scores and student achievement, significance was found in both grade levels and subjects studied. In fourth grade ELA, Standards 3, 7, 8, and 10 at the exemplary level were significant predictors of student achievement as indicated by the GMAS MSS. Standard 8 at the needs improvement level was a significant predictor of having a negative impact on scale scores. Standards 1, 2, 7, and 8 at the exemplary level were found to have a positive impact on GMAS MSS in fourth grade math. In fifth grade ELA, Standards 7 (needs development and exemplary), 8 (exemplary), and 9 (exemplary) were found to have a positive impact on student achievement. In fifth grade math, Standards 1, 7, 8, and 9 at the exemplary level were found to have an impact on achievement.

Further, in question 3, there was a significant difference in student achievement between the levels of the percentage of ED students in fourth and fifth grade ELA and math. Students in schools with a higher level of ED students had significantly lower achievement scores. ED and setting were found significant only in fourth grade math. Students in departmentalized settings scored higher than students in self-contained settings.

When interpreting components for TKES Standards, two or three latent variables were created. In fourth grade ELA, two latent variables were created, including variable 1 (Standards 1, 2, 3, 4, 5, 6, and 8) variable 2 (Standards 7, 9, and 10). In fourth grade math, three variables were created, including variable 1 (Standards 1, 5, and 6), variable 2 (Standards 3, 4, and 8), and variable 3 (Standards 2, 7, 9, and 10). In fifth grade ELA, three variables were created, including variable 1 (Standards 1, 2, 4, 5, and 6), variable 2 (Standards 9 and 10), and variable 3 (Standards

3, 7, and 8). In fifth grade math, three variables were created, including variable 1 (Standards 1, 2, 3, 7, and 8), variable 2 (Standards 9 and 10), and variable 3 (Standards 4, 5, and 6).

Chapter V

INTERPRETATIONS, CONCLUSIONS, AND RECOMMENDATIONS

Introduction

As the needs of students change, educators must continually grow in their ability to educate students. Educational leaders believe an effective evaluative tool may catalyze teacher growth (Wise et al., 1984). While developing and sustaining a reliable evaluation tool has been a source of debate for many years, educational leaders caution against a tool that fails to address research-based teaching practices and encourages professional growth (Marzano et al., 2011). A combination of models is necessary to assess teaching practice, including observations, teaching artifacts, and frequent walkthroughs that provide evidence of daily classroom performance (Milanowski, 2011).

Educational leaders in Georgia believe they have created a valuable evaluative system. The Georgia Teacher Keys Effectiveness System was developed based on sound teaching practices (Georgia's Teacher Keys Effectiveness System Meaningful Feedback Professional Growth Flexibility to Innovate, 2018). Although measures indicate it to be a valid and reliable tool (Elder, Wang, & Cramer, 2015), the connection to student growth and achievement must be established for educators to confidently use evaluation results to make informed decisions that impact American schools (Marzano et al., 2011).

The overarching purpose of this study was to determine the predictive power of the TKES instrument on student growth and achievement. Additionally, the study determined the significance of levels of percentage of ED students and academic setting on student achievement. The study aided in understanding the number of factors present among the TKES Standards. Information gleaned from this study can assist administrators in understanding the role TKES plays in student growth and achievement. The information also provided insight into teacher behaviors on which to base professional development.

Literature Review

Since the beginning of formal education in the United States, educators have sought effective evaluative systems to improve the public education system (Barrette et al., 1995). As public education became more valued during the latter part of the nineteenth century and the early twentieth century, the need for more critical teacher evaluations intensified. If evaluated, teachers were typically assigned a grade from A to F for specific lessons, and administrators provided little to no constructive feedback (Marzano et al., 2011).

Following World War II, although administrative duties remained primarily managerial, many leaders recognized their role as instructional leaders. The shift in roles emphasized the need for a more clinical approach to teacher assessment. Scholars began to focus on forms of evaluation meant to increase teachers' pedagogical growth (Marzano et al., 2011). In 1969, Goldhammer published a five-phase process of supervision, including reflection (Goldhammer, 1969). Cogan (1973) also emphasized continual instructional improvement as the priority of teacher evaluations. Although scholars began to understand the need for evaluative systems that focused on teacher growth, widespread evaluative systems still failed to provide instructional growth and were only used in extreme cases for remediation and dismissal purposes (Weisburg et al., 2009).

In the 1980s, Madeline Hunter published a 7-step model that provided clear guidance for administrators, including supervisory conferences (Hunter, 1980.) Hunter initiated the use of professional development to create a common language for evaluations (Hosford, 1984). Glatthorn, McGreal, and Glickman further facilitated the transition to modern teacher evaluation by calling for a differentiated approach to teacher evaluations, including teacher input and professional development (Marzano et al., 2011). Glickman (1985) confirmed the most crucial function of the teacher evaluation must be for instructional improvement.

Charlotte Danielson's work, in 1996, highlighted 76 elements of quality teaching and set the stage for teacher evaluation using four levels of performance (unsatisfactory, basic, proficient, and distinguished). Continuing the trend toward modern evaluative systems, Tucker and Stronge (2005) advocated for including student achievement with classroom observations when evaluating teachers. They believed student achievement was an integral component when considering teacher effectiveness.

Despite the further development of evaluative practices, *The Widget Effect*, a mixed-method study based on data from over 1,300 administrators and 15,000 teachers, heavily criticized teacher evaluation practices in the United States. The study highlighted ineffective evaluative practices, including failing to identify exceptional teachers. The report also indicated a consistent lack of feedback that promoted professional growth (Weisburg et al., 2009).

As educators continued to strive for an improvement in the evaluative process, debates consistently arose regarding the models by which teachers were evaluated. Examples of models include value-added models (VAMs), classroom observations, clinical models, i.e., Marzano's, Focused Teacher Evaluation Model (2020) or Danielson's Model (2014), or various combinations of these models. Researchers of the value-added model, or models driven by student growth on formal assessments, remained concerned that the model alone would not provide the complex information necessary for teachers to grow professionally. There were concerns that value-added measures were unstable and standardized test scores often fluctuated for reasons outside the scope of the teacher's control (Raudys, 2018). Baker, Barton, Darling-

Hammond, Haertel, Ladd, Linn, Ravitch, Rothstein, Shavelson, and Shepard (2010), suggested teacher effectiveness predicted only 4% to 16% of the variation in a teacher's ratings on formal assessments from one year to the next. Few teachers ranked as high performers maintained the ranking from year to year. Researchers also agreed value-added models lacked feedback (Baker et al., 2010).

Evaluations based solely on classroom observation, another frequently used evaluative model, are commonly used. Well-designed observation instruments included rubrics that promote consistent, reliable results and multiple raters (Milanowski, 2011; Raudys, 2018; Robinson, n.d.). During teacher observations, administrators observe teaching practices for a set period of time. Although not without merit, the disadvantages of observations included the administrator time required and observation of only a small part of the instructional day. Further, busy administrators sometimes fail to provide quality feedback (Raudys, 2018).

Other popular models emerged that attempted to capture the necessary elements of teacher evaluation. For example, Danielson's model was modified in 2013 (Danielson, 2014). The Marzano Focused Teacher Evaluation Model also became popular. Observers systematically scored numerous elements and domains. Researchers determined that observed elements were significantly correlated to student gains. According to Marzano, the observation scores were significant predictors of student growth on state assessments (The focused teacher evaluation model, 2020).

Milanowski (2011) believes a combination of systems are necessary to assess teaching practice, including observations, teaching artifacts, and frequent walkthroughs that provide evidence of daily classroom performance. A combination might provide the information necessary for optimal professional development (Robinson, n.d.). Milanowski (2011) suggested student assessment measures and measures of teacher behaviors are very different and should both be analyzed both separately and systematically to determine teacher effectiveness. The Georgia TKES model is an example of an evaluative model that combines aspects of multiple teacher evaluative models ("Georgia Teacher Keys Effectiveness System," n.d.).

The components of the evaluative system must be consistent and fair, but they must also lead teachers to the use of research-based teaching practices. If teacher evaluation's pervasive purpose is to increase student achievement, the most impactful strategies must be the assessment's driving force. Unfortunately, the debate over effective teaching practices only adds complexity to the debate over teacher evaluations (Marzano et al., 2011).

John Hattie (2008) combined thousands of theories and studies to develop a robust list of effective teaching practices. He maintained that teachers must become experts on effective strategies and implement them consistently. Hattie and his colleagues (2017) assigned a ranking to show the effect of each force on students. Any effect size greater than .4 was considered beneficial to students. Some influences were ranked with negative effect sizes, meaning students were negatively impacted or regressed. Most negative influences fell outside of the teacher's control, except for influences the teacher might potentially have mitigated, such as motivation, students feeling disliked, retention, boredom, and performance goals (Hattie, 2008). Examples of effective strategies included collective teacher efficacy, self-reported grades, teacher estimates of achievement, cognitive tasks analysis, and response to intervention (Hattie et al., 2017). Hattie and his colleagues (2017) also found teachers' attitudes toward themselves and their students significantly impacted student growth (Hattie et al., 2017).

Najimi and colleagues (2013) warned against ignoring contributing factors when evaluating teachers. The failure to consider those factors in teacher effectiveness could be irresponsible on the part of educational leaders. Although some factors hinge on school-based decisions and settings, many are beyond the school systems' control. Examples include the level of students deemed as E.D., attendance rate, parental support of students, etc. (Najimi et al., 2013).

There are many internal factors affecting teaching and learning. These include curricular decisions, program decisions, physical building environment, building leadership, staff collaboration, and staff development. Building administration may not have the ability to control these characteristics. However, some of the factors are within the scope of school administrators (Najimi et al., 2013). Classroom setting is an example of a factor that influences the learning environment within the scope of administrators' control. Many schools remain self-contained throughout the primary years, while others have departmentalized models. Whether the factors are internal or external, they can affect teaching and learning and must be considered when evaluating teachers (Najimi, Sharifirad, Amini, & Meftagh, 2013).

The Georgia Department of Education developed the Teacher Keys Effectiveness System (TKES) to standardize and strengthen the teacher effectiveness system in Georgia and bridge the gap between effective teaching practices and student achievement. Under the system, teachers, assistant principals, and principals receive an overall effectiveness rating annually. In agreement with Marzano's (2011) beliefs, the evaluation system is based on multiple measures prioritizing student growth and achievement (H.B. 244, 2014).

There are three primary components of TKES, including the Teacher Assessment on Performance Standards (TAPS), Student Growth, and Professional Growth. A Teacher Effectiveness Measure (TEM) is calculated based on scores within these components. TAPS provides evaluating administrators with rubrics on which to base teacher ratings as evidenced by teacher observations and other supporting artifacts. The TAPS portion of the TKES evaluation is 50% of the TEM score. The TKES Standards are as follows: Standard 1 (Professional Knowledge), Standard 2 (Instructional Planning), Standard 3 (Instructional Practices), Standard 4 (Differentiation), Standard 5 (Assessment Planning), Standard 6 (Assessment Uses), Standard 7 (Positive Learning Environment), Standard 8 (Academically Challenging Environment), Standard 9 (Professionalism), and Standard 10 (Communication) ("Georgia's Teacher Keys Effectiveness System," n.d).

The second component of TKES is the student growth measure, which accounts for 30 percent of the Teacher Effectiveness Measure (TEM) score. Student growth percentiles are an integral part of the TKES system (*A guide to the Georgia student growth model*, 2014). The final component, professional growth, is 20% of the TEM score. Professional growth is measured by the teacher completing the professional growth plan or goal. Teachers are assigned a plan or goal based on years of experience and success in the classroom. New teachers or teachers that need remediation more intensive plan while more experienced and successful teachers have flexibility in working toward the professional learning goals (Georgia's Teacher Keys Effectiveness System Meaningful Feedback Professional Growth Flexibility to Innovate, 2018). Based on these elements, teachers are provided an overall rating: exemplary, proficient, needs development, or ineffective (Georgia's Teacher Keys Effectiveness System Meaningful Feedback Professional Growth Flexibility to Innovate, 2018).

Methodology

The study employed a non-experimental, quantitative design to determine the extent to which TKES scores predict student growth and achievement. The study also examined the difference academic setting (self-contained or departmentalized) and levels of the percentages of ED students are expected to make in student achievement. Additionally, the reliability and interpretability of the TKES components were examined. No data manipulation was used for the dependent or independent variables.

The independent variables included the teachers' summative scores on each of the ten TKES Standards, levels of the percentage of students identified as ED, and academic setting (self-contained or departmentalized). The levels of ED were divided at the mean for each dataset into two groups. The academic setting was categorized as self-contained or departmentalized. The dependent variables for the first research question were levels of growth percentiles. The levels of student growth were low growth (1-34), typical growth (35-65), and high growth (66-99). Scale scores on the GMAS served as the dependent variable in questions two and three.

This study's target population was fourth and fifth grade English Language Arts and math teachers across Georgia. These teachers were responsible for either ELA or math instruction or both. A teacher's mean growth percentile is based on their students' growth percentiles. Data from the previous year were compared to the following year to determine growth to calculate student growth percentiles. Because students in Georgia do not take the Milestones Assessment until third grade, student growth percentiles cannot be calculated for students until fourth grade. Therefore, only teachers in the fourth and fifth grades have student growth percentiles with levels of performance (A guide to the Georgia Student Growth Model, 2014).

The accessible population included a random sample generated by the Georgia Department of Education of fourth and fifth grade English Language Arts and math teachers with student growth percentiles. The GaDOE delivered a random sampling of 4,000 teachers with student growth percentile and mean scale score data, categorized by grade and subject in fourth or fifth grade English Language Arts and math. The GaDOE provided the summative scores for each of the ten standards, the academic setting for each teacher (departmentalized or self-contained), and levels of the percentage of students identified as ED for each respective school.

RQ1: Are summative scores on TKES Standards (Professional Knowledge, Instructional Planning, Instructional Strategies, Differentiated Instruction, Assessment Strategies, Assessment Uses, Positive Learning Environment, Academically Challenging Environment, Professionalism, and Communication) significant predictors of the teacher's SGP level on the Georgia Milestones Assessment System (GMAS)? For research question 1, an ordinal regression model was used to determine the predictive power of TKES scores on SGP's.

RQ2: Are summative scores on TKES Standards (Professional Knowledge, Instructional Planning, Instructional Strategies, Differentiated Instruction, Assessment Strategies, Assessment Uses, Positive Learning Environment, Academically Challenging Environment, Professionalism, and Communication) significant predictors of the teacher's mean scale score on the Georgia Milestones Assessment System (GMAS)? A multiple regression analysis was used in question 2 to determine the predictive power of TKES scores on the GMAS MSS.

RQ3: Is there a significant difference in academic setting (departmentalized or self-contained) by level of economically disadvantaged (ED) students on the teacher's mean scale score on the Georgia Milestones Assessment System (GMAS)? Research question 3 employed a factorial ANOVA to determine if there was a significant difference in the academic setting (departmentalized or self-contained) by the level of ED students on the GMAS MSS.

RQ4: How many reliable and interpretable components are there among the following variables: Professional Knowledge, Instructional Planning, Instructional Strategies, Differentiated Instruction, Assessment Strategies, Assessment Uses, Positive Learning Environment, Academically Challenging Environment, Professionalism, and Communication? A factor analysis was used in question 4 to determine the number of reliable components among the TKES Standards.

Findings

The results of question 1 revealed inconsistencies between fourth and fifth grades for the ability of TKES Standards to predict student growth in fourth and fifth grades. In fourth grade ELA, Standard 4 (Differentiation) at the proficient and exemplary level and Standard 8 (Academically Challenging Environment) at the exemplary level were found to have a positive impact on student growth. Standard 9 (Professionalism) at the proficient level and Standard 3 (Instructional Practices) at the proficient and exemplary level negatively impacted student growth. In fourth grade math, Standard 1 (Professional Knowledge) and 3 (Instructional Planning) at the exemplary level were found to have a positive impact on student growth, while no standards were noted as negatively impacting SGP's. No TKES Standards predicted student growth in fifth grade ELA or math.

In the analysis for question 2, standards were noted as significant predictors of student achievement. In fourth grade ELA, the following standards at the exemplary level were found to have a significant predictive power of positively impacting achievement scores: Standard 3 (Instructional Planning), Standard 7 (Positive Learning Environment), Standard 8 (Academically

Challenging Environment) and Standard 10 (Communication). Standard 8 (Academically Challenging Environment) at the Needs Improvement Level was found to have a negative impact on achievement. In fourth grade math, the following standards, at the exemplary level, were found significant for a positive impact on student achievement: Standard 1 (Professional Knowledge), Standard 2 (Instructional Planning), Standard 7 (Positive Learning Environment), and Standard 8 (Academically Challenging Environment). No standards were found to have a negative impact on student achievement. In fifth grade ELA, Standards 7 (Positive Learning Environment) at the needs development and exemplary level, Standard 8 (Academically Challenging Environment) at the exemplary level, and Standard 9 (Professionalism) at the exemplary level were found to positively impact scores. In fifth grade math, Standard 1 (Professional Knowledge), Standard 7 (Positive Learning Environment), Standard 8 (Academically Challenging Environment), and Standard 9 (Professionalism) at the exemplary level were found to positively impact scores. No standards were found to have a negative impact.

In research question 3, while there was a significant difference in student achievement between the levels of the percentage of ED students across fourth and fifth grade ELA and math, there was no consistent outcome between setting and student achievement. The levels of ED and the GMAS MSS were significantly different in fourth and fifth grade ELA and math. However, in fourth grade math, there was also a significant difference in setting. Students in departmentalized settings were scored higher than students in self-contained settings.

Through the factor analysis, two or three latent variables were created and adequately reduced the dimensionality of the TKES Standards. In fourth grade ELA, two latent variables were created, including variable 1 (Standards 1, 2, 3, 4, 5, 6, and 8) variable 2 (Standards 7, 9, and 10). In fourth grade math, 3 variables were created, including variable 1 (Standards 1, 5, and

6), variable 2 (Standards 3, 4, and 8), and variable 3 (Standards 2, 7, 9, and 10). In fifth grade ELA, three variables were created, including variable one (Standards 1, 2, 4, 5, and 6), variable 2 (Standards 9 and 10), and variable three (Standards 3, 7, and 8). In fifth grade math, three variables were created, including variable one (Standards 1, 2, 3, 7, and 8), variable two (Standards 9 and 10), and variable three (Standards 4, 5, and 6).

Limitations and Assumptions

There were limitations of the study. The dataset was based on a random sampling provided by the Georgia Department of Education. Because the student growth percentiles are computed using the current and previous year's data, portions of the dataset obtained for this study were based on calculations that utilized the 2018 and the 2019 administration of the GMAS. Further, there was no GMAS in 2020 due to the COVID pandemic.

The overall subjectivity of the TKES ratings may also be considered a limitation. Although interrelated reliability training is frequently offered to Georgia educators, it is not required. Failure to continually participate in reliability practices may result in inconsistency among raters. However, leaders are required annually to recertify before beginning teacher observations.

The subjectivity of the administrator ratings may also be considered a limitation. Inconsistencies arise based on the administrator's views, experiences, etc. Based on these aspects of an administrator's views or experiences, discrepancies in ratings may be noted.

Factors outside of the teacher's locus of control may also affect the student's scores on the GMAS. For example, socioeconomics, attendance, student health, traumatic events in the child's life, and overall support for education at home may impact student scores. Additionally, parents may choose to refuse GMAS testing. Therefore, scores for those students are not available.

Students that do not test affect the overall mean for the GMAS. School factors, again outside of the educator's control, may also affect teacher scores, i.e., teachers may not have access to quality sources of professional development or instructional tools.

Suggestions for Future Research

Educators will continue to strive for increased knowledge and growth in their ability to educate students. Future research in this area will shape leaders' decisions in all aspects of teacher leadership. Ideas for future research include duplicating the parameters of this study to incorporate additional grade levels to examine the predictive ability of TKES scores on student growth and student achievement in the sixth, seventh, and eighth grades. Science and Social Studies could also be included as those subjects are added to tested subjects in the upper grades.

Another idea for future research includes studying the implications of teacher evaluation instruments from other states and comparing those results to this study. For example, are other teacher observation tools more or less predictive of student growth and achievement? A qualitative analysis that identified teacher and administrative opinions on the relationship between various teacher observation tools and perceived growth and achievement could add a valuable perspective. Educators could also be surveyed to understand how feedback from teacher evaluation systems affects teacher growth.

Similar to research question 4, interpreting the factors present in other popular teacher observation tools is another idea for future research. Comparing the results of other observation tools to the outcome of this study could provide valuable information regarding the effectiveness of the Georgia TKES compared to other states' observation tools. Further, decisions regarding the use of specific tools may be made dependent upon the effectiveness of alternative measures.

Conclusion

This study provides critical information on the predictive power of the TKES on student growth and achievement. The study also examined the impact of ED and academic settings on student achievement. Additionally, the study investigated the factorability of TKES Standards. Research question 1 provided an understanding of the predictive power of the TKES Standards on student growth. Research question 2 examined the predictive power of the TKES Standards on student achievement. Research question 3 analyzed the difference between achievement scores between levels of ED and classroom setting. Research question 4 assessed reducing the number of variables among the standards.

As noted in the research, there is a longstanding attempt for educators to improve their craft and enhance the quality of the education provided to the nation's students (Wise et al., 1984). While there are many schools of thought regarding how to improve public education, most leaders agree that effectively evaluating teachers is a critical step (Milanowski, 2011). However, how to best evaluate teachers is the subject of an enduring debate (Wise et al., 1984).

TKES is Georgia's form of educator evaluation. The evaluation system is based on multiple measures prioritizing student growth and achievement (H.B. 244, 2013-2014). One of the three primary components of TKES is the Teacher Assessment on Performance Standards (TAPS). TAPS provides evaluating administrators with rubrics on which to base teacher ratings following walkthrough observations. While there is discretion when assigning the summative scores at the end of the school year, observation scores generally inform the summative assessment ratings. The TAPS portion of the TKES evaluation is 50% of the overall teacher effectiveness rating. ("Georgia's Teacher Keys Effectiveness System," n.d).

Teachers are rated on four levels for each standard. The levels are exemplary (level IV), proficient (level III), needs development (level II), or ineffective (level I). The standards by which teachers are evaluated for observations include Standard 1: Professional Knowledge, Standard 2: Instructional Planning, Standard 3: Instructional Strategies, Standard 4: Differentiated Instruction, Standard 5: Assessment Strategies, Standard 6: Assessment Uses, Standard 7: Positive Learning Environment, Standard 8: Academically Challenging Environment, Standard 9: Professionalism, and Standard 10: Communication ("Georgia's Teacher Keys Effectiveness System," n.d).

The summative scores for each of the 10 TKES scores served as the independent variables in research questions 1 and 2 to determine the predictive power of TKES on student growth and achievement. The levels of the percentage of ED and academic setting (self-contained or departmentalized) served as the independent variables in question 3 to determine if there was a significant difference in achievement based on the levels of percentage of ED or setting. SGPs and GMAS MSS served as the dependent variables. The dataset used for the study consisted of a random sample of 4,000 fourth and fifth grade teachers. For each grade, data represented 800 self-contained ELA teachers and 600 departmentalized teachers for both ELA and math.

The findings of this study have significant implications for school leaders. Administrators may make decisions that address specific school needs based on these implications. For example, schools with low achievement may be more interested in promoting school growth explored in research question 1. In contrast, schools with higher achievement may be more inclined to explore the results of the second research question. In schools with existing high levels of student achievement, exponential student growth may not be feasible, and their focus may be to maintain current levels of achievement. Further, the results of question 3 may be helpful in forming master schedules and designing delivery models. Again, the implications may differ based on school needs, i.e., socioeconomic level. In question 4, leaders may gain insight into teacher behaviors by understanding the correlation between the standards. To correctly utilize the findings in this study for school impact, however, leaders should explore the results of the questions by grade level and subject as the implications are different based on these areas.

The results of research question 1 indicate that, while no standards in fifth grade are predictive of student growth, there are standards in fourth grade ELA and math that are predictive of student growth. In fourth grade ELA, Standard 4 at the proficient and exemplary level and Standard 8 at the exemplary level were found to have a positive impact on student growth. Standard 9 at the proficient level and Standard 3 at the proficient and exemplary level negatively impacted student growth. Administrators seeking to improve student growth in fourth grade ELA, perhaps in student populations considered at risk with low achievement scores, may choose to provide additional training on Standards 4 and 8 and the associated behaviors as indicated by the performance rubrics. For example, these two standards center around providing a student-centered environment with high expectations based on specific learning needs identified by student data. Because Standards 3 and 9 were associated with negative growth, the administrator may seek to isolate the factors with the negative impact and how these factors affect growth within specific school environments. Examples might include instructional strategies gained through professional development that were used incorrectly or not as intended. In fourth grade math, Standard 1 and 3 at the exemplary level were found to positively impact student growth, while no standards were noted as negatively impacting SGP's. Again, specific behaviors, as indicated by TKES rubrics, could increase growth in fourth grade math. Examples include the continual exploration and use of research-based strategies to promote rigor and student learning. Hiring decisions may be made based on high TKES scores for these standards. There are no implications of the results on fifth grade ELA or math.

In research question 2, the study indicated the TKES Standards that are predictive of student achievement. In fourth grade ELA, Standards 3, 7, 8, and 10 at the exemplary level were significant predictors of student achievement as indicated by the GMAS MSS. Standard 8 at the ineffective/needs improvement level was a significant predictor of having a negative impact on scale scores. Standards 1, 2, 7, and 8 at the exemplary level were found to positively impact GMAS MSS in fourth grade math. Because Standards 7 and 8 at the exemplary level are predictive of overall student achievement in fourth grade, educational leaders may study the associated behaviors and provide additional professional support so that teachers are highly successful at implementing the associated behaviors. As a fourth grade administrator, concentrating on professional development in Standards 7 and 8 could benefit both ELA and math. Standards 7 (needs development and exemplary), Standard 8 (exemplary), and Standard 9 (exemplary) were significant predictors of achievement in fifth grade ELA. Standards 1, 8, 9, and 9, at the exemplary levels, were significant predictors of achievement in fifth grade math. Because Standards 7, 9, and 9 predict student achievement in both subjects, an administrator may use that information to provide professional development opportunities and analyze rubrics associated with these standards to capitalize on these TKES Standards affecting both subjects. When seeking candidates, hiring candidates with high scores in these standards could greatly benefit student achievement in fifth grade.

It is essential to note the standards impacting both student growth and achievement. In fourth grade ELA, Standard 8 (Academically Challenging Environment) at the exemplary level had a significant impact on growth and achievement. In fourth grade math, Standard 1 (Professional Knowledge) at the exemplary level significantly impacted student growth and achievement. Administrators making hiring or professional development decisions in fourth grade could use this knowledge to potentially impact student growth and achievement. No standards were associated with growth and achievement in fifth grade ELA or math.

The results of research question 3 indicated a significant difference in the MSS based on the level of the percentage of ED. However, there is no significant difference based on academic setting except fourth grade math. Although interesting, the significance of the fourth grade math setting will most likely not change educational practice. To provide a departmentalized setting, teachers must exclusively teach ELA or math. Because the setting is not significant in ELA, it is not feasible to differentiate for only one setting. Administrators should consider additional instructional support for educators serving populations with higher levels of ED students and consider means by which the teachers may become content specialists. Because content specialists had a more significant impact on math achievement in fourth grade, professional development for fourth grade teachers may provide the additional benefit of a departmentalized classroom but in a self-contained setting.

In research question 4, through the factor analysis, 2 or 3 latent variables were created and adequately reduced the dimensionality of the TKES Standards. In fourth grade ELA, 2 latent variables were created, including variable 1 (Standards 1, 2, 3, 4, 5, 6, and 8) variable 2 (Standards 7, 9, and 10). In fourth grade math, 3 variables were created, including variable one (Standards 1, 5, and 6), variable 2 (Standards 3, 4, and 8), and variable 3 (Standards 2, 7, 9, and 10). In fifth grade ELA, 3 variables were created, including variable one (Standards 1, 2, 4, 5, and 6), variable 2 (Standards 9 and 10), and variable 3 (Standards 3, 7, and 8). In fifth grade math, 3 variables were created, including variable one (Standards 1, 2, 3, 7, and 8), variable 2 (Standards 9 and 10), and variable 3 (Standards 4, 5, and 6). Because there is little consistency among the TKES Standards contained within the latent variables across the grade level and subject, individual grade levels must be treated separately. There are, however, exceptions. Standards 9 and 10 and Standards 5 and 6 were included on the same latent variable across all four datasets. It is interesting to note that the indicators in these standard pairs generally appear together in teacher behavior. For example, in Standards 5 and 6, teachers who have appropriate assessment strategies typically understand how to effectively use the assessment data in their classrooms to drive instructional decisions. For Standards 9 and 10, behaviors associated with professionalism, Standard 9, are also understood throughout the indicators in the communication standard, or Standard 10. Further, administrators may seek to isolate the behaviors within the newly created variables for each grade to further understand teacher professional development and growth.

Although there were statistically significant findings of value gained in this study, the results were generally not consistent among grade levels and subjects. While administrators and educators may use the information linked to student growth and achievement, few broad statements may be made regarding the relationships between standards and student growth and achievement. The existence of these inconsistencies supports the longstanding debate over how to evaluate teachers best. While Georgia TKES was created to provide a comprehensive assessment for educators, more work is needed to develop and maintain an evaluative tool that not only concisely and correctly identifies teacher behaviors that promote student growth and achievement, but also provides avenue for professional growth.

References

250 + influences on student achievement. (2017). Visible Learning+. Retrieved from https://visible-learning.org/hattie-ranking-influences-effect-sizes-learning-achievement/

A Guide to the Georgia Student Growth Model. (2014). Retrieved from www.gadoe.org/Curriculum-Instruction-and-Assessment/Assessment/Documents/GSGM/FY18/SGPGuide 06052017.pdf

American Statistical Association. (2014). *ASA statement on using value-added models for educational assessment*. Retrieved from https://www.amstat.org/asa/files/pdfs/POL-ASAVAM-Statement.pdf

Anderson, L., Butler, A., Palmiter, A., & Arcaira, E. (2016). *Study of emerging teacher evaluation systems*. Retrieved from https://www2.ed.gov/rschstat/eval/teaching/emerging-teacher-evaluation/report.pdf

Assessing the validity and reliability of the Teacher Keys Effectiveness System (TKES) and the Leader Keys Effectiveness System (LKES) of the Georgia Department of Education. (2014). Georgia Center for Assessment. Retrieved from https://gaedassessment.files.wordpress.com/2018/05/tkes-lkes_reliability-validity_report_v2-student-perception-surveys-included.pdf

Baker, E., Barton, P., Darling-Hammond, L., Haertel, E., Ladd, H., Linn, R., Ravitch, D., Rothstein, R., Shavelson, R., & Shepard, L. (2010). *Problems with the use of student test scores to evaluate teachers*. EPI briefing paper #278. http://www.epi.org. Washington, D.C.: Economic Policy Institute.

Barge, J. (2013). *Teacher Keys Effectiveness System*. Retrieved from https://www.gadoe.org/School-Improvement/Teacher-and-Leader-Effectiveness/Documents/TKES%20Handbook%20FINAL%207-18-2013.pdf

Barrette, C., Morton, E., & Tozcu, A. (1995). An overview of teacher evaluation. *SLAT Student Association*, *3*(1), 36–48. Retrieved from file:///C:/Users/9923100/Downloads/21487-40326-1-PB (4).pdf

Bellani, L., & Bia, M. (2018). The long-run effect of childhood poverty and the mediating role of education. *Journal Of The Royal Statistical Society*, *182*, 37-68.

Berliner, D. C. (2018). Between Scylla and Charybdis: Reflections on and problems associated with the evaluation of teachers in an era of metrification. *Education Policy Analysis Archives*, *26*(54), 1–25. doi: 10.14507/epaa.26.3820

Boatwright, P., & Metcalf, L. (2020). The effects of poverty on lifelong learning: Important lessons for educators. *The Delta Kappa Gamma Bulletin: International Journal For Professional Educators*, *85*(3), 52-57.

Carbough, C., Marzano, R., & Toth, M. (2017). *The Marzano focused teacher evaluation model.* Retrieved from https://www.learningsciences.com/wp/wp-content/uploads/2017/06/Focus-Eval-Model-Overview-2017.pdf

Cheung, A., & Slavin, R. (2016). How methodological features affect effect sizes in education. *Educational Researcher*, *45*(5), 283-292.

Coe, R., Aloisi, C., Higgins, S., & Major, L. (2014). *What makes great teaching?* Retrieved from http://www.suttontrust.com/research-paper/great-teaching

Cogan, M. (1973). *Clinical supervision*. Boston: Houghton Mifflin.

Cosner, S., Kimball, S. M., Barkowski, E., Carl, B., & Jones, C. (2015). Principal roles, work demands, and supports needed to implement new teacher evaluation. *Mid-Western Educational Researcher*, *27*(1), 76-95.

Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika*, *16*(3), 297–334. doi: 10.1007/bf02310555

Danielson, C. (1996). *Enhancing professional practice: A framework for teaching*. Alexandria, VA: Association for Supervision and Curriculum Development.

Danielson, C. (2014). *The framework for teaching evaluation instrument 2013 edition* [Ebook]. Retrieved from http://www.loccsd.ca/~div15/wp-content/uploads/2015/09/2013-framework-for-teaching-evaluation-instrument.pdf

Darling-Hammond, L. (2009). Recognizing and enhancing teacher effectiveness. *The International Journal of Educational and Psychological Assessment*, 3, 1-24.

Darling-Hammond, L. (2013). *Getting teacher evaluation right*. New York: Teachers College Press.

Darling-Hammond, L., Hyler, M. E., & Gardner, M. (2017*). Effective teacher professional development*. Retrieved from https://learningpolicyinstitute.org/product/effective-teacher-professional-development-brief

Derrington, M. L. (2014). Teacher evaluation initial policy implementation: superintendent and principal perceptions. *Planning & Changing*, *45*(1/2), 120-137.

Derrington, M. L., & Campbell, J. W. (2015). Implementing new teacher evaluation systems: Principals' concerns and supervisor support. *Journal of Educational Change*, 1-22.

Donaldson, M. L., & Papay, J. P. (2014). Teacher evaluation reform: Policy lessons for school principals. *Principal's Research Review*, *9*(5), 1.

Dymond, A. (2017). *A case study on the comparison of fourth-grade students' mathematical achievement as evidenced by the measures of academic progress assessment: Self-contained vs. departmentalized settings* (Unpublished doctoral dissertation). Gardner-Webb, Boiling Springs, NC.

Elder, T., Wang, J., & Cramer, S. (2015). An evaluation of the validity and reliability of the Intern Keys Assessment. Retrieved from https://coehp.columbusstate.edu/docs/accreditation/an-evaluation-of-the-validity-and-reliability-of-the-intern-keys-assessment.pdf

Every Student Succeeds Act (ESSA). (2017). Retrieved from https://www.ed.gov/ESSA

Every Student Succeeds Act (ESSA) reporting requirements. (n.d.). Retrieved from https://schoolgrades.georgia.gov/reference

Fisher, D., Frey, N., & Hattie, J. (2016). *Visible learning for literacy, grades K-12: Implementing the practices that work best to accelerate student learning*. Thousand Oaks: CA: Corwin Press.

Fuhrman, S. (2010). *Tying teacher evaluation to student achievement*. Retrieved from http://68.77.48.18/RandD/Education%20Week/Tying%20Teacher%20Eval%20to%20Student%20Achievement%20-%20Fuhrman.pdf

Gaderman, A., Guhn, M., & Zumbo, B. (2012). Estimating ordinal reliability for Likert-type and ordinal item response data: A conceptual, empirical, and practical guide. *Practical Assessment, Research, and Evaluation*, *17*(3), 1-13. doi: https://doi.org/10.7275/n560-j767.

Georgia State University Law Review, Education Quality Basic Education Act: Enacted, 1 Ga. St. U. L. Rev. (2012). Available at: https://readingroom.law.gsu.edu/gsulr/vol1/iss2/27

Georgia-Student-Growth-Model (2018). Retrieved from https://www.gadoe.org/Curriculum-Instruction-and-Assessment/Assessment/Pages/Georgia-Student-Growth-Model.aspx

Georgia's Teacher Keys Effectiveness System: Meaningful feedback professional growth flexibility to innovate. (2018) Atlanta, GA: Georgia Department of Education. Retrieved from https://www.gadoe.org/School-Improvement/Teacher-and-Leader-Effectiveness/Documents/TKES%20LKES%20Do

Georgia's Teacher Keys Effectiveness System. (n.d.). Retrieved from https://www.gadoe.org/School-Improvement/Teacher-and-Leader-Effectiveness/Documents/TKES LKES Documents/TKESHandbook2018.2019final.pdf

Glatthorn, A. (1984). *Differentiated supervision*. Alexandria, VA: Association for Supervision and Curriculum Development.

Glickman, C. (1985). Development as the aim of instructional supervision. Retrieved from https://files.eric.ed.gov/fulltext/ED263655.pdf

Goldhader, D., Brewer, D., & Anderson, D. (1999). A three-way error components analysis of educational productivity. *Education Economics*, *7*(3), 199-208. doi: 10.1080/09645299900000018

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

H.B. 244. Georgia (2013-2014).

Haertel, E. (2013). *Reliability and validity of inferences about teachers based on test scores*. Princeton, NJ: Educational Testing Services. Retrieved from https://www.ets.org/Media/Research/pdf/PICANG14.pdf

Hattie, J. (2008). *Visible Learning: A Synthesis of over 800 Meta-Analyses Relating to Achievement;* Abingdon, Oxon: Routledge.

Hattie, J., Fisher, D., & Frey, N. (2017). *Visible learning for mathematics, grades K-12: What works best to optimize student learning*. Thousand Oaks, CA: Corwin Press.

Ho, A., & Kane, T. (2013). *The reliability of classroom observations by school personnel*. Retrieved from https://danielsongroup.org/sites/default/files/inline-files/Reliabiilty_Observations_School_Personnel_Gates.pdf

Hosford, P. (1984). *Using what we know about teaching*. Association for Supervision and Curriculum Development, Alexandria, VA.

Hunter, M. (1980). Six types of supervisory conferences. *Educational Leadership, 37*(5), 408–412.

Introduction to generalized linear models. (n.d.). Retrieved from https://online.stat.psu.edu/stat504/node/216/

Jack, D. (2014). *Self-contained versus departmentalized settings in urban elementary schools: An analysis of fifth-grade student mathematics performance* (Unpublished doctoral dissertation). Mercer University. Macon, GA.

Jensen, E. (2010). *Teaching with poverty in mind: what being poor does to kids brains and what schools can do about it*. Association for Supervision and Curriculum Development.

Kane, M. (2009). Validating the interpretations and uses of test scores. *The concept of validity: Revisions, new directions, and applications* (1-73) Charlotte, NC: Information Age Publishing, Inc.

Kane, M. (2013). Validating the interpretations and uses of test scores. *Journal of Educational Measurement, 50*(1), 1-73. doi: 10.1111/jedm.12000.

Kane, T., Taylor, E., Tyler, J., & Wooten, A. (2011). Identifying effective classroom practices using student achievement data. *The Journal of Human Resources*, *46*(3), 587-613.

Kowalski, T., & Dolph, D. (2015). Principal dispositions regarding the Ohio teacher evaluation system. *AASA Journal of Scholarship and Practice*, *11*(4), 4-20.

Leader-Keys-Effectiveness-System. (2018). Retrieved from https://www.gadoe.org/School-Improvement/Teacher-and-Leader-Effectiveness/Pages/Leader-Keys-Effectiveness-System.aspx

Leader Keys Effectiveness System: Fact sheets. (2012). Retrieved from http://www.gadoe.org/School-Improvement/Teacher-and-Leader-Effectiveness/Documents/LKES%20Fact%20Sheets%20September%202012.pdf

Marzano, R. J., Frontier, T., & Livingston, D. (2011). *Effective supervision: Supporting the art and science of teaching*. Alexandria, VA: Association for Supervision and Curriculum Development.

McGreal, T. (1983). *Successful teacher evaluation*. Alexandria, VA: Association for Supervision and Curriculum Development.

Mertler, C., & Vannatta, R. (2013). *Advanced and multivariate statistical methods*. Glendale, CA: Pyrczak Publishing.

Milanowski, A. (2011). Strategic measures of teacher performance. *The Phi Delta Kappan, 92*(7), 19-25.

Mitchell, V. (2013). *Departmentalized or self-contained: The relationship between classroom configuration and student achievement* (Unpublished doctoral dissertation). California State University. Long Beach, CA.

Montero, M. (2020). *The pros and cons of departmentalization in elementary schools*. Retrieved from https://teachingwithamountainview.com/the-pros-and-cons-of/

Najimi, A., Sharifirad, G., Amini, M., & Meftagh, S. (2013). Academic failure and students' viewpoint: The influence of individual, internal, and external organizational factors. *Journal of Education and Health Promotion*, *2*(1), 22.

Nelson, K. (2014). *A study comparing fifth grade student achievement in mathematics in departmentalized and non-departmentalized and non-departmentalized settings* (Unpublished doctoral dissertation). Liberty University. Lynchburg, VA.

Nye, B., Konstantopoulos, S., & Hedges, L. (2004). How large are teacher effects?. Retrieved from https://journals.sagepub.com/doi/abs/10.3102/01623737026003237

Patterson, C. (2019, March 19). *Measuring the lasting impact of a nation at risk*. Retrieved from https://www.waltonfamilyfoundation.org/stories/k-12-education/measuring-the-lasting-impact-of-a-nation-at-risk

Primer: Education Issues – Variables affecting student achievement. (2014). Retrieved from http://weac.org/articles/primer-variable/

Putman, H., Ross, E. & Walsh, K. (2018). *Making a difference: Six places where teacher evaluation systems are getting results.* Retrieved from: https://www.nctq.org/publications/Making-a-Difference

Raudys, J. (2018). *4 teacher evaluation models to use (with examples!).* Retrieved from https://www.prodigygame.com/main-en/blog/teacher-evaluation

Robinson, S. (n.d.). Teacher evaluations - Why teacher performance matters. Retrieved from https://www.frontlineeducation.com/teacher-evaluation/

Schmoker, M. (2014). *Focus*: *Elevating the essentials to radically improve student learning.* Alexandria, VA: Association for Supervision and Curriculum Development.

Skedsmo, G. G., & Huber, S. G. (2018). Teacher evaluation: The need for valid measures and teacher involvement. *30*(2), 1-5. *Educational Assessment Evaluation and Accountability*. doi: 10.1007/s11092-018-9273-9.

Strauss, V. (2014). *Statistician slam popular teacher evaluation method. The Washington Post.* Retrieved from *http://www.washingtonpost.com.*

Teacher-Keys-Effectiveness-System. (n.d.). Retrieved from https://www.gadoe.org/School-Improvement/Teacher-and-Leader-Effectiveness/Pages/Teacher-Keys-Effectiveness-System.aspx

TEM Scoring Guide. (2014). Retrieved from https://www.gadoe.org/School-Improvement/Teacher-and-Leader-Effectiveness/Documents/TEM%20Scoring%20Guide%206-18-14Final1.pdf

Terhart, E. (2011). Has John Hattie really found the holy grail of research on teaching? An extended review of Visible Learning. *Journal of Curriculum Studies*, *43*(3), 425-438.

The effect of poverty on student achievement. (2009). *Information Capsule: Research Services*, *0901*. Retrieved from https://files.eric.ed.gov/fulltext/ED544709.pdf

The focused teacher evaluation model: Summary and implementation. (2020). Retrieved from https://www.marzanocenter.com/wp-content/uploads/sites/4/2018/10/MC06-14-FTEM-White-Paper-1-16-18-Digital-4.pdf

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

The No Child Left Behind Act of 2001. (2002). States Department of Education. Retrieved from https://www2.ed.gov/nclb/overview/intro/execsumm.pdf

Toch, T., & Rothman, R. (2008). *Rush to judgment: Teacher evaluation in public education*. Washington, DC: Education Sector.

Tucker, P. D., & Stronge, J. H. (2005). *Linking teacher evaluation and student learning*. Alexandria, VA: Association for Supervision and Curriculum Development.

Turnell, J. L. (2018). *An explorational study of educational leaders ' attitudes, opinions, understanding, and application of student growth percentile (SGP) data*. (Unpublished doctoral dissertation) Kennesaw University. Kennesaw, GA.

Understanding the Georgia Milestones. (2020). Retrieved from https://www.gadoe.org/Curriculum-Instruction-and-Assessment/Assessment/Pages/achievement_levels.aspx

Vallaster, J. (2019). *Recognizing and supporting the forgotten poverty frontier: Exploring suburban school poverty in elementary schools* (Unpublished doctoral dissertation). George Washington University. Washington, DC.

Value-Added Modeling 101. (2020). Retrieved from https://www.rand.org/education-and-labor/projects/measuring-teacher-effectiveness/value-added-modeling.html

Warnock, D. B. (2015). *An evaluation of principals' perceptions of Georgia's teacher KEYS effectiveness system*. (Unpublished doctoral dissertation). Georgia Southern University, Atlanta, GA.

Weiner, R. & Jacobs, A. (2011). Designing and implementing teacher performance management. http://aspeninstitute.org/education. Washington, D.C.: Aspen Institute.

Weisberg, D., Sexton, S., Mulhern, J., & Keeling, D. (2009). *The widget effect: Our national failure to acknowledge and act on differences in teacher effectiveness*. Brooklyn, NY: Grantee.

Whitehurst, G. J., Chingos, M. M., & Lindquist, K. M. (2015). Getting classroom observations right. *Education Digest*, *80*(7), 20-28.

Wise, E., Darling-Hammond, L., McLaughlin, M., & Bernstein, H. (1984). *Teacher evaluation: A study of effective practices*. Santa Monica, CA: RAND.

Woods, K. (2017). *The impact of departmentalized classrooms versus self-contained classrooms on fourth and fifth-grade students' reading and math achievement* (Ed. D.). Wingate University.

APPENDIX A:

R Code for Ordinal Regression

R Code for Ordinal Regression

Set Working Directory

```
setwd("C:/Users/9923100/OneDrive - Cherokee County School District/VSU/dissertation/Question 1")


#install.packages("tidyverse")
#install.packages("readxl")
#install.packages("tidymodels")
#install.packages("MASS")
#install.packages("DescTools")
#install.packages("ggplot2")
#install.packages("Hmisc")
#install.packages("recipes")
#install.packages("EnvStats")
#install.packages("AER")
#install.packages("psych")
#install.packages("effects")
#install.packages("polycor")
library(tidyverse)
library(readxl)
library(tidymodels)
library(MASS)
library(DescTools)
library(ggplot2)
library(Hmisc)
library(recipes)
library(EnvStats)
library(AER)
library(psych)
library(effects)
library(polycor)
```

```
#read in dataset
RQ1E4<-read_excel("GADOE data rq1.4E edited.xlsx")

head(RQ1E4)

OLR_data <- RQ1E4

summary(OLR_data)

#Ordering the dependent variable
OLR_data$`ELA SGP Level` <- factor(OLR_data$`ELA SGP Level`,levels=c(1,2,3,4),
        labels=c("Ineffective", "Needs Development", "Proficient",
              "Exemplary"))

#Ordering the independent variable
OLR_data$PS1 <- as.factor(OLR_data$PS1)

OLR_data$PS2 <- as.factor(OLR_data$PS2)

OLR_data$PS3 <- as.factor(OLR_data$PS3)

OLR_data$PS4 <- as.factor(OLR_data$PS4)

OLR_data$PS5 <- as.factor(OLR_data$PS5)

OLR_data$PS6 <- as.factor(OLR_data$PS6)

OLR_data$PS7 <- as.factor(OLR_data$PS7)
```

```
OLR_data$PS8 <- as.factor(OLR_data$PS8)

OLR_data$PS9 <- as.factor(OLR_data$PS9)

OLR_data$PS10 <- as.factor(OLR_data$PS10)

#Descriptive Analysis
#Exploratory data analysis
#Summarizing the data
summary(OLR_data)

#Frequency Table by standard
table(OLR_data$`ELA SGP Level`,OLR_data$PS1)
table(OLR_data$`ELA SGP Level`,OLR_data$PS2)
table(OLR_data$`ELA SGP Level`,OLR_data$PS3)
table(OLR_data$`ELA SGP Level`,OLR_data$PS4)
table(OLR_data$`ELA SGP Level`,OLR_data$PS5)
table(OLR_data$`ELA SGP Level`,OLR_data$PS6)
table(OLR_data$`ELA SGP Level`,OLR_data$PS7)
table(OLR_data$`ELA SGP Level`,OLR_data$PS8)
table(OLR_data$`ELA SGP Level`,OLR_data$PS9)
table(OLR_data$`ELA SGP Level`,OLR_data$PS10)

#TKES Matrix
tkesmatrix=OLR_data[,3:12]

#generate Correlation matrix
tkescorr=polychoric(tkesmatrix)
tkescorr$rho
```

```
#Run Bartlett's on raw dataframe
cortest.bartlett(tkescorr$rho,n=1400)

#Calculate KMO
KMO(tkescorr$rho)

#Determinate of correlation matrix
det(tkescorr$rho)

OLR_data <- RQ1E4

#Ordering the dependent variable
OLR_data$`ELA SGP Level` <- factor(OLR_data$`ELA SGP Level`,levels=c(1,2,3,4),
                        labels=c("Ineffective", "Needs Development", "Proficient",
                              "Exemplary"))

#Ordering the independent variable
OLR_data$PS1 <- factor(OLR_data$PS1,levels = c(2,3,4),labels =
                 c("Ineffective/Needs Improvement", "Proficient",
                  "Exemplary"))



OLR_data$PS2 <- factor(OLR_data$PS2,levels = c(2,3,4),labels =
                 c("Ineffective/Needs Improvement","Proficient",
                  "Exemplary"))

OLR_data$PS3 <- factor(OLR_data$PS3,levels = c(2,3,4),labels =
                 c("Ineffective/Needs Improvement", "Proficient",
                  "Exemplary"))
```

```
OLR_data$PS4 <- factor(OLR_data$PS4,levels = c(2,3,4),labels =
                    c("Ineffective/Needs Improvement", "Proficient",
                      "Exemplary"))



OLR_data$PS5 <- factor(OLR_data$PS5,levels = c(2,3,4),labels =
                    c("Ineffective/Needs Improvement", "Proficient",
                      "Exemplary"))



OLR_data$PS6 <- factor(OLR_data$PS6,levels = c(2,3,4),labels =
                    c("Ineffective/Needs Improvement", "Proficient",
                      "Exemplary"))



OLR_data$PS7 <- factor(OLR_data$PS7,levels = c(2,3,4),labels =
                    c("Ineffective/Needs Improvement", "Proficient",
                      "Exemplary"))



OLR_data$PS8 <- factor(OLR_data$PS8,levels = c(2,3,4),labels =
                    c("Ineffective/Needs Improvement", "Proficient",
                      "Exemplary"))

OLR_data$PS9 <- factor(OLR_data$PS9,levels = c(2,3,4),labels =
                    c("Ineffective/Needs Improvement", "Proficient",
                      "Exemplary"))

OLR_data$PS10 <- factor(OLR_data$PS10,levels = c(2,3,4),labels =
                     c("Ineffective/Needs Improvement",  "Proficient",
                    "Exemplary"))

#Build ordinal logistic regression model
OLRmodel <- polr(`ELA SGP Level` ~ PS1 + PS2 + PS3 + PS4 + PS5 + PS6 +
              PS7 + PS8 + PS9 + PS10, data = OLR_data, Hess = TRUE)
summary(OLRmodel)
coeftest(OLRmodel)

#check for multicollinearity
vif(OLRmodel)
vif_values <- vif(OLRmodel)

#Confidence intervals in log odds scale
(ci <- confint(OLRmodel))

#confidence intervals in odds ratio scale
exp(coef(OLRmodel))
exp(cbind(OR = coef(OLRmodel), ci))

#Check the overall model fit
#Run a "only intercept" model
OIM <- polr(`ELA SGP Level` ~ 1, data = OLR_data)
summary(OIM)

#Compare the our test model with the "only intercept" model
anova(OIM,OLRmodel)

#Check the model fit information
#Check the predicted probability for each program
OLRmodel$fitted.values
```

```
#We can get the predicted result by using predict function
predict(OLRmodel)

#Test the goodness of fit
chisq.test(OLR_data$`ELA SGP Level`,predict(OLRmodel))

#Compute confusion table and misclassification error
predictOLR <- predict(OLRmodel,OLR_data)
cTab <- table(OLR_data$`ELA SGP Level`, predictOLR)
mean(as.character(OLR_data$`ELA SGP Level`) != as.character(predictOLR))

(CCR <- sum(diag(cTab)) / sum(cTab)) # Calculate the classification rate

#Measuring strength of association
#Calculate the R Square
PseudoR2(OLRmodel, which = c("CoxSnell"))
PseudoR2(OLRmodel, which = c("Nagelkerke"))
PseudoR2(OLRmodel, which = c("McFadden"))

#We can use the coef() function to check the parameter estimates
(OLRestimates <- coef(summary(OLRmodel)))

#Parameter Estimates
#Add the p-value to our parameter estimates table
p <- pnorm(abs(OLRestimates[, "t value"]), lower.tail = FALSE) * 2
(OLRestimates_P <-cbind(OLRestimates, "pvalue" = p))
options(scipen = 999)

#Calculating Expected Values
```

```
#Building the newteacher entry to test our model
newteacher <- data.frame("subject" = 1, `ELA SGP Level`= NA,
                         "PS1" = "Proficient", "PS2" = "Proficient",
                         "PS3" = "Proficient", "PS4" = "Proficient",
                         "PS5" = "Proficient", "PS6" = "Proficient",
                         "PS7" = "Proficient", "PS8" = "Proficient",
                         "PS9" = "Proficient", "PS10" = "Proficient")

#Use predict function to predict the new teacher's situation
predict(OLRmodel, newteacher)

#Checking proportional odds assumption

sf <- function(y) {
  c('Y>=1' = qlogis(mean(y >= 1)),
    'Y>=2' = qlogis(mean(y >= 2)),
    'Y>=3' = qlogis(mean(y >= 3)),
    'Y>=4' = qlogis(mean(y >= 4)))
}

(s <- with(OLR_data, summary(as.numeric(`ELA SGP Level`) ~ PS1 + PS2 + PS3 + PS4 + PS5 +
                       PS6 + PS7 + PS8 + PS9 + PS10, fun=sf)))

#GLM
glm(I(as.numeric(`ELA SGP Level`) >= 2) ~ PS1, family="binomial", data = OLR_data)
glm(I(as.numeric(`ELA SGP Level`) >= 3) ~ PS1, family="binomial", data = OLR_data)
glm(I(as.numeric(`ELA SGP Level`) >= 4) ~ PS1, family="binomial", data = OLR_data)

glm(I(as.numeric(`ELA SGP Level`) >= 2) ~ PS2, family="binomial", data = OLR_data)
glm(I(as.numeric(`ELA SGP Level`) >= 3) ~ PS2, family="binomial", data = OLR_data)
glm(I(as.numeric(`ELA SGP Level`) >= 4) ~ PS2, family="binomial", data = OLR_data)

glm(I(as.numeric(`ELA SGP Level`) >= 2) ~ PS3, family="binomial", data = OLR_data)
glm(I(as.numeric(`ELA SGP Level`) >= 3) ~ PS3, family="binomial", data = OLR_data)
glm(I(as.numeric(`ELA SGP Level`) >= 4) ~ PS3, family="binomial", data = OLR_data)

glm(I(as.numeric(`ELA SGP Level`) >= 2) ~ PS4, family="binomial", data = OLR_data)
glm(I(as.numeric(`ELA SGP Level`) >= 3) ~ PS4, family="binomial", data = OLR_data)
glm(I(as.numeric(`ELA SGP Level`) >= 4) ~ PS4, family="binomial", data = OLR_data)

glm(I(as.numeric(`ELA SGP Level`) >= 2) ~ PS5, family="binomial", data = OLR_data)
glm(I(as.numeric(`ELA SGP Level`) >= 3) ~ PS5, family="binomial", data = OLR_data)
glm(I(as.numeric(`ELA SGP Level`) >= 4) ~ PS5, family="binomial", data = OLR_data)

glm(I(as.numeric(`ELA SGP Level`) >= 2) ~ PS6, family="binomial", data = OLR_data)
glm(I(as.numeric(`ELA SGP Level`) >= 3) ~ PS6, family="binomial", data = OLR_data)
glm(I(as.numeric(`ELA SGP Level`) >= 4) ~ PS6, family="binomial", data = OLR_data)

glm(I(as.numeric(`ELA SGP Level`) >= 2) ~ PS7, family="binomial", data = OLR_data)
glm(I(as.numeric(`ELA SGP Level`) >= 3) ~ PS7, family="binomial", data = OLR_data)
glm(I(as.numeric(`ELA SGP Level`) >= 4) ~ PS7, family="binomial", data = OLR_data)

glm(I(as.numeric(`ELA SGP Level`) >= 2) ~ PS8, family="binomial", data = OLR_data)
glm(I(as.numeric(`ELA SGP Level`) >= 3) ~ PS8, family="binomial", data = OLR_data)
glm(I(as.numeric(`ELA SGP Level`) >= 4) ~ PS8, family="binomial", data = OLR_data)

glm(I(as.numeric(`ELA SGP Level`) >= 2) ~ PS9, family="binomial", data = OLR_data)
glm(I(as.numeric(`ELA SGP Level`) >= 3) ~ PS9, family="binomial", data = OLR_data)
glm(I(as.numeric(`ELA SGP Level`) >= 4) ~ PS9, family="binomial", data = OLR_data)
glm(I(as.numeric(`ELA SGP Level`) >= 2) ~ PS10, family="binomial", data = OLR_data)
glm(I(as.numeric(`ELA SGP Level`) >= 3) ~ PS10, family="binomial", data = OLR_data)
glm(I(as.numeric(`ELA SGP Level`) >= 4) ~ PS10, family="binomial", data = OLR_data)

#Extension
s[, 5] <- s[, 5] - s[, 4]
s[, 4] <- s[, 4] - s[, 3]

[, 3] <- s[, 3] - s[, 3]
s # print

##Plots
# with the SGP levels
par("mar")
par(mar=c(1,1,1,1))
plot(s, which=1:4, pch=1:4, xlab='logit', main=' ', xlim=c(-6,0))
with(OLR_data,table(PS1,`ELA SGP Level`))

### Bootstrap Sample
bootstraps(OLR_data, times = 1e3)

set.seed(12345)
fit_intervals <- reg_intervals(`ELA SGP Level` ~ PS1 + PS2 + PS3 + PS4 + PS5 + PS6 + PS7 + PS8 +
PS9 + PS10,

                          data = OLR_data,

                          type = "percentile",

                          keep_reps = TRUE)
```

```
# RUN keep_reps = FALSE ABOVE Generates point estimates with error bars

fit_intervals %>%

  mutate(term = str_remove(term, "TRUE"),

         term = fct_reorder(term, .estimate)) %>%

  ggplot(aes(.estimate, term)) +

  geom_vline(xintercept = 0, size = 1.5, lty = 2, color = "gray80") +

  geom_errorbarh(aes(xmin = .lower, xmax = .upper),

                 size = 1.5, alpha = 0.5, color = "midnightblue") +

  geom_point(size = 3, color = "midnightblue") +

  labs(y = NULL)

# RUN keep_reps = TRUE ABOVE   Generates histogram of estimates
fit_intervals %>%
  mutate(term = str_remove(term, "TRUE"),
         term = fct_reorder(term, .estimate)) %>%
  unnest(.replicates) %>%
  ggplot(aes(estimate, fill = term)) +
  geom_vline(xintercept = 0, size = 1.5, lty = 2, color = "gray50") +
  geom_histogram(alpha = 0.8, show.legend = FALSE) +
  facet_wrap(vars(term))
```

APPENDIX B:

R Code for Multiple Regression

R Code for Multiple Regression

```
###Set Working Directory
setwd("C:/Users/9923100/OneDrive - Cherokee County School District/VSU/dissertation/Question 2")

###Install and load packages
install.packages("car")
install.packages("DescTools")
install.packages("psych")
install.packages("QuantPsyc")
install.packages("lmtest")
install.packages("lsr")
install.packages("readxl")
install.packages("knitr")
install.packages("tidymodels")
install.packages("devtools")
install.packages("stringr")
install.packages("forcats")
install.packages("ggplot2")
install.packages("dplyr")
install.packages("tidyr")
install.packages("reshape2")
install.packages("tidyverse")
install.packages("Rtools")
install.packages("rsample")
install.packages("rmarkdown")
install.packages("tinytex")
library(car)
library(DescTools)
library(psych)
library(QuantPsyc)
library(lmtest)
library(lsr)
library(readxl)
library(knitr)
library(tidymodels)
library(devtools)
library(stringr)
library(forcats)
library(ggplot2)
library(dplyr)
library(tidyr)
library(reshape2)
library(tidyverse)
library(rsample)
library(rmarkdown)
library(tinytex)

###read in dataset
RQ24E<-read_excel("rq2.4E.xlsx")

###Generates Pearson correlations
wfc.cor <- round(cor(RQ24E), 2)
wfc.cor

# Correlation effort to get Polychoric correlations.
poly <- mixedCor(data=RQ24E, c=1, p = 2:11, d=NULL, smooth=TRUE, correct=FALSE,
global=TRUE, ncat=4, use="pairwise", method="pearson", weight=NULL)
poly$rho <- round((poly$rho), 2)
poly$rho
```

```
#Changes to a factor variable and relevels to make "3" the reference level
RQ24E$PS1 = as.factor(RQ24E$PS1)
RQ24E$PS1 <- factor(RQ24E$PS1,levels = c(3,2,4))

RQ24E$PS2 = as.factor(RQ24E$PS2)
RQ24E$PS2 <- factor(RQ24E$PS2,levels = c(3,2,4))

RQ24E$PS3 = as.factor(RQ24E$PS3)
RQ24E$PS3 <- factor(RQ24E$PS3,levels = c(3,2,4))

RQ24E$PS4 = as.factor(RQ24E$PS4)
RQ24E$PS4 <- factor(RQ24E$PS4,levels = c(3,2,4))

RQ24E$PS5 = as.factor(RQ24E$PS5)
RQ24E$PS5 <- factor(RQ24E$PS5,levels = c(3,2,4))

RQ24E$PS6 = as.factor(RQ24E$PS6)
RQ24E$PS6 <- factor(RQ24E$PS6,levels = c(3,2,4))

RQ24E$PS7 = as.factor(RQ24E$PS7)
RQ24E$PS7 <- factor(RQ24E$PS7,levels = c(3,2,4))

RQ24E$PS8 = as.factor(RQ24E$PS8)
RQ24E$PS8 <- factor(RQ24E$PS8,levels = c(3,2,4))

RQ24E$PS9 = as.factor(RQ24E$PS9)
RQ24E$PS9 <- factor(RQ24E$PS9,levels = c(3,2,4))

RQ24E$PS10 = as.factor(RQ24E$PS10)
RQ24E$PS10 <- factor(RQ24E$PS10,levels = c(3,2,4))
```

```
str(RQ24E) # checks the coding of the variable
head(RQ24E) # Generates the first five rows of output.

summary(RQ24E)
describe(RQ24E$ELASS)

### Checks normality
shapiro.test(RQ24E$ELASS)
JarqueBeraTest(RQ24E$ELASS)

### Checks z scores for outliers
z <- scale(RQ24E$ELASS, center = TRUE, scale = TRUE)
#print(z) #I do not run this line unless I have a smaller dataset.
summary(z)
describe(z)

###This reduces the z score output. Rounds to two decimal places
X <- round(z, 2)
xtabs(~X) # Look at this output for outliers

# Generates descriptive statistics for the dataset
summary(RQ24E)    # Generates descriptive statistics
describe(RQ24E)   # Generates descriptive statistics

# To dummy code the factor variables and transform the DV.
set.seed(1234)
bc_rec <- recipe(ELASS ~ PS1 + PS2 + PS3 + PS4 + PS5 + PS6 + PS7 + PS8 + PS9 + PS10, data =
RQ24E) %>%
  step_dummy(all_predictors()) %>%
  step_BoxCox(ELASS)
bc_rec


bc_prep <- prep(bc_rec, retain = TRUE)
bc_prep


bc_juiced <- juice(bc_prep)
bc_juiced


# This line shows the descriptive statistics for the transformed variable.
describe(bc_juiced)


#Check normality after the data transformation.
shapiro.test(bc_juiced$ELASS)
JarqueBeraTest(bc_juiced$ELASS)


z <- scale(bc_juiced$ELASS, center = TRUE, scale = TRUE)
#print(z) #I do not run this line unless I have a smaller dataset.
summary(z)
describe(z)
# This produces the z score output. Rounds to two decimal places
X <- round(z, 2)
xtabs(~X) # Look at this output for outliers



###################
# Multiple Regression #
###################
# In the next sections of code, I am beginning to check the assumptions for the multiple regression analysis.
# Checking the DV for univariate normality using multiple procedures. Evaluate the output
# for each procedure.

hist(bc_juiced$ELASS)
qqPlot(bc_juiced$ELASS, distribution ="norm", pch=20, col="red", lwd=2, main="Normal Q-Q Plot")

##################
# The MR Analysis    #
#Generates the correct regression model with dummy coded items
fit <- lm(ELASS ~ PS1_X2 + PS1_X4 + PS2_X2 + PS2_X4 + PS3_X2 + PS3_X4 + PS4_X2 + PS4_X4
+ PS5_X2 + PS5_X4 +
          PS6_X2 + PS6_X4 + PS7_X2 + PS7_X4 + PS8_X2 + PS8_X4 + PS9_X2 + PS9_X4 + PS10_X2
+ PS10_X4, data = bc_juiced)
summary(fit)

# Produces the unstandardized regression coefficients and confidence intervals
coefficients(fit)
confint(fit)

# Generates the standardized regression coefficients
standardCoefs(fit)  # Generates unstandardized and standardized coefficients

### Bootstrap Sample
bootstraps(bc_juiced, times = 1e3)

set.seed(12345)
fit_intervals <- reg_intervals(ELASS ~ PS1_X2 + PS1_X4 + PS2_X2 + PS2_X4 + PS3_X2 + PS3_X4 +
PS4_X2 + PS4_X4 + PS5_X2 + PS5_X4 + PS6_X2 + PS6_X4 + PS7_X2 + PS7_X4 + PS8_X2 +
PS8_X4 + PS9_X2 + PS9_X4 + PS10_X2 + PS10_X4,
                    data = bc_juiced,

                    type = "percentile",

                    keep_reps = FALSE)

# RUN keep_reps = FALSE ABOVE Generates point estimates with error bars

fit_intervals %>%

  mutate(term = str_remove(term, "TRUE"),

        term = fct_reorder(term, .estimate)) %>%

  ggplot(aes(.estimate, term)) +

  geom_vline(xintercept = 0, size = 1.5, lty = 2, color = "gray80") +

  geom_errorbarh(aes(xmin = .lower, xmax = .upper),

                size = 1.5, alpha = 0.5, color = "midnightblue") +

  geom_point(size = 3, color = "midnightblue") +

  labs(y = NULL)

# RUN keep_reps = TRUE ABOVE   Generates histogram of estimates
fit_intervals %>%
  mutate(term = str_remove(term, "TRUE"),
        term = fct_reorder(term, .estimate)) %>%
  unnest(.replicates) %>%
  ggplot(aes(estimate, fill = term)) +
  geom_vline(xintercept = 0, size = 1.5, lty = 2, color = "gray50") +
  geom_histogram(alpha = 0.8, show.legend = FALSE) +
  facet_wrap(vars(term))

################################################################################

# Regression model checks for the overall model for cases and for residuals

# Need to look up what all the temrs mean in the assumptions output

################################################################################

vif(fit) # Checks for multicollinearity

residuals(fit)

plot(fit, which = 1) # Plots the residuals versus fitted model

plot(fit, which = 2) # Plots the standardized residuals versus quantiles (Q-Q plot)

plot(fit, which =3) # Plots the standardized residuals versus the fitted (observed) model

outlierTest(fit) # Identifies outliers

durbinWatsonTest(fit) #Test if there is correlation among the residuals

options(max.print=10000)
```

```
inflRes <- influence.measures(fit)

summary(inflRes)

# Identifies potential influential observations to remove?

influenceIndexPlot(fit)

#influence.measures(fit) #Shows the influence measures for all cases. Remove the # if you want to view.

cooksDst <- cooks.distance(fit)

summary(cooksDst)

estnd <- rstandard(fit)

estud <- rstudent(fit)

par(mar=c(5, 4.5, 4, 2)+0.1)

hist(estud, main="Histogram Studentized Residuals", breaks="FD", freq=FALSE)

curve(dnorm(x, mean=0, sd=1), col="red", lwd=2, add=TRUE)

qqPlot(estud, distribution="norm", pch=20, main="QQ-Plot Studentized Residuals")

JarqueBeraTest(estud) #Checking normality of the residuals

shapiro.test(estud)
```

```
ncvTest(fit) # test for nonconstant error variance


bptest(fit) # Breusch-Pagan test of constant variance. If ncvTest is significant and this test is not, use th
```

APPENDIX C:

R Code for Factorial ANOVA

R Code for Factorial ANOVA

```
### Import the data
library(readxl)
setwd("C:/Users/9923100/OneDrive - Cherokee County School District/VSU/dissertation/Question 3")
RQ3E4<-read_excel("RQ3E4.xlsx")

install.packages("ARTool")
install.packages("emmeans")
install.packages("rcompanion")
install.packages("ggplot2")
install.packages("psych")
install.packages("multcomp")
install.packages("lsr")
install.packages("tidymodels")
install.packages("userfriendlyscience")
install.packages("rstatix")
install.packages("ggpubr")
library(ARTool)
library(emmeans)
library(rcompanion)
library(ggplot2)
library(psych)
library(car)
library(DescTools)
library(psych)
library(lsr)
library(multcomp)
#library(userfriendlyscience)
library(nortest)
library(tidymodels)
library(rstatix)
library(ggpubr)
######################################
# Factorial ANOVA -Two-Way ANOVA  #
######################################
# We are using our binned data to create our groups for this analysis
attach(RQ3E4)

str(RQ3E4)

#Note creates a factor (nominal level) from a numerical variable
RQ3E4$setting = as.factor(RQ3E4$setting)
RQ3E4$teatype = as.factor(RQ3E4$teatype)
RQ3E4$grade = as.factor(RQ3E4$grade)

str(RQ3E4)
summary(RQ3E4)
describe(RQ3E4)

RQ3E4$ed2<-cut(RQ3E4$ed,breaks=c(-1, 73.68, 100),labels = c("1","2"))
table(RQ3E4$ed2)

RQ3E4$ed2 <- as.factor(RQ3E4$ed2)
str(RQ3E4)

#create Z scores within groups
RQ3E4$zscore<-ave(RQ3E4$elamss,RQ3E4$setting,
        RQ3E4$ed2, FUN=scale)

#Show values of Z scores for outliers only
```

```
outliers<-RQ3E4%>%
  group_by(RQ3E4$setting, RQ3E4$ed2) %>%
  identify_outliers(zscore)
print(outliers,n=35,width=Inf)
help(identify_outliers)

xtabs(~RQ3E4$setting+RQ3E4$ed2)

RQ3E4 %>%
  group_by(setting,ed2) %>%
  identify_outliers(zscore)
print(outliers,n=100,width=Inf)

#displays all Z scores sorted by grouping and zscore

#too big to display
#print(RQ3E4[order(RQ3E4$setting,RQ3E4$ed2,RQ3E4$zscore),],n= 1500)
write.csv(RQ3E4[order(RQ3E4$setting,RQ3E4$ed2,RQ3E4$zscore),],file="RQ3E4_outliers.csv",row.names = F)

###Visualization for ela
bxp <- ggboxplot(
  RQ3E4, x = "setting", y = "elamss",
  color = "ed2", palette = "jco"
)
bxp

describeBy(RQ3E4$elamss,RQ3E4$ed2)
describeBy(RQ3E4$elamss,RQ3E4$setting)
describeBy(RQ3E4$elamss,RQ3E4$ed2:RQ3E4$setting)
```

```
# These next 9 - 10 lines do the YeoJohnson data transformation
set.seed(1234)
yj_rec <- recipe(elamss ~ setting + ed2, data = RQ3E4) %>%
  step_log (elamss)
yj_rec

yj_prep <- prep(yj_rec, retain = TRUE)
yj_prep

yj_juiced <- juice(yj_prep)

# This line shows the descriptive statistics for the transformed variable.
describe(yj_juiced)

#Generates descriptive statistics by binned variable: each separately and then the combination
describeBy(yj_juiced$elamss,yj_juiced$ed2)
describeBy(yj_juiced$elamss,yj_juiced$setting)
describeBy(yj_juiced$elamss,yj_juiced$ed2:yj_juiced$setting)

###Check for outliers
outliers<-yj_juiced %>%
  group_by(setting, ed2) %>%
  identify_outliers(elamss)
print(outliers,n=100)

###Visualization for ela
bxp <- ggboxplot(
  yj_juiced, x = "setting", y = "elamss",
  color = "ed2", palette = "jco"
)
bxp
#Untransformed, regular ANOVA
###################
# Factorial ANOVA #
###################
DP <- aov(elamss ~ ed2 + setting + ed2:setting, data= RQ3E4, na.action= na.exclude)
summary(DP)
TukeyHSD(DP)  # Post hoc test of significance by group
model.tables(DP, type = "means")
plot(TukeyHSD(DP))
drop1(DP,~.,test="F")   # type III SS and F Tests

Anova(DP, type="II")


RQ3E4$resids<-DP$residuals
#Procedures to check the normality assumption statistically
shapiro.test(RQ3E4$resids)
by(RQ3E4$resids, RQ3E4$ed2:RQ3E4$setting,shapiro.test)

#Procedures to check the homogeneity of variance assumption
leveneTest(RQ3E4$resids ~ RQ3E4$ed2:RQ3E4$setting,center=median)
leveneTest(RQ3E4$resids ~ RQ3E4$ed2:RQ3E4$setting,center=mean)

#Transformed, regular ANOVA
###################
# Factorial ANOVA #
###################
DP <- aov(elamss ~ ed2 + setting + ed2:setting, data= yj_juiced, na.action= na.exclude)
summary(DP)
```

```
TukeyHSD(DP)  # Post hoc test of significance by group
model.tables(DP, type = "means")
plot(TukeyHSD(DP))
drop1(DP,~.,test="F")   # type III SS and F Tests

Anova(DP, type="II")


yj_juiced$resids<-DP$residuals
#Procedures to check the normality assumption statistically
shapiro.test(yj_juiced$resids)
by(yj_juiced$resids, yj_juiced$ed2:yj_juiced$setting,shapiro.test)



#Procedures to check the homogeneity of variance assumption
leveneTest(yj_juiced$resids ~ yj_juiced$ed2:yj_juiced$setting,center=median)
leveneTest(yj_juiced$resids ~ yj_juiced$ed2:yj_juiced$setting,center=mean)

#Untransformed, Aligned ranks transformation ANOVA
model <- art(elamss~ed2+setting+ed2:setting,data=RQ3E4)
model

###Conduct ANOVA
anova(model)

###Posthoc Tests
library(emmeans)
model.lm = artlm(model, "ed2")
marginal = emmeans(model.lm,~ ed2)
pairs(marginal, adjust = "tukey")
cld(marginal, alpha=0.05, Letters=letters, adjust="tukey")
```

```
model.lm = artlm(model, "setting")
marginal = emmeans(model.lm,~ setting)
pairs(marginal, adjust = "tukey")
cld(marginal, alpha=0.05, Letters=letters, adjust="tukey")

model.diff = artlm(model, "ed2:setting")
marginal = emmeans(model.diff, ~ed2:setting)
contrast(marginal, method="pairwise", interaction=TRUE)
contrast(marginal, method="pairwise", adjust="tukey")

###Eta squared
Result = anova(model)
Result$part.eta.sq = with(Result, `Sum Sq`/(`Sum Sq` + `Sum Sq.res`))
Result

#Transformed, Aligned ranks transformation ANOVA
model1 <- art(elamss~ed2+setting+ed2:setting,data=yj_juiced)
model1

###Conduct ANOVA
anova(model1)

###Posthoc Tests
library(emmeans)
model.lm = artlm(model1, "ed2")
marginal = emmeans(model.lm,~ ed2)
pairs(marginal, adjust = "tukey")
cld(marginal, alpha=0.05, Letters=letters, adjust="tukey")
```

~~~
model.lm = artlm(model1, "setting")
marginal = emmeans(model.lm,~ setting)
pairs(marginal, adjust = "tukey")
cld(marginal, alpha=0.05, Letters=letters, adjust="tukey")

model.diff = artlm(model1, "ed2:setting")
marginal = emmeans(model.diff, ~ed2:setting)
contrast(marginal, method="pairwise", interaction=TRUE)
contrast(marginal, method="pairwise", adjust="tukey")

###Eta squared
Result = anova(model1)
Result$part.eta.sq = with(Result, `Sum Sq`/(`Sum Sq` + `Sum Sq.res`))
Result



```
~~~

APPENDIX D:

R Code for Factor Analysis

R Code for Factor Analysis

```
###Set Working Directory
setwd("C:/Users/9923100/OneDrive - Cherokee County School District/VSU/dissertation/Question 4")

###Install and load packages
#install.packages("psych")
#install.packages("car")
#install.packages("GPArotation")
#install.packages("readxl")
library(psych)
library(readxl)
library(GPArotation)
library(car)

#read in dataset
RQ4ELA<-read_excel("GADOE data rq4 ELA.xlsx")

#Shows structure of the variables as they are read in
str(RQ4ELA)

###Descriptive Statistics
#describe (RQ4ELA)
table(RQ4ELA$PS1)
table(RQ4ELA$PS2)
table(RQ4ELA$PS3)
table(RQ4ELA$PS4)
table(RQ4ELA$PS5)
table(RQ4ELA$PS6)
table(RQ4ELA$PS7)
table(RQ4ELA$PS8)
table(RQ4ELA$PS9)

table(RQ4ELA$PS10)


prop.table(table(RQ4ELA$PS1))*100

prop.table(table(RQ4ELA$PS2))*100

prop.table(table(RQ4ELA$PS3))*100

prop.table(table(RQ4ELA$PS4))*100

prop.table(table(RQ4ELA$PS5))*100

prop.table(table(RQ4ELA$PS6))*100

prop.table(table(RQ4ELA$PS7))*100

prop.table(table(RQ4ELA$PS8))*100

prop.table(table(RQ4ELA$PS9))*100

prop.table(table(RQ4ELA$PS10))*100


#Test number of factors in data

vss(RQ4ELA)

fa.parallel(RQ4ELA)


#check for multicollinearity

poly <- mixedCor(data=RQ4ELA, c=1, p = 1:10, d=NULL, smooth=TRUE, correct=FALSE,
global=TRUE, ncat=4, use="pairwise", method="pearson", weight=NULL)

poly$rho <- round((poly$rho), 2)

poly$rho


#Factor Analysis

factors_1 <- fa(r = poly$rho, nfactors =1, fm="wls")

factors_1

print(factors_1$loadings, cutoff = 0.3)

factors_2 <- fa(r = poly$rho, nfactors =2, fm="wls")

factors_2
```

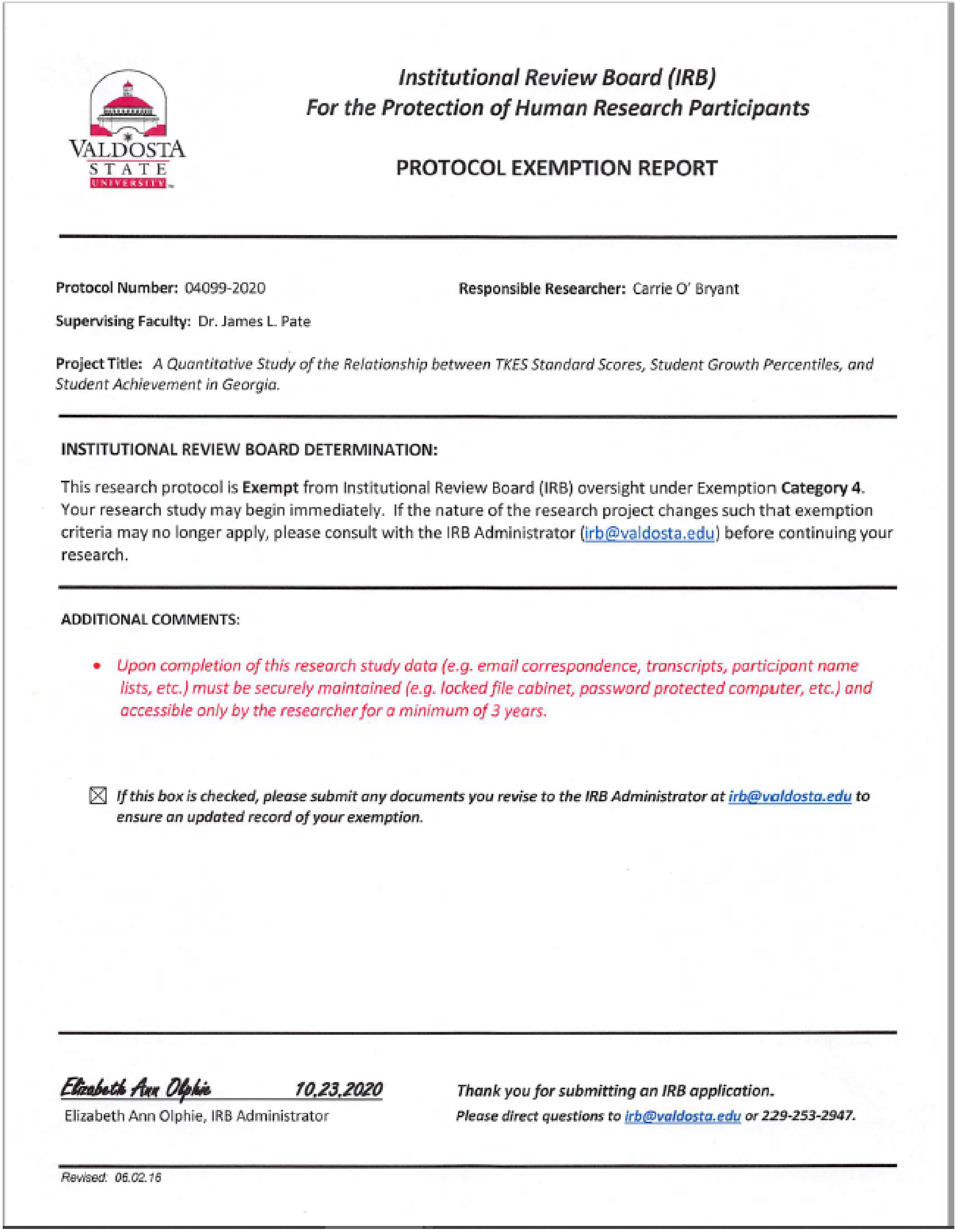
```
print(factors_2$loadings, cutoff = 0.3)
factors_3 <- fa(r = poly$rho, nfactors =3, fm="wls")
factors_3
print(factors_3$loadings, cutoff = 0.3)
factors_4 <- fa(r = poly$rho, nfactors =4, fm="wls")
factors_4
print(factors_4$loadings, cutoff = 0.3)
factors_5 <- fa(r = poly$rho, nfactors =5, fm="wls")
factors_5
print(factors_5$loadings, cutoff = 0.3)
factors_10 <- fa(r = poly$rho, nfactors =10, fm="wls")
factors_10
print(factors_10$loadings, cutoff = 0.3)

#Factor Analysis Diagrams
plot(factors_1)
fa.diagram(factors_1)
plot(factors_2)
fa.diagram(factors_2)
plot(factors_3)
fa.diagram(factors_3)
plot(factors_4)
fa.diagram(factors_4)
plot(factors_5)
fa.diagram(factors_5)
plot(factors_10)
fa.diagram(factors_10)

#Principal Components Analysis
prin1 <- principal(r = poly$rho, nfactors = 1, rotate="Promax")
```

```
prin1
print(prin1$loadings, cutoff = 0.3)

prin2 <- principal(r = poly$rho, nfactors = 2, rotate="Promax")
prin2
print(prin2$loadings, cutoff = 0.3)

prin3 <- principal(r = poly$rho, nfactors = 3, rotate="Promax")
prin3
print(prin3$loadings, cutoff = 0.3)

prin4 <- principal(r = poly$rho, nfactors = 4, rotate="Promax")
prin4
print(prin4$loadings, cutoff = 0.3)

prin5 <- principal(r = poly$rho, nfactors = 5, rotate="Promax")
prin5
print(prin5$loadings, cutoff = 0.3)

prin10 <- principal(r = poly$rho, nfactors = 10, rotate="Promax")
prin10
print(prin10$loadings, cutoff = 0.3)

#Scree Plot
plot(prin1$values, type = "b")
plot(prin2$values, type = "b")
plot(prin3$values, type = "b")
plot(prin4$values, type = "b")
plot(prin5$values, type = "b")
plot(prin10$values, type = "b")
```

```
#Run Bartlett's on raw dataframe
cortest.bartlett(RQ4ELA)

#Run Bartlett's on the correlation matrix
cortest.bartlett(poly$rho, n = 2800)

#Calculate KMO
KMO(RQ4ELA)

#Determinate of correlation matrix
det(poly$rho)
```

APPENDIX E:

IRB Protocol Exemption Report



VALDOSTA STATE UNIVERSITY

Institutional Review Board (IRB)
For the Protection of Human Research Participants

PROTOCOL EXEMPTION REPORT

Protocol Number: 04099-2020 **Responsible Researcher:** Carrie O' Bryant

Supervising Faculty: Dr. James L. Pate

Project Title: *A Quantitative Study of the Relationship between TKES Standard Scores, Student Growth Percentiles, and Student Achievement in Georgia.*

INSTITUTIONAL REVIEW BOARD DETERMINATION:

This research protocol is **Exempt** from Institutional Review Board (IRB) oversight under Exemption **Category 4**. Your research study may begin immediately. If the nature of the research project changes such that exemption criteria may no longer apply, please consult with the IRB Administrator (irb@valdosta.edu) before continuing your research.

ADDITIONAL COMMENTS:

- *Upon completion of this research study data (e.g. email correspondence, transcripts, participant name lists, etc.) must be securely maintained (e.g. locked file cabinet, password protected computer, etc.) and accessible only by the researcher for a minimum of 3 years.*

☒ ***If this box is checked, please submit any documents you revise to the IRB Administrator at irb@valdosta.edu to ensure an updated record of your exemption.***

Elizabeth Ann Olphie *10.23.2020*
Elizabeth Ann Olphie, IRB Administrator

Thank you for submitting an IRB application.
Please direct questions to irb@valdosta.edu or 229-253-2947.

Revised: 06.02.16