THE UTILITY OF A STANDARDS-BASED TEACHER EVALUATION AS A MEASURE OF EFFECTIVENESS

By

Richard P. Akers

Dissertation Submitted

in Partial Fulfillment of Requirements for the Degree

Doctor of Education

College of Education

Frostburg State University

December 2016

THE UTILITY OF A STANDARDS-BASED TEACHER EVALUATION

AS A MEASURE OF EFFECTIVENESS

By

Richard P. Akers

The undersigned, appointed by the Dean of the College of Education, have examined and approved this dissertation submitted in partial fulfillment of requirements for the degree of Doctor of Education.

Chair — October 26, 2016

Date

John Stoothoff, Ph.D.
Associate Professor of Education
Frostburg State University

Jamelyn Tobery-Nystrom, Ed.D.
Associate Professor of Education
Frostburg State University

April Bishop, Ed.D.
Supervisor of Mathematics and Science
Washington County Public Schools

Acknowledgments

I extend my sincere gratitude for the support and guidance provided to me by my Committee Chair, Dr. John Stoothoff. Our discussions and the insightful feedback you provided to me throughout the dissertation process were invaluable.

I also extend my gratitude to my committee members, Dr. Jamelyn Tobery-Nystrom and Dr. April Bishop. I greatly appreciate your time, expertise, and encouragement.

I also extend my gratitude to Dr. Glenn Thompson for his leadership, teaching, and unique ability to provide timely encouragement.

To all who assisted with the collection of data or provided other assistance or encouragement, I thank you.

I am forever grateful for the encouragement provided by my parents, Bob and Verlene Akers and my brothers Tony and Jeff Akers.

I am especially grateful for the inspiration provided for this and all my endeavors by my wife, Anne, and my sons Alex, Nieko, and Mitchell.

Abstract

THE UTILITY OF A STANDARDS-BASED TEACHER EVALUATION AS A MEASURE OF EFFECTIVENESS

By

Richard P. Akers

The purpose of this study was to examine the correlation between teachers' evaluation ratings and their students' attainment of projected growth scores in reading. The independent variables were the Charlotte Danielson *Framework for Teaching* (FFT) component ratings assigned to teachers of third-, fourth-, and fifth-grade students in a western Maryland school district during the 2013–2014 and 2014–2015 school years. The dependent variable was the percentage of each evaluated teachers' students attaining projected growth scores in reading, measured by the MAP reading assessment. A multiple correlation analysis determined teachers' performance ratings on all 22 FFT component ratings, explaining 11.5% (R^2 = .115) of the variance in the percentage of teachers' students attaining projected growth scores. The adjusted R^2 was .5% (adj. R^2 = .005) and was not statistically significant ($p > .05$). The unstandardized coefficient for one component, 3d: Using Assessment in Instruction, was 8.139 and was statistically significant ($p < .05$), meaning teachers rated Distinguished had an 8.139% higher mean percentage of their students attaining projected growth scores in reading than teachers rated Basic or Proficient. Teachers' ratings on 10 other FFT components had positive correlations to students' attainment of growth projections in reading; however, the correlations were not statistically significant ($p > .05$).

Keywords: education reform, teacher evaluation, teacher effectiveness

Table of Contents

List of Tables

Chapter 1 – Introduction

In recent years, the education reform-and-accountability movement has encompassed teacher evaluation as a strategy to improve public education. In 2009, the Race to the Top (RTTT) grant encouraged public education agencies to use rigorous teacher evaluations to improve the quality of instruction provided to public school students (U.S. Department of Education [USDE], 2009). In 2011, the USDE offered waivers that exempted recipients from 10 Elementary and Secondary Education Act (ESEA, 1965) accountability provisions (USDE, 2016). To receive an ESEA waiver, states had to agree to implement new reform strategies, including the use of rigorous teacher evaluations (USDE, 2016). The waiver required teacher evaluations to include an assessment of professional practices and student growth to provide summative ratings of *ineffective*, *effective*, or *highly effective* to inform personnel and professional-development decisions (USDE, 2016). By 2015, 46 states had received ESEA waivers (USDE, 2016, p. 1).

Including student-growth measures in teacher evaluations was and continues to be controversial due to concerns about the validity of student-achievement tests, methods used to control for categorical differences among students, and the high percentage of teachers who do not teach assessed subjects or grade levels (Darling-Hammond, Amrein-Beardsley, Haertel, & Rothstein, 2012; Ehlert, Koedel, Parsons, & Podgursky, 2013). Despite the ESEA waiver requirement to include measures of student growth in the new evaluations (USDE, 2016), local education agencies continued to rely on observation-based evaluations to assess the performance of professional practices and to inform decisions regarding tenure, promotions, terminations, and professional development (Goldring et al., 2015).

Most states that received ESEA waivers implemented new teacher-evaluation systems to measure the performance of professional practices (Steinberg & Donaldson, 2014). Many of those states implemented standards-based evaluations grounded in standards developed by professional organizations and research on teacher effectiveness (Steinberg & Donaldson, 2014). The use of professional standards and teacher-effectiveness research is intended to ensure evaluations measure professional practices that impact student learning (Morgan, Hodge, Trepinksi, & Anderson, 2014; Steinberg & Donaldson, 2014). Although each standards-based evaluation model has a distinctive approach and emphasis for measuring teacher effectiveness, most models organize various teaching practices into four or more categories and provide rubrics and rating scales that differentiate three or more levels of performance (Danielson, 2007; Marshall, 2009; Marzano, 2007; Morgan, et al. 2014; Steinberg, & Donaldson, 2014; Stronge, 2012).

Hundreds of school districts in states that received ESEA waivers (Danielson Group, 2011) chose to use Danielson's (2007) *Framework for Teaching* (FFT) standards-based evaluation system. Danielson asserted that the FFT aligns to standards developed by the Interstate New Teacher Assessment and Support Consortium (InTasc; Council of Chief State School Officers, 2011). The FFT assesses the performance of 22 professional practices, referenced as components, and each component aligns with one or more of the InTasc standards. Danielson also asserted that the relevance of the components to student learning is supported by teacher-effectiveness research. Research on the utility of standards-based evaluations as measures of effectiveness is limited (Balch & Koedel, 2014; Good & Lavigne, 2014; Morgan et al., 2014). The present study examines the correlation between teachers' evaluation ratings on the FFT and student learning.

Statement of the Problem

Before recent efforts to increase the rigor of teacher evaluations, 98% of teachers received performance ratings of *satisfactory* (Weisberg, Sexton, Mulhern, & Keeling, 2009). The failure of teacher evaluations to differentiate levels of performance prevented school districts from using the evaluations to inform personnel or professional-development decisions (Kane, 2012; Marzano, 2012). The recent influence of professional standards and teacher-effectiveness research on teacher evaluations has not significantly improved the ability of the evaluations to identify variations in teacher performance (Kraft & Gilmour, 2016). In 19 states that implemented new evaluations to meet ESEA waiver requirements, only 3% of teachers received ratings of less than effective (Kraft & Gilmour, 2016, p.1).

In contrast to the homogeneous evaluation ratings assigned to teachers, significant variations exist in the ability of teachers to facilitate the academic achievement of students (Hanushek, 1971; Marzano, 2003; Rockoff, 2004; Wright, Horn, & Sanders, 1997). As much as 7% of the variation in student achievement is due to the effectiveness of the students' school, and 13% of the variation is due to the effectiveness of the students' teachers (Marzano, 2003, p. 74). Variations in teacher effectiveness have a greater impact on the academic achievement of students than class size or categorical differences among students (Hanushek, 1971; Rockoff, 2004; Wright et al., 1997). Also, positive correlations exist between specific professional practices and increased student learning (Brophy & Good, 1986; Kane, Taylor, Tyler, & Wooten, 2011). Despite efforts to use professional standards and teacher effectiveness research to improve teacher evaluation systems, the dichotomy between performance ratings and the actual effects teachers have on student learning persists (Kraft & Gilmour, 2016).

The ESEA waiver requires evaluations to provide summative ratings of *ineffective*, *effective*, or *highly effective* (USDE, 2016) to indicate a level of competence at facilitating student learning. Evaluation systems that rate nearly all teachers *effective* or *highly effective* are probably not providing accurate effectiveness ratings (Kraft & Gilmour, 2016; Weisberg et al., 2009). The problem with using evaluations that do not accurately differentiate levels of teacher effectiveness is that school districts cannot reward highly effective teachers, provide targeted professional development to less effective teachers, or justify personnel decisions related to competency (Weisberg et al., 2009).

Purpose and Rationale of the Study

The purpose of this study was to examine the correlation between teachers' evaluation ratings and their students' attainment of projected growth scores in reading. The independent variables were the FFT component ratings assigned to teachers of third-, fourth-, and fifth-grade students in a western Maryland school district during the 2013–2014 and 2014–2015 school years. The dependent variable was the percentage of each of the evaluated teachers' students attaining projected growth scores in reading, measured by the Measures of Academic Progress (MAP) reading assessment. A multiple correlation analysis determined the proportion of the variance in student attainment of projected growth scores in reading explained by teachers' ratings on all 22 components and by teachers' ratings for each individual component. The rationale for the study is to contribute to understanding of the relationship between standards-based evaluation criteria and students learning. An examination of the correlation between all 22 FFT components and a nationally normed value-added measure of student learning would broaden understanding of the strength of FFT's relationship to student learning and the

FFT's potential utility for informing personnel and professional-development decisions.

Significance of the Study

Most of the 46 states that received ESEA waivers now use standards-based teacher-evaluation systems (Steinberg & Donaldson, 2014). By 2011, Arkansas, Delaware, Idaho, and South Dakota had implemented the FFT statewide (Danielson Group, 2011). Additionally, hundreds of school districts in California, Florida, Illinois, Maryland, Michigan, New Jersey, New York, Ohio, Pennsylvania, and Washington are using the FFT to evaluate teachers (Danielson Group, 2011; Maryland State Department of Education [MSDE], 2014a; Michigan Department of Education, 2014). With the number of school districts that rely on the FFT evaluation to inform personnel and professional-development decisions, it is important to investigate its utility toward that purpose.

Conceptual Framework

The framework of this study was predicated on three related concepts:

1. Teacher evaluations should provide accurate and relevant information that can be used to improve teacher performance, make personnel decisions, and ultimately improve student learning (Danielson & McGreal, 2000; Gullickson, 2009; Joint Committee on Standards for Educational Evaluation [JCSEE], 1988).
2. The relevance of the information provided by teacher evaluations rests on the strength of the association between the evaluation criteria and student learning (Gullickson, 2009; Marzano, 2012; Milanowski, 2004; Milanowski, Kimball, & White, 2004; Nye, Konstantopoulos, & Hedges, 2004).
3. The strength of the association between evaluation criteria and student learning can be determined by measuring the strength of the correlation between changes

in the level of performance of the evaluation criteria and changes in the level of student learning (Kane, 2012; Kane & Cantrell, 2013; Marzano, 2012; Milanowski et al., 2004).

The variables in this study were teacher ratings on evaluation criteria and a measure of student learning. The independent variables were the FFT component ratings assigned to teachers of third-, fourth-, and fifth-grade students in a western Maryland school district during the 2013–2014 and 2014–2015 school years. The dependent variable was the percentage of each of the evaluated teachers' students attaining projected growth scores in reading, measured by the MAP reading assessment.

Research Design and Overview

This study applied a multiple correlation analysis to FFT performance ratings assigned to teachers during evaluations and the percentage of each evaluated teacher's students attaining projected growth scores in reading. The independent variables were the FFT component ratings assigned during the 2013–2014 and 2014–2015 school years to the teachers of third-, fourth-, and fifth-grade students in a western Maryland School district. The dependent variable was the percentage of each of the evaluated teachers' students attaining projected growth scores in reading, measured by the MAP reading assessment. The multiple correlation analysis determined the proportion of the variance in the percentage of students attaining projected growth scores in reading explained by teachers' performance ratings on all 22 FFT components and by the teachers' performance ratings on individual components.

Research Questions

1. Do teachers' performance ratings on all 22 FFT components explain a significant proportion of the variance in the percentage of teachers' students attaining

projected growth scores in reading?

2. Do teachers' performance ratings on any individual FFT component explain a significant proportion of the variance in the percentage of teachers' students attaining projected growth scores in reading?

Assumptions

One assumption is that interrater reliability among administrators who conducted the teacher evaluations was sufficient to provide accurate teacher ratings. All evaluators held Maryland Administrator I or II certificates. All evaluators performed teacher evaluations as a function of positions held as principals, assistant principals, or subject supervisors. The year before implementation, representatives of the Danielson Group trained school district evaluators to use the FFT, including performance-level descriptors (Danielson, 2009). During the first 2 years of implementation, the evaluators continued training with the school district's Coordinator of Teacher Evaluation and Professional Development.

Another assumption is that student growth in reading provides an adequate measure of teacher effectiveness. Teaching reading was a primary responsibility of teachers included in the study, and the MAP growth projections in reading are based on the median level of growth among students in the same grade with the same score at the beginning of the instructional year in the Northwest Evaluation Association's (NWEA, 2012) norming sample. In general, 50% of a teacher's students are expected to fall below the median level of growth, and about 50% of a teacher's students are expected to fall above the median level of growth (Jensen, 2013). This study determined if higher or lower FFT evaluation ratings relate to a higher or lower percentage of students meeting growth projections.

Findings

Regarding Research Question 1, a multiple correlation analysis determined that teachers' collective component ratings explained 11.5% (R^2 = .115) of the variance in the percentage of students' attaining projected growth scores. The adjusted R^2 was .5% and was not statistically significant ($p > .05$). The multiple correlation analysis determined that the FFT evaluation system does not provide relevant information that can be used to improve teacher performance, make personnel decisions, and ultimately improve student learning.

Regarding Research Question 2, a multiple correlation analysis determined teachers rated *Distinguished* on component *3d: Using Assessment in Instruction* had an 8.139% higher mean percentage of students attaining projected growth scores in reading than teachers rated *Basic* or *Proficient*. The correlation coefficients for the other 21 FFT components ranged from -5.293 to 4.723, and were not statistically significant ($p > .05$). The multiple correlation analysis determined only one *FFT* component provided relevant information that can be used to improve teacher performance, make personnel decisions, and ultimately improve student learning.

Limitations

This study had several limitations:

- I examined teacher-evaluation data generated in the first 2 years of implementation of the FFT in Maryland.
- I collected data from one school district and results may not be applicable to school districts with different curricula or resources.
- I only assumed interrater reliability among evaluators.
- Study analysis included only teachers of third-, fourth-, and fifth-grade students.

- The only measure of teacher effectiveness was growth in reading.
- I did not control for variations in teacher assignments such as class size or the student demographic composition of schools or classrooms.
- I did not control for variations in teacher qualifications such as years of experience or educational attainment.

Delimitations

To control for outside variables, I confined data collection and analysis to the evaluative ratings assigned by administrators using the FFT to evaluate teachers of third-, fourth-, and fifth-grade students in one school district. I did not include evaluations lacking ratings for all FFT components in the study. I did not include MAP data that did not include scores from the beginning and end of the year in the study.

Definitions

Basic: An FFT rating that indicated a teacher understands the concepts underlying a particular component or domain, but is not successful in its implementation (Danielson, 2007, p. 39).

Component: An evaluated teaching practice and a subcategory of a domain of teaching (Danielson, 2007, p. 23).

Distinguished: An FFT rating that indicated a teacher has attained mastery of a particular component or domain and was functioning at the highest level of performance (Danielson, 2007, p. 40).

Domains: The four categories of professional practices that include *planning and preparation, the classroom environment, instruction,* and *professional responsibilities* (Danielson, 2007, p. 22).

Framework For Teaching (FFT): A standards-based evaluation instrument that

assesses 22 professional practices, referenced as components (Danielson, 2007).

Levels of performance: The ratings teachers received based on the competence with which they performed particular components or domains. The levels include *unsatisfactory, basic, proficient,* and *distinguished* (Danielson, 2007, p. 38).

Levels of performance descriptors: Rubrics that describe teacher and student behaviors that indicate the level of competence with which teachers perform particular components or domains (Danielson, 2007).

Measure of academic progress (MAP): The NWEA's (2012) assessment system.

Projected growth scores: The growth in reading educators expect students to achieve based on the median level of growth achieved by past students in the same grade with the same score at the beginning of an instructional year (NWEA, 2012).

Proficient: An FFT rating that indicates the teacher clearly understands the concepts underlying a particular component or domain and implements it well (Danielson, 2007, p. 40).

Unsatisfactory: An FFT rating that indicates the teacher does not understand the concepts underlying a particular component or domain (Danielson, 2007, p. 39).

Summary

Federally led education reforms now include the use of rigorous teacher evaluations. Many states have implemented standards-based evaluations to meet the requirements of a federal grant and the ESEA waiver (USDE, 2009, 2016). The use of professional standards does not appear to have improved the utility of teacher evaluations for informing personnel and professional-development decisions (Kraft & Gilmour, 2016). School districts in many states, including Maryland, have implemented the FFT standards-based evaluation (Danielson Group, 2011). The utility of a teacher-evaluation

system for informing personnel and professional-development decisions depends on the relationship between the criteria used to interpret and judge the performance of teachers and student learning (Gullickson, 2009; Kane, 2012; Marzano, 2012). Stakeholders can determine the accuracy and value of the information provided by teacher evaluations by measuring the strength of the association between the level of performance of the evaluation criteria and the level of student learning (Kane, 2012; Kane & Cantrell, 2013; Marzano, 2012; Milanowski et al., 2004).

The purpose of this study was to examine the correlation between teachers' evaluation ratings and their students' attainment of projected growth scores in reading. The independent variables were the FFT component ratings assigned to teachers of third-, fourth-, and fifth-grade students in a western Maryland school district during the 2013–2014 and 2014–2015 school years. The dependent variable was the percentage of each of the evaluated teachers' students attaining projected growth scores in reading, measured by the MAP reading assessment. A multiple correlation analysis determined the proportion of the variance in student attainment of projected growth scores in reading explained by teachers' performance ratings on all 22 FFT components and by teachers' performance ratings on each individual component.

Organization of the Study

This study follows a five-chapter format. Chapter 1 introduced the topic of teacher evaluations and the purpose of the study. Chapter 2 presents a review of the literature apposite to teacher evaluations. Chapter 3 explains the methodology of this study. Chapter 4 presents the findings of this study. In Chapter 5, I discuss the conclusions and implications of the research. References and appendices follow Chapter 5.

Chapter 2 – Review of the Literature

Introduction

The purpose of this study was to examine the correlation between teachers' evaluation ratings and their students' attainment of projected growth scores in reading. The independent variables were the FFT component ratings assigned to teachers of third-, fourth-, and fifth-grade students in a western Maryland school district during the 2013–2014 and 2014–2015 school years. The dependent variable was the percentage of each of the evaluated teachers' students attaining projected growth scores in reading, measured by the MAP reading assessment. A multiple correlation analysis determined the proportion of the variance in the percentage of students attaining projected growth scores in reading, explained by teachers' performance ratings on all 22 FFT components and by teachers' performance ratings on each individual component. Chapter 2 reviews the work of authors and researchers regarding the conceptual framework of this study, the FFT, the MAP reading assessment, and topics apposite to teacher evaluation and measures of student learning.

Conceptual Framework

The framework for this study rested on three related concepts. First, teacher evaluations should provide accurate and relevant information that can be used to improve teacher performance, make personnel decisions, and ultimately improve student learning (Danielson & McGreal, 2000; Gullickson, 2009; JCSEE, 1988). Using teacher evaluation to inform personnel and professional-development decisions as a strategy for improving public education (USDE, 2016) assumes evaluations provide accurate assessments of the performance of relevant criteria. The expectation that evaluations provide accurate and relevant information is well-established (Gullickson, 2009; JCSEE, 1988). JCSEE

established 27 standards for the evaluation of educators (Gullickson, 2009), organized into four categories: propriety, utility, feasibility, and accuracy (Gullickson, 2009). Two of the six utility standards for personnel evaluations address the value of evaluation criteria and the accuracy and relevance of reports regarding the performance of the evaluation criteria (Gullickson, 2009; JCSEE, 1988). The JCSEE listed the *explicit criteria* and *functional reporting* utility standards among the most applicable evaluation standards for selecting an evaluation system and for directing staff development (Gullickson, 2009). The JCSEE also lists the explicit criteria standard among the most applicable evaluation standards for making personnel decisions (Gullickson, 2009). The explicit criteria utility standard requires evaluations to identify and justify the criteria used to interpret and judge the performance of personnel (Gullickson, 2009). The functional reporting utility standard requires evaluations to provide reports that are accurate, relevant, and of value (Gullickson, 2009).

A second concept contributing to the framework of this study is that the relevance of the information provided by teacher evaluations rests on the strength of the association between the evaluation criteria and student learning (Kane, 2012; Marzano, 2012; Milanowski, 2004; Milanowski et al., 2004). However, teacher evaluations often measure criteria that are not relevant to student learning (Kraft & Gilmour, 2016; Weisberg et al., 2009). The third concept contributing to the framework of this study is that the strength of the association between evaluation criteria and student learning can be determined by measuring the strength of the correlation between changes in the level of performance of the evaluation criteria and changes in the level of student learning (Kane & Cantrell, 2013; Marzano, 2012; Milanowski, 2004; Milanowski et al., 2004).

The framework for teaching and the Maryland-model evaluation. The

independent variables in this study were the FFT component ratings assigned to 200 teachers of third-, fourth-, and fifth-grade students in a western Maryland school district. The FFT is grounded on standards developed by the Educational Testing Service (ETS) for the Praxis III new teacher assessments and aligns to the InTasc standards (Danielson, 2007). Each domain of the FFT is grounded on empirical and theoretical research and the recommendations and writings of experts and leading authorities in the fields of teaching and evaluation (Danielson, 2007).

The FFT measures four categories of professional practices, referred to as domains. The four domains are (a) *Planning and Preparation*, (b) *Classroom Environment*, (c) *Instruction*, and (d) *Professional Responsibilities* (Danielson, 2007). Each domain includes five or six components. Components are measureable professional practices, and teachers receive ratings of distinguished, proficient, basic, or unsatisfactory for the performance of each of the 22 components (Danielson, 2007). The education agencies that use the FFT determine the value of each component rating and domain rating to calculate a summative rating. For example, on the Maryland state model evaluation, the component ratings convert to numerical scores. The numerical component scores are added together to determine ratings for each domain. The equally weighted domain ratings convert to numerical scores that are added together to determine an effectiveness rating of highly effective, effective, or ineffective for professional practices. The rating for professional practices combined with a rating for student growth, scored with student learning objectives (SLOs) determine a summative evaluation rating of highly effective, effective, or ineffective (MSDE, 2014a).

The FFT is an evaluation system. In addition to providing an observation instrument listing components categorized by domains to be evaluated, the FFT

evaluation system provides descriptions that differentiate levels of performance for each component. The *level of performance* (Danielson, 2007) rubrics describe each level of performance for two to five elements for each component. The element shown in Table 1, *student participation*, is one of three elements for component *3b: using questioning and discussion techniques* (Danielson, 2007).

Table 1

Level of Performance Description for One Element

Element	Unsatisfactory	Basic	Proficient	Distinguished
Student Participation	A few students dominate the discussion	Teacher attempts to engage all students in the discussion, but with only limited success	Teacher successfully engages all students in the discussion	Students themselves ensure that all voices are heard.

Note. From *Enhancing Professional Practice: A Framework for Teaching* (p. 82), by C. Danielson, 2007, Alexandria, VA: Association for Supervision and Curriculum Development. Copyright 2007 by the Association for Supervision and Curriculum Development. Reprinted with permission.

The FFT evaluation system also provides a process for collecting evidence to determine levels of performance, a process for conducting evaluations, and a process for conducting preobservation and postobservation conferences with teachers (Danielson, 2007). The conferences provide teachers the opportunity to reflect on and strengthen the performance of their professional practices (Danielson, 2007). Appendix A lists the FFT domains and the associated components.

Even before receiving an ESEA waiver, most of Maryland's school districts were in the process of implementing new teacher-evaluation systems as a condition of the RTTT grants awarded under the American Recovery and Reinvestment Act of 2009 (ARRA, 2009; MSDE, 2012, 2013, 2014a). After receiving an ESEA waiver, the MSDE developed a model teacher evaluation, and school districts had the option to use the state model or submit a locally developed evaluation model for approval by MSDE (2013, 2014a).

Maryland's model teacher evaluation weights professional practice as 50% of the evaluation and various measures of student growth as the other 50% of the evaluation (MSDE, 2013, 2014a). The state model uses the FFT to assess the performance of professional practices and weights each FFT domain at 12.5% of the total evaluation. The four domain ratings are based on the ratings assigned to the components associated with each domain. The rating scale for components ranges from 1 to 4: Unsatisfactory = 1, Basic = 2, Proficient = 3, and Distinguished = 4 (MSDE, 2014a). The school district from which the data for this study were collected uses the same domain weighting and component-scoring system for professional practice as the Maryland state model. The performance ratings that teachers in the school district received for each FFT component during the 2013–2014 and 2014–2015 school were the independent variables in this study.

Research on the framework for teaching. Research on the relationship between the FFT and student achievement has generally focused on less than half of the components and has found weak to moderate correlations between those components and various measures of student learning. The student achievement variable in most of the studies was a value-added score based on a comparison of predicted scores to actual scores on standardized assessments. The predicted scores were based on students' scores the previous year (Borman & Kimball, 2005; Holtzapple, 2003; Kane & Staiger, 2012; Milanowski et al., 2004; Sartain, Stoelinga, & Brown, 2011).

Holtzapple (2003) conducted a study on the correlation between FFT ratings and student achievement in the Cincinnati Public Schools system and found weak to moderate correlations between teachers' summative FFT ratings and student achievement on state and district assessments. The students of teachers who had the lowest FFT scores

generally demonstrated lower growth, and students of teachers who had the highest FFT scores generally demonstrated higher growth. The correlation coefficients ranged from 0.28 to 0.37, depending on the subject matter (Holtzapple, 2003).

Milanowski et al. (2004) conducted studies in three locations, the Cincinnati Public Schools system in Ohio, the Vaughn Next Century Learning Center charter school in California, and the Washoe County School District in Nevada, to determine the strength of the correlation between student achievement and teachers' ratings on evaluations using varying portions of the FFT. All three studies found a weak to moderate positive correlation between the evaluation ratings and student achievement. None of the studies examined all FFT components individually (Milanowski et al., 2004).

The Cincinnati Public Schools system study examined the correlation between teacher-evaluation ratings and student achievement in science, mathematics, and reading for students in Grades 3 through 8 (Milanowski et al., 2004). The Cincinnati Public Schools used all four FFT domains in the evaluations. The researchers used the average of the four domain ratings as an overall composite rating and the value-added student achievement data to calculate the correlations. The average correlation coefficients between the composite evaluation ratings and student achievement were .27 for science, .43 for mathematics, and .32 for reading (Milanowski et al., 2004).

The Vaughn Next Century Learning Center study examined the correlation between evaluation ratings and student achievement in language arts, reading, and mathematics (Milanowski et al., 2004). The Vaughn Next Century Learning Center evaluated teachers on 12 domains, including subject-specific domains and some FFT domains. The study focused on five domains: *Lesson Planning, Classroom Management, Literacy, Language Development,* and *Mathematics*. The researchers used the average

domain ratings and the value-added student-achievement data to calculate the correlations. The average correlation coefficients between the evaluation ratings and student achievement were .18 for language arts, .50 for reading, and .21 for mathematics (Milanowski et al., 2004).

The Washoe County School District study examined the correlation between evaluation ratings and student achievement in reading and mathematics. The Washoe County School District evaluates probationary teachers on all four FFT domains and evaluates experienced teachers on either one or two domains, based on an established evaluation cycle. The researchers used the average of the domain scores and the value-added student-achievement data to calculate the correlations. The average correlation coefficients between the evaluation ratings and student achievement were .21 for reading, and .19 for mathematics (Milanowski et al., 2004).

Borman and Kimball (2005) conducted another study in Washoe County School District in Nevada to examine the correlation between FFT evaluation ratings and student achievement in reading and mathematics for subcategories of students. The purpose of the study was to determine if teachers with higher ratings on their evaluations were more effective at closing achievement gaps than teachers with lower evaluation ratings (Borman & Kimbell, 2005). The researchers found no evidence that teachers with high evaluation ratings were more effective at closing achievement gaps than teachers with low evaluation ratings. The researchers examined changes to the achievement gaps between historically low-achieving and high-achieving students, students from low-income and high-income families, and minority and White students (Borman & Kimbell, 2005).

Sartain et al. (2011) conducted a study in the Chicago Public Schools system to

examine the correlation between teacher evaluation ratings and student achievement on mathematics and reading assessments. The study focused on the component ratings of Domains 2 and 3 of the FFT. The researchers found that, on average, students of teachers rated *unsatisfactory* have lower achievement scores than students of teachers rated *basic*, students of teachers rated *basic* have lower achievement scores than students of teachers rated *proficient*, and students of teachers rated *proficient* have lower achievement scores than students of teachers rated *distinguished* (Sartain et al., 2011). The study did not analyze the components of Domains 1 and 4.

Another study analyzed five evaluation instruments, including the FFT, to evaluate 3,000 teachers in seven school districts from seven states (Kane & Staiger, 2012). The researchers focused on the teaching practices that were observable in the classroom. For the FFT, the researchers focused on Domains 2 and 3. The researchers used the average of the domain scores and the value-added student-achievement data to calculate correlations. The study found a weak to moderate correlation between observation ratings and student-achievement gains and found that none of the observation instruments were as predictive of student-achievement gains as a teacher's past record of student-achievement gains (Kane & Staiger, 2012).

Kane and Staiger (2012) reported the findings in terms of the increase in the number of months of academic growth students experienced relative to the ratings teachers received on the two FFT domains. One month of growth is the median level of growth students achieved in a school year divided by nine, which is the number of months students are in school. The difference in student-learning gains between teachers rated in the top 25% and teachers rated in the bottom 25% on Domains 2 and 3 of the FFT was the equivalent of approximately 2.7 months of schooling in mathematics and

less than 1 month in English-language arts. For comparison, the researchers calculated the difference in student learning based on teachers with the highest and lowest years of experience and between teachers with and without master's degrees and found the most experienced teachers increased student learning by .5 months and teachers with master's degrees increase learning by 1 month (Kane & Staiger, 2012).

To determine if using multiple measures of teacher effectiveness improved the accuracy of the evaluation ratings, Kane and Staiger (2012) combined teachers' ratings from student surveys, past records of student achievement gains, and FFT ratings. The difference in student gains between teachers in the top 25% and teachers in the bottom 25% of scores on the evaluations with multiple measures of teacher effectiveness increased to 8 months of schooling in mathematics and 2.5 months in English-language arts (Kane & Staiger, 2012).

In summary, researchers demonstrated a correlation between teacher ratings on a subset of FFT components and a variety of value-added achievement data with correlation coefficients that range from .18 to .50 (Borman & Kimbell, 2005; Holtzapple, 2003; Kane & Staiger, 2012; Milanowski et al., 2004; Sartain et al., 2011). An examination of the correlation between all 22 FFT components and a nationally normed measure of student learning would broaden understanding of the strength of the relationship between the FFT and student learning and the FFT's potential utility for informing personnel and professional-development decisions.

Measures of academic progress. The dependent variable in this study was the percentage of each of the evaluated teachers' students attaining projected growth scores in reading, measured by the MAP reading assessment. The school district in which this study was conducted assesses student growth in reading with the NWEA's MAP reading

assessment (NWEA, 2012). The MAP assessments are computer-adaptive achievement tests in mathematics and reading. The computer adjusts the difficulty of the questions so each student takes a unique test. The difficulty of questions change based on how well the student has answered previous questions. After students take the beginning of the instructional year MAP reading assessment, the school district receives projected growth scores from NWEA for each student who took the assessment. The projected growth scores are set at the median level of growth among students in the same grade with the same score at the beginning of the instructional year in the NWEA's norming sample (NWEA, 2012). Almost 20% of U.S. school districts use the NWEA's MAP assessments (Cordray, Pion, Brandt, Molefe, & Toby, 2012). The NWEA maintains the largest repository of student growth data in the U.S. (Cordray, et al., 2012). Researchers analyzed MAP testing data from 10 states and found assessment scores were consistent across grade levels in all states (Wang, McCall, Hong, & Harris, 2013). Some school districts incorporate MAP assessments into teacher evaluations by setting goals for the percentage of a teacher's students who meet individual growth projections (Jensen, 2013: Steinberg & Donaldson, 2014).

Several studies focused on or used MAP assessments (Andren, 2010; Gilbert, 2014; Martin, 2014; Washington, 2011). Andren (2010) found that the MAP reading assessment consistently predicted student performance on the New England Common Assessment Program reading assessment for the general student population and for students receiving supplemental reading support. Martin (2014) found that the MAP reading-assessment scores of first-grade students moderately related to scores on the Woodcock-Johnson Tests of Achievement reading assessment in the spring of second grade. Martin also found that MAP reading scores of second-grade students were

moderately to strongly related to their second grade Woodcock-Johnson reading-assessment scores (Martin, 2014). Washington (2011) found that MAP reading and mathematics scores moderately correlated to teachers' summative-evaluation ratings. MAP reading-assessment scores were the dependent variable in a study that examined the extent to which teachers' summative evaluation ratings and other teacher characteristics predicted student learning in reading (Gilbert, 2014).

Teacher-Effectiveness Research

The analyses of relationships between professional practices and student-achievement gains have played a significant role in research on teacher effectiveness (Brophy & Good, 1986; Tyler, Taylor, Kane, & Wooten, 2010). Researchers examined correlations between the academic growth of students and teacher behaviors (Brophy & Good, 1986; Tyler et al., 2010) and the correlations between student-achievement gains and teacher characteristics such as level of education and years of experience (Kane & Staiger, 2012; Nye et al., 2004). Researchers also examined differences in student-achievement gains among students assigned to different teachers (Hanushek, 1971; Nye et al., 2004; Rockoff, 2004; Wright et al., 1997).

Most studies on teacher effectiveness have relied on student achievement on standardized tests as a measure of teacher effectiveness (Nye et al., 2004). For example, Hanushek (1971) used the results of students' Stanford Achievement Test scores to determine the impact of teachers and other factors on student achievement. Wright et al. (1997) and Rockoff (2004) used the results of the Comprehensive Test of Basic Skills to measure teacher effects on student achievement.

Standards-Based Teacher Evaluations

Standards-based teacher evaluations build on concepts of teaching developed

using empirical and theoretical research on effective teaching behaviors (Danielson, 2007; Danielson & McGreal, 2000; Hull, 2013; Kimball & Milanowski, 2009; McGreal, 1995; Scriven, 1988). Assessing the performance of teaching practices directly linked to standards may improve instruction and student learning (Steinberg & Donaldson, 2014). Several organizations and researchers have issued standards for teacher-evaluation instruments (Steinberg & Donaldson, 2014).

In 1988, the JCSEE issued standards for personnel evaluations. As previously referenced, the JCSEE organized 21 standards into four categories: propriety, utility, feasibility, and accuracy (JCSEE, 1988). The propriety standards address expectations regarding the legal and ethical issues involved in evaluation. The utility standards address expectations regarding the transparency and usefulness of evaluations for informing personnel and professional-development decisions. The feasibility standards address expectations regarding the practical, political, and fiscal viability of evaluations. The accuracy standards address expectations for the validity and reliability of the information evaluations provide regarding teacher performance (JCSEE, 2016; Stufflebeam & Shinkfield, 2007). In 2008, the JCSEE updated personnel evaluation standards. The updated utility standards include constructive orientation, defined uses, evaluator qualifications, explicit criteria, functional reporting, and professional development (Gullickson, 2009).

The ETS developed standards for teaching through theoretical and field research to use as the basis for the Praxis Series, licensing tests and classroom performance assessments for new teachers (Dwyer, 1998). ETS organizes the standards into four groups: organizing content knowledge for student learning, creating an environment for student learning, teaching for student learning, and teacher professionalism (Dwyer,

1998). The ETS standards significantly influenced the development of the FFT.

The National Board for Professional Teaching Standards (NBPTS) established standards teachers must meet to earn national certification (NBPTS, 2016). Participation in the national certification process is voluntarily. The NBPTS applies one of 25 different sets of standards to candidates for certification, depending on the age of students and the subjects candidates teach. All standards align with the NBPTS's Five Core Propositions:

1. Teachers are committed to students and their learning…
2. Teachers know the subjects they teach and how to teach those subjects to students…
3. Teachers are responsible for managing and monitoring student learning…
4. Teachers think systematically about their practice and learn from experience…
5. Teachers are members of learning communities (NBPTS, 2016, para 3–7).

The FFT is grounded on standards developed by InTasc, which provides 10 standards divided into four categories: the learner and learning, content knowledge, instructional practice, and professional responsibility (Council of Chief State School Officers, 2011). Danielson (2007) asserted that the use of standards provides credibility to the domains and components assessed with FFT.

Critics of using comprehensive standards as the basis for teacher evaluation asserted that observation rubrics should not include every successful teaching practice, especially those that occur outside the classroom and argued that evaluating too wide a scope of teaching practices may diminish the correlation between evaluation ratings and student achievement (The New Teacher Project, 2013). Kane and Cantrell asserted "Teachers shouldn't be asked to expend effort to improve something that doesn't help them achieve better outcomes for their students" (2013, p. 15). Measuring too many

teaching practices wastes the effort of evaluators and teachers and does not add "precision to ratings or feedback" (The New Teacher Project, 2013, p. 6).

The History of Teacher Evaluations

Until the middle of the 19th century, teaching was not a professional discipline, and teacher evaluations rested on local criteria, not pedagogical expertise (Marzano, Frontier, & Livingston, 2011). School systems grew larger and more complex during the latter half of the 19th century (Marzano et al., 2011) and the establishment of normal schools and teacher institutes met the growing demand for teachers with subject knowledge and pedagogical skills (Spring, 2011).

From the late 19th century until World War II, the status of teaching as a profession improved. The number of states that required teachers to have licenses or certifications increased from four in 1898 to 42 in 1933 (Spring, 2012). Two views on education influenced expectations for teachers. Dewey argued that schools should reflect democratic ideals, and teachers should provide student-centered and individualized instruction to prepare students for citizenship in a democracy (as cited in Marzano et al., 2011; Spring, 2011). Thorndike and Cubberley argued that schools should apply scientific management principals, and teachers should provide efficient and structured instruction to prepare students for the 20th century (as cited in Marzano et al., 2011; Spring, 2011). In the 1920s, Wetzel recommended the use of aptitude tests, measurable objectives for every course, reliable assessments to measure learning, and the use of assessment data to evaluate schools and teachers (as cited in Marzano et al., 2011).

After World War II, educators shifted their focus from the scientific management of schools to supervisory practices such as classroom observations and postobservation conferences to improve teacher performance (Marzano et al., 2011). Before 1950,

teachers were rarely evaluated (Shinkfield & Stufflebeam, 1995). In the 1960s and 1970s, research on teacher effects, learning theory, and relationships between teaching practices and the acquisition of basic skills led to the development of more precise observation instruments and clinical-supervision practices (Brophy & Good,1986; Danielson & McGreal, 2000). Hunter designed a set of prescriptive teaching practices based on teacher effects and learning theory research, and state and local education agencies developed evaluation tools that encouraged highly structured and teacher-centered classrooms based on Hunter's work (as cited in Danielson & McGreal, 2000).

In the 1980s, cognitive-learning theory emphasized the social nature of learning and learning context; higher order thinking skills led to an expanded view of good teaching (Marzano et al., 2011). Research on teacher effects and cognitive-learning theory suggested that teacher-evaluation systems should measure a variety of proven teaching practices, engagement in professional discussion and reflection, and provision of targeted professional development (Danielson & McGreal, 2000). Concerns about the quality of public education in the 1980s and 1990s led to increased emphasis on teacher evaluations (Shinkfield & Stufflebeam, 1995). In the 1990s, Danielson developed a complex description of teaching, and an FFT evaluation model that was the most "comprehensive approach to evaluation at that time" (Marzano et al., 2011). In the 21st century, the focus of teacher evaluation shifted from teacher performance to student achievement, and evaluators aim to measure professional practices and student achievement (Marzano et al., 2011).

Federal Influence on Evaluations

The federal government influenced many recent changes to teacher evaluations. In 2011, the federal government incentivized the use of new, more rigorous teacher

evaluations that include student-achievement data as a new strategy to improve public education (USDE, 2016). Historically, many federal government goals and strategies for improving public education have been pursued through the ESEA (1965; Spring, 2012). The ESEA provides categorical aid to states that agree to comply with its provisions and associated regulations. The initial purpose of the ESEA was to improve the quality of public schools that serve high populations of low-income students (ESEA, 1965; Spring, 2012).

In response to national concerns about flagging student achievement and global economic competition, the Improving America's Schools Act of 1994 and the No Child Left Behind Act of 2001 (NCLB) reauthorizations extended the scope of the ESEA to address the quality of public education for all students (Improving America's Schools, 1994; NCLB, 2001; Spring, 2011, 2012). Improving America's Schools required states to annually assess students in mathematics and reading. NCLB held states, school districts, and individual schools accountable for closing achievement gaps among subcategories of students and improving the performance of all students based on adequate yearly progress goals. Adequate yearly progress goals increased each year on a trajectory set to achieve 100% proficiency on the required assessments by 2014 (NCLB, 2001). Although NCLB required school districts to provide students with highly qualified teachers, it did not require teachers to be held accountable for student performance on the required assessments.

In 2011, the USDE offered states a flexibility waiver that conditionally exempted recipients from 10 provisions of the ESEA, including the provision that required states to achieve 100% proficiency on state assessments (USDE, 2016). To receive an ESEA flexibility waiver, states agreed to implement four principals: a curriculum and

assessment system based on college- and career-ready standards, a system of accountability and support for schools, a process for eliminating redundant reporting, and a rigorous system for evaluating teachers and principals. The ESEA waiver specified that teacher-evaluation systems must use student growth and other measures of professional practice to differentiate levels of teacher effectiveness and that the evaluations were to be used to inform professional development and personnel decisions, thereby improving the educational outcomes of students (USDE, 2016).

By 2015, 46 states had received ESEA flexibility waivers (USDE, 2016, p. 1). Several states, including Maryland, chose to use the FFT observation instrument to evaluate teaching practices (Danielson Group, 2011). Most of Maryland's school districts were already in the process of implementing the FFT as a condition of the RTTT grants they had received under the ARRA (2009; Ballou & Springer, 2015; MSDE, 2013, 2014a). The Every Student Succeeds Act of 2015, a reauthorization of the ESEA (1965), passed into law on December 10, 2015 (ESSA, 2015). The Every Student Succeeds Act of 2015 does not provide criteria for teacher evaluations (ESSA, 2015). On December 18, 2015, the USDE issued an open letter stating the ESEA flexibility waivers would remain in effect through August 1, 2016 (USDE, 2015). When the ESEA waivers expire, no federal policy will compel states to include student-growth measures in teacher evaluations.

The General Performance of Teacher Evaluations

In contrast to the research that shows significant differences in teacher effects on student learning, teacher evaluations have historically categorized 98% of teachers as satisfactory, and teacher evaluation ratings have rarely been used to inform districtwide, schoolwide, or individualized professional development or to make personnel decisions

(Kane, 2012; Weisberg et al., 2009).

Most teacher evaluation systems require two or fewer observations and do not significantly differentiate levels of effectiveness (Weisberg et al., 2009). Districts that use evaluations with two performance levels rate 99% of teachers satisfactory, and districts that use evaluations with four or more performance levels rate 70% of teachers at the highest performance level and 24% of teachers at the second highest level (Weisberg et al., 2009). Student achievement does not affect evaluation ratings: "On average, over the last three years, only 10% of failing schools [schools that did not meet adequate yearly progress goals] issued at least one unsatisfactory rating to a tenured teacher" (Weisberg et al., 2009, p. 12). Additionally, the use of evaluations to improve teacher performance is limited. A survey of 15,000 teachers found only 26% of experienced teachers and 43% of teachers with less than 5 years experience received recommendations for improving teaching practices during the evaluation process (Weisberg et al., 2009)

Strunk, Weinstein, and Makkonen (2014) analyzed the correlation between student academic-growth data and the observation-based evaluation ratings of approximately 200 teachers in the Los Angeles Unified School District (Strunk et al., 2014). The Los Angeles Unified School District used the Teaching and Learning Framework (TLF) evaluation instrument to evaluate teachers and the Academic Growth Over Time standardized assessment to measure the academic growth of students (Strunk et al., 2014). The purpose of the study was to determine if observation ratings and student-assessment data provide teachers and evaluators conflicting evidence of teacher effectiveness and to determine if a subset of teaching practices measured by the TLF are more predictive of student achievement than the overall summative ratings (Strunk et al., 2014).

The study found weak to moderate consistency between teachers' overall ratings on the TLF and their students' performance on the Academic Growth Over Time assessments and found that the ratings for a small subset of teaching practices were more predictive of student achievement than the TLF's overall ratings for English-language arts teachers, but not for mathematics teachers (Strunk et al., 2014). The subset included the following teaching practices: Expectations for Learning, Achievement and Student Ownership of Their Work, Using Questioning and Discussion Techniques, Structures to Engage Students in Learning, and Using Assessment in Instruction to Advance Student Learning. One teaching practice, Analysis and Use of Assessment Data for Planning, closely aligned with student achievement in mathematics and English-language-arts teachers. Strunk et al. (2014) recommended school districts identify the teaching practices in their observation instruments that most closely align with student achievement and focus their professional development on those teaching practices.

Using Multiple Measures of Effectiveness

In an effort to strengthen the relationship between teacher-evaluation ratings and student learning, some states include student-achievement data from standardized assessments in their teacher-evaluation systems (Ballou & Springer, 2015; Eckert & Dabrowski, 2010; Grossman, Loeb, Cohen, & Wyckoff, 2013; Kane, McCaffrey, Miller, & Staiger, 2013; Steinberg & Donaldson, 2014). Additionally, state and local merit pay programs, NCLB, RTTT, and the ESEA flexibility waiver have all compelled educators to develop various processes to measure student learning and attribute student learning to school districts, schools, and teachers (Dillon, 2008; Grossman et al., 2013; Mueller, 2011; Steinberg & Donaldson, 2014). Deciding which measures of student growth are appropriate for use in teacher evaluations has been heavily debated (Ehlert et al., 2013).

Variables other than teacher quality affect student performance, and practices such as teachers monitoring their own students during testing and verifying their own student rosters may affect the validity of the testing data used in teacher evaluations (Ballou & Springer, 2015).

NCLB established the use of status models and growth-to-standard models to measure student achievement to hold states, school districts, and schools accountable for improving student outcomes. Status models and growth-to-standard models measure student learning by the percentage of students who achieve a specified performance level by a specified point in time (Betebenner, 2008). Critics of status and growth-to-standard models argue that schools serve different student populations and should be held accountable for student growth rather than an arbitrary level of achievement (Ladd & Lauen, 2010). Student characteristics such as socioeconomic background affect status and growth in standard models more than in growth models (Ryser & Rambo-Hernandez, 2014). In 2005, the USDE responded to concerns about status models and initiated the Growth Model Pilot Program, eventually allowing seven states to develop plans to use student growth measures rather than status models to meet NCLB's accountability requirements (Betebenner, 2008; Ho, Lewis, & MacGregor Farris 2009). Evaluators have applied the same arguments for using growth models rather than status models for school accountability to teacher evaluations, and states have used a variety of student growth models to incorporate student test data into teacher evaluations (Ehlert et al., 2013; Steinberg & Donaldson, 2014).

Some states use value-added models to calculate student growth and evaluate teachers (Steinberg & Donaldson, 2014). The "Tennessee Value-Added Assessment System (TVAAS) is one of the most sophisticated and respected value-added models in

use" (Eckert & Dabrowski, 2010, p. 89). The TVAAS was initially designed to provide student growth data for a merit-pay system, but Tennessee's Education Improvement Act of 1992 specified the TVAAS would be used to calculate student growth for the state's accountability system, including teacher evaluations (Sanders & Horn, 1998).

TVAAS uses a statistical analysis of longitudinal student-achievement data from state assessments to project or estimate a trajectory for each student's future academic growth (Eckert & Dabrowski, 2010). Evaluators assess school districts, schools, and teachers based on the percentage of students who exceed, maintain, or fall short of their predicted academic growth. The TVAAS does not control for categorical differences among students because each student's growth is compared to that student's own past growth (Sanders & Horn, 1998). The SAS Institute's Education Value-Added Assessment System is built on the TVAAS model. North Carolina, Ohio, Pennsylvania, and Tennessee and several school districts in 16 other states use the SAS system statewide (Ballou & Springer, 2015; Collins, 2014; Eckert & Dabrowski, 2010).

Critics of the TVAAS and similar value-added systems point to quantitative studies that suggested the populations teachers serve affect teacher ratings and teacher ratings provided by such systems are unstable, often changing dramatically from year to year (Armour-Garb, 2009; Newton, Darling-Hammond, Haertel, & Thomas, 2010; Papay, 2011). Regarding the latter concern, it is "misleading" and "erroneous" (Good & Lavigne, 2014, p. 6) to assume teacher ratings should be stable from year to year.

Some states use student-growth-percentiles (SGP) models to incorporate student-testing data into teacher evaluations (Ballou & Springer, 2015; Betebenner, 2008; Ryser & Rambo-Hernandez, 2014; Steinberg & Donaldson, 2014). Evaluators traditionally have used normative growth models or SGPs to describe a student's achievement on an

assessment relative to all other students taking the same assessment. For accountability purposes, SGP models can control for differences among students by describing a student's achievement relative only to those students with historically similar scores (Ryser & Rambo-Hernandez, 2014). Evaluators can assess school districts, schools, and teachers on the percentage of students who achieve below, near, or above the mean level of growth for students with historically similar scores on the same standardized tests. Critics of normative growth models point out that nearly 50% of students will always fall below the median and nearly 50% will always fall above the median. As a result, SGPs can obscure changes in the rate of growth, in that student growth percentiles above the median do not necessarily indicate a student is making adequate progress toward an acceptable level of performance by a specific point in time (Betebenner, 2008; Jensen, 2013).

Some states and school districts use SLOs to incorporate student growth into teacher evaluations (Lachlan-Haché, Cushing, & Bivona, 2012; Steinberg & Donaldson, 2014). Denver Public Schools started using SLOs in 1999 to evaluate teacher performance for a merit-pay program (USDE, n.d.). SLOs are learning goals set collaboratively by the teacher being evaluated and the administrator conducting the evaluation. SLOs identify the subject content to be taught, the time period the instruction is to occur, the population of students to be assessed, the assessment that will be used to measure student learning, and the goal for student growth or achievement (MSDE, 2014a). Critics asserted that the evaluation ratings provided by SLOs are subjective because evaluators apply no psychometric methodology to ensure SLOs are valid or reliable (Lachlan-Haché et al., 2012). In an effort to provide more objectivity to SLOs, some school districts are using MAP growth projections to set SLO performance targets

for teachers (Jensen, 2013).

Using multiple measures of effectiveness adds to the strength of the correlation between summative evaluation ratings and student learning (Kane & Cantrell, 2013). Different models for weighting ratings on classroom observations, student surveys, and the prior history of the performance of each teacher's students on state assessments provide different levels of accuracy in predicting student performance on state assessments and tests of higher order thinking skills (Kane & Cantrell, 2013). The most accurate prediction of student performance on state assessments for middle school English-language arts teachers weighted state testing results at 81%, student surveys at 17%, and teacher observations at 2%; however, that weighting model did not provide consistent ratings for the same teacher from year to year (Kane & Cantrell, 2013). Weighting state testing results at 50%, student surveys at 25%, and observations at 25% was slightly less accurate at predicting performance on state assessments, but was more accurate at predicting performance on assessments of higher order thinking skills and provided ratings that are consistent for the same teacher year to year. The combinations of weighting ratings on classroom observations, student surveys, and a prior history of each teacher's student performance on state assessment yielded similar results for teachers of most subjects and grade levels (Kane & Cantrell, 2013).

Although the inclusion of measures of student growth in teacher evaluations may strengthen the relationship between evaluation ratings and student learning, evaluators rarely use test data to inform personnel and professional-development decisions (Goldring et al., 2015). Surveys and interviews of principals and central office administrators in six urban school districts indicated that school districts relied more heavily on teacher observations than student growth data to make personnel decisions in

all subject areas, even though observation ratings accounted for only 25% to 50% of teachers' summative-evaluation ratings. The researchers suggested that states and school districts consider shifting some focus and resources from improving student-growth measures to developing tools for high-quality observation systems (Goldring et al., 2015).

Recent Performance of Teacher Evaluations

The states that received ESEA waivers implemented the required college- and career-ready curricula, accountability systems, and teacher-evaluation systems (Steinberg, & Donaldson, 2014; USDE, 2013, 2016). Researchers collected the evaluation data reported by 19 states that implemented new evaluation systems and found that only 3% of teachers were rated less than effective (Kraft & Gilmour, 2016). The Georgia Department of Education (GDE, 2014) reported that 96.9% of the 40,000 teachers who were evaluated during the 2012–2013 school year received proficient or exemplary ratings on the Teacher Assessment on Performance Standards (TAPS) Georgia used to measure teaching practices. Less than 0.1% of teachers were rated ineffective, 3.0% were rated needs development, 93.5% were rated proficient, and 3.4% were rated exemplary. Regarding individual standards, 95% of teachers were rated proficient or exemplary on eight of the 10 TAPS standards, 92.2% of teachers were rated proficient or exemplary on Standard 4: Differentiation, and 94.3% of teachers were rated proficient or exemplary on Standard 8: Academically Challenging Environment. Georgia's teacher-evaluation system uses student-growth percentiles on state assessments and SLOs to calculate student growth scores for teachers. The GDE reported that student-growth scores were significantly lower than TAPS scores; however, the GDE reported a weak, positive correlation emerged between TAPS scores and student-growth scores (GDE, 2014).

The Michigan Department of Education (2014) reported that about 23% of teachers were rated highly effective, 75% were rated effective, 2% were rated minimally effective, and less than 1% were rated ineffective. Each Michigan school district developed its own evaluation system and approximately 50% of the 897 school districts chose to use the FFT to evaluate teaching practices (Michigan Department of Education, 2014). The Tennessee Department of Education reported that 31.5% of the state's evaluated teachers were rated at the highest of the five levels of performance on teacher observations, 43.2% were rated at the second-highest level, and 22.3% at the third-highest level. About 3% of teachers were rated at the two lowest levels of performance (Tennessee Department of Education, 2015). Tennessee's evaluation system differentiated performance levels to a greater extent than Georgia's or Michigan's; however, their teacher-observation scores were still skewed toward higher performance ratings (Tennessee Department of Education, 2015).

In a report to the Maryland State Board of Education, the MSDE (2014b) stated evaluators assessed 43,805 teachers during the 2012–2013 school year, and that 40.8% of teachers were rated highly effective, 56.4% were rated effective, and 2.8% were rated ineffective. Although state testing data were not included in the teachers' actual evaluations, the MSDE report showed that if the testing data had been included in the evaluations, 86.6% of evaluation ratings would have stayed the same, 3.2% would have increased by one level, and 10.2% would have decreased by one level, primarily from highly effective to effective. The report noted that schools in the highest quartile of poverty had higher numbers of ineffective teachers and lower numbers of highly effective teachers than schools in the lowest quartile of poverty (MSDE, 2014b). The new teacher evaluation systems in Georgia, Michigan, Tennessee, and Maryland did not differentiate

levels of teacher performance to a much greater extent than previous evaluation systems (Kraft & Gilmour, 2016; Weisberg et al., 2009).

Summary

Chapter 2 reviewed the work of authors and researchers regarding the conceptual framework of this study, topics apposite to teacher evaluation and measures of student learning, including research on the correlation between the ratings teacher received on the FFT and student learning. Researchers of the FFT showed a correlation between teacher ratings on a subset of components and a variety of value-added achievement data with correlation coefficients that ranged from .18 to .50 (Borman & Kimball, 2005; Holtzapple, 2003; Kane & Staiger, 2012; Milanowski et al., 2004; Sartain et al., 2011). Absent from the literature was an examination of the correlation between all 22 FFT components and a nationally normed value-added measure of student learning. Chapter 3 explains the research design and methodology of this study that examined all 22 FFT and used the nationally normed MAP reading assessment to measure student learning.

The purpose of this study was to examine the correlation between teachers' evaluation ratings and their students' attainment of projected growth scores in reading. The independent variables were the FFT component ratings assigned to teachers of third-, fourth-, and fifth-grade students in a western Maryland school district during the 2013–2014 and 2014–2015 school years. The dependent variable was the percentage of each of the evaluated teachers' students attaining projected growth scores in reading, measured by the MAP reading assessment. A multiple correlation analysis determined the proportion of the variance in the percentage of students attaining projected growth scores in reading, explained by teachers' performance ratings on all 22 FFT components and by teachers' performance ratings on each individual component.

Chapter 3 – Research Design and Methodology

Introduction

Teacher evaluations have historically lacked utility in that ratings provided by evaluations did not differentiate levels of effectiveness based on student learning, and educators rarely used evaluations to inform professional development or personnel decisions (Kane, 2012; Weisberg et al., 2009). In response to federally led reform efforts, many states recently implemented new standards-based evaluation models (Steinberg & Donaldson, 2014). The new evaluations have not significantly improved the utility of teacher evaluations (Kraft & Gilmour, 2016). Hundreds of school districts implemented Danielson's (2007) FFT. The purpose of this study was to examine the correlation between teachers' evaluation ratings and their students' attainment of projected growth scores in reading.

Research Design

This study applied a multiple correlation analysis to FFT performance ratings assigned to teachers during evaluations and the percentage of each evaluated teacher's students attaining projected growth scores in reading. The independent variables were the FFT component ratings assigned during the 2013–2014 and 2014–2015 school years to the teachers of third-, fourth-, and fifth-grade students in a western Maryland School district. The dependent variable was the percentage of each of the evaluated teachers' students who attained projected growth scores in reading, measured by the NWEA's MAP reading assessment administered at the beginning and end of each school year. A multiple correlation analysis determined the proportion of the variance in student attainment of projected growth scores in reading explained by teachers' performance ratings on all 22 FFT components and by teachers' performance ratings on each

individual component. The multiple correlation analysis was conducted using the IBM Statistical Package for the Social Sciences (SPSS). The following research questions and hypotheses guided this study:

Research Questions

1. Do teachers' performance ratings on all 22 FFT components explain a significant proportion of the variance in the percentage of teachers' students attaining projected growth scores in reading?
2. Do teachers' performance ratings on any individual FFT component explain a significant proportion of the variance in the percentage of teachers' students attaining projected growth scores in reading?

Null Hypotheses

1. H_0: *The* FFT component ratings do not explain a significant proportion of the variance in the percentage of teachers' student attaining projected growth scores in reading.
2. H_0: No FFT component rating explains a significant proportion of variance in the percentage of teachers' student attaining projected growth scores in reading.

Research Methodology

Multiple correlation analysis is a common statistical procedure used to determine the extent to which multiple independent variables explain variance in one dependent variable (McMillan & Schumacher, 2010; Rovai, Baker, & Ponton, 2014). Multiple correlation analysis has been used in research to determine the extent to which MAP reading and mathematics scores predict summative teacher-evaluation ratings (Washington, 2011) and to determine the extent to which a teacher's summative-evaluation rating and other teacher characteristics predict student learning in reading

(Gilbert, 2014).

In this quantitative study, a multiple correlation analysis determined the proportion of the variance in the percentage of teachers' students attaining projected growth scores in reading, explained by teachers' performance ratings on FFT components. The dependent variable was a measure of student learning based on the percentage of a teachers' students who met their growth projections in reading, expressed as a percentage. Using projected growth scores rather than achievement scores controlled for variations in students' reading ability at the beginning of the instructional year. The independent variables were the average of two to four FFT ratings (*Distinguished* = 4, *Proficient* = 3, *Basic* = 2, and *Unsatisfactory* = 1) for each component assigned to each teacher during evaluative classroom observations. Averages were rounded to the nearest whole number.

Population

I selected all teachers in a western Maryland district's 25 elementary schools who were the primary reading instructors of third-, fourth-, and fifth-grade students and were evaluated during either or both the 2013–2014 and 2014–2015 school years for the study. I did not include teachers of kindergarten, first, and second grade in the study because students do not read the MAP assessments until the third grade. I did not include secondary school teachers in the study due to the probability that several teachers influence secondary students' growth in reading. The research included teachers from only one school district to limit the effect of variations in curriculum and instructional resources that are based on local decisions and preferences. I evaluated all teachers included in the study with the FFT.

Sample Selection

This study analyzed the FFT component ratings of 200 teachers evaluated during the 2013–2014 and 2014–2015 school years in a western Maryland school district. The teachers all taught third-, fourth-, or fifth-grade students who took the beginning of the year and end of the year MAP reading assessments. The study included teachers from 25 elementary schools. School and district-level administrators evaluated the teachers.

Data Source

The independent variables in the study were each teacher's average rating for each of the 22 FFT components. Administrators formally observed teachers a minimum of two times during an evaluation year and assigned component ratings during each observation. Principals, assistant principals, or district-level supervisors assigned the ratings using the FFT observation instrument. Evaluators rated teachers as unsatisfactory, basic, proficient, or distinguished on each component. I did not include teachers with any missing ratings in the study.

The dependent variable in this study was the percentage of each of the evaluated teachers' students attaining projected growth scores in reading, measured by the MAP reading assessment. The projected growth scores were based on the median level of growth among students in the same grade with the same score at the beginning of the instructional year in the NWEA's norming sample (NWEA, 2012). After students take the MAP reading assessment at the end of the instructional year, the school district receives a report from NWEA that indicates which students attained growth projections (NWEA, 2012). The NWEA provided a report to the school district with the data used to calculate the percentage of each teacher's students attaining individual growth projections in reading.

Data-Collection Procedures

After obtaining approval from the school district, I collected the data for the study. First, I collected the independent variable data. An administrator in the school district's Teacher Evaluation and Professional Development Office (TEPD) copied the names, summative-evaluation ratings, and component ratings of the district's teachers of third-, fourth-, and fifth-grade students from the school district's data-management system onto an Excel spreadsheet and copied the spreadsheet to a USB storage device. The TEPD administrator gave the USB storage device to the district's director of elementary schools, who removed the names of teachers who were not the students' primary reading instructor and returned the USB storage device to the TEPD administrator.

The TEPD administrator gave the USB storage device with the independent-variable data to an administrator in the Testing and Accountability Office. The Testing and Accountability Office administrator provided a list of each teacher's students with a *did not attain projected growth score* or *attained projected growth score* code of 0 and 1, respectively, to the Excel spreadsheet and returned the USB storage device to the TEPD administrator.

The NWEA provides the Testing and Accountability Office administrator the data to determine which students attained projected growth scores. The projected growth scores are based on the median level of growth among students in the same grade with the same score at the beginning of the instructional year in the NWEA's norming sample (NWEA, 2012). The TEPD administrator calculated the percentage of each teacher's students attaining individual growth projections in reading by dividing the number of each teacher's students attaining individual-growth projections by the total number of

each teacher's students who took the beginning of the year and end of the year MAP reading assessments.

For demonstration purposes only, Table 2 provides a simulated example of the calculation of the percentage of students attaining projected growth scores. The example shows a teacher with four students of seven attaining projected growth scores in reading, resulting in a dependent variable value of 57.14%.

Table 2

Calculating the Percentage of Students Attaining Growth Projections

Students	Fall MAP reading score	Spring MAP reading score	Growth score	Growth projections
Student 0001	248	322	74	1
Student 0002	377	401	24	0
Student 0003	531	565	34	0
Student 0004	389	470	81	1
Student 0005	433	543	110	1
Student 0006	455	509	54	0
Student 0007	502	592	90	1

Note. The scores in the sample are not based on real data. Percentage of Students Attaining Projected Growth Scores: 57.14%.

The TEPD administrator entered the percentage of each teacher's students attaining growth projections onto the spreadsheet as percentages to the hundredths of a percent. The TEPD then calculated the average component ratings for all 22 components for each teacher and entered the average rating onto the spreadsheet. The TEPD administrator then deleted students' names and individual component ratings. Finally, the TEPD administrator converted the names of teachers to three digit numbers, then gave the USB storage device with teacher codes, teachers' average component ratings, and the percentage of each teacher's students attaining projected growth scores to me. I then made random changes to the teacher codes to prevent reassociation of the teacher codes

to teacher names.

Data-Analysis Procedures

I conducted all statistical tests using SPSS. I entered each teacher code into SPSS as a case. Each case had a corresponding dependent variable entered into SPSS as a scale measure, indicating the percentage of each teacher's students attaining projected growth scores and 22 independent variables entered into SPSS as a categorical measure with either a 0 to indicate a basic or proficient performance rating or a 1 to indicate a distinguished performance rating. After entering the data, I conducted a regression analysis.

The first step in interpreting the multiple correlation analysis was to confirm the variable data met the following six assumptions:

1. Independence of observations must exist, meaning the errors (residuals) of adjacent observations in the multiple regression are not correlated. I tested the independence of observations using the Durbin–Watson statistic. A Durbin–Watson test provides a value between 0 and 4, and a value close to 2 indicates no significant correlation among residuals.
2. A linear relationship must exist between the dependent variable and the collective independent variables. I tested the assumption of linearity between the dependent variable and the independent variables collectively by visual inspection of a scatter plot of studentized residuals and unstandardized predicted values (see Appendix B) generated during the multiple correlation using the SPSS. A horizontal band of the residual data points characterized a linear relationship (Laerd Statistics, 2015). The individual independent variables were all categorical variables and therefore I did not test for linearity (Laerd Statistics, 2015; Rovai et

al., 2014).

3. Homoscedasticity of residuals must exist, meaning the residuals of each level of the independent variables have similar variances (Rovai et al., 2014). I tested the assumption of homoscedasticity by a visual inspection of the plot of studentized residuals and unstandardized predicted values (see Appendix B). Homoscedasticity is characterized by an approximately even spread of points on the scatter plot with no pattern of increasing or decreasing data points across the predicted values (Laerd Statistics, 2015).
4. Multicollinearity between or among variables must exist, meaning there must not be high correlations between or among two or more independent variables. I tested multicollinearity with a tolerance value generated with SPSS. Tolerance values greater than 0.1 indicate multicollinearity in the data set is not probable (Laerd Statistics, 2015; Rovai et al., 2014).
5. No significant outliers must exist, meaning dependent or independent variable data points that deviate far enough from the norm to distort the strength of the relationship between the independent and dependent variables (Laerd Statistics, 2015; Rovai et al., 2014). I conducted a casewise diagnostic with SPSS tests for outliers greater than ± 3standard deviations among the dependent or independent variables. The multiple correlation analysis also provides leverage values to test for untenably high leverage values. Leverage values above .5 are considered high. I tested the Cook's distance value to test for untenably influential points among variables. Values above 1.0 are considered untenable (Laerd Statistics, 2015; Rovai et al., 2014).
6. Normal distribution of residuals must exist. I tested the distribution of residuals

with a visual inspection of a histogram (see Appendix C) with a normal curve superimposed (Laerd Statistics, 2015).

After confirming that the variable data met the required assumptions, I interpreted the results of the multiple correlation analysis. To address the first research question, the multiple correlation analysis procedure calculated the coefficient of multiple determination (R^2) to determine the proportion of the variance in the percentage of teachers' students attaining projected reading scores. The coefficient of multiple determination (R^2) is the percentage of variance explained by the regression model, a model based on teachers' FFT component ratings, over and above a model based on the mean percentage of students attaining projected growth scores in reading (Laerd Statistics, 2015). The multiple correlation analysis procedure also calculated the adjusted coefficient of multiple determination (adj. R^2). The adjusted coefficient of multiple determination (adj. R^2) is the percentage of variance that is generalizable to a larger population. The multiple correlation analysis procedure conducted with the SPSS provided an analysis of variance (ANOVA). The ANOVA value indicates whether the model using the independent variable data is significantly ($p < .05$) better at predicting the percentage of students attaining growth projections than the mean model.

To address the second research question, the multiple correlation procedure provided unstandardized coefficients for each FFT component. The unstandardized coefficient is the change in the mean percentage of students attaining projected growth scores when teachers receive a distinguished rating compared to teachers receiving a basic or proficient rating on each component when the ratings on all other components remain constant. The multiple correlation procedure also provided the statistical significance ($p < .05$) for each unstandardized coefficient.

Validity and Reliability

The use of extant data minimized internal and external threats to validity as I did not interact with participants (evaluators, teachers, or students) represented in the analyzed data. The school district collected independent variable data from actual teacher evaluations conducted by principals, assistant principals, and supervisors. The school district had already collected dependent-variable data from students' MAP assessment data. The school district collected the data used for the independent and dependent variables from the 2013–2014 school year 2 years before I conducted this research, and collected the 2014–2015 data approximately 1 year before I conducted this research. The study's population was limited to teachers of reading for third-, fourth-, and fifth-grade students because teachers read the MAP reading assessment to students in lower grades and secondary students have several teachers.

Role of the Researcher

I am a doctoral student in the Educational Leadership program at Frostburg State University. Additionally, I have been employed for 25 years by the school district from which data were collected. I did not evaluate or supervise and was not evaluated or supervised by any of the teachers or evaluators represented in the data. I did not evaluate or supervise and was not evaluated or supervised by the administrators in the Teacher Evaluation and Professional Development or Testing and Accountability Offices from which the data were obtained.

Measures of Ethical Protection

No threat could accrue to the human participants in this research. I worked exclusively with extant data and had no interaction with evaluators, teachers, or students represented in the data. An administrator in the Teacher Evaluation and Professional

Development Office converted the names of teachers to numerical codes and paired the teacher code with each teacher's average component ratings, and the percentage of that teacher's students attaining individual reading growth projections. I changed the order and numerical teacher codes before conducting the analysis of the data. Administrators in the Offices of Testing and Accountability and Teacher Evaluation and Professional Development removed all personally identifiable information. I have stored the USB storage device used to maintain the data in a secured lockbox in a secured drawer in a secured building and will maintain it for 5 years and then wipe it. I will make no copies of the USB storage device. I followed all school district and Institutional Review Board processes.

Summary

The purpose of this study was to examine correlation between teachers' evaluation ratings and their students' attainment of projected growth scores in reading. The independent variables were the FFT component ratings assigned to teachers of third-, fourth-, and fifth-grade students in a western Maryland school district during the 2013–2014 and 2014–2015 school years. The dependent variable was the percentage of each of the evaluated teachers' students attaining projected growth scores in reading, measured by the MAP reading assessment. A multiple correlation analysis determined the proportion of the variance in student attainment of projected growth scores in reading, explained by teachers' performance ratings on all 22 FFT components and by teachers' performance ratings on each individual component. The collection and analysis of data followed appropriate procedures for the ethical protection of participants represented in the data and the validity and reliability of the analysis. Chapter 4 presents the findings of this study.

Chapter 4 – Results

Introduction

The purpose of this study was to examine the correlation between teachers' evaluation ratings and their students' attainment of projected growth scores in reading. The independent variables were the FFT component ratings assigned to teachers of third-, fourth-, and fifth-grade students in a western Maryland school district during the 2013–2014 and 2014–2015 school years. The dependent variable was the percentage of each of the evaluated teachers' students attaining projected growth scores in reading, measured by the MAP reading assessment.

A multiple correlation analysis determined the proportion of the variance in student attainment of projected growth scores in reading explained by teachers' performance ratings on all 22 FFT components and by teachers' performance ratings on each individual component. I confirmed the assumptions regarding the data and analyzed the results of the multiple correlation using SPSS. Chapter 4 presents the descriptive statistics and the findings from the multiple correlation analysis.

Research Questions

1. Do teachers' performance ratings on all 22 FFT components explain a significant proportion of the variance in the percentage of teachers' students attaining projected growth scores in reading?
2. Do teachers' performance ratings on any individual FFT component explain a significant proportion of the variance in the percentage of teachers' students attaining projected growth scores in reading?

Null Hypotheses

1. H_0: FFT component ratings do not explain a significant proportion of the variance

in the percentage of teachers' students attaining projected growth scores in reading.

2. H_0: No FFT component rating explains a significant proportion of variance in the percentage of teachers' students attaining projected growth scores in reading.

Descriptive Statistics

I collected but did not use the summative evaluation ratings in the multiple correlation analysis. The distribution of the summative ratings of teachers in the study was ineffective: 0, effective: 61, and highly effective: 139. The independent variable data used in the multiple correlation analysis included 4,400 component ratings assigned to the 200 teachers in the study for the performance of the 22 FFT components.

Table 3 shows the distribution of component ratings for Domains 1 and 2 of the FFT, and Table 4 shows the distribution of component ratings for Domains 3 and 4 of the FFT. Teacher performance was never rated unsatisfactory, was rated basic 65 times, was rated proficient 2,603 times, and was rated distinguished 1,732 times. The highest number of basic ratings (8) and the lowest number of distinguished ratings (24) assigned to teachers on any component was for component 1f: *Designing Student Assessments*. More than half of the teachers received ratings of distinguished on components 2a: *Creating an Environment of Respect and Rapport*, 2d: *Managing Student Behavior*, 3a: *Communicating with Students*, and 4a: *Reflecting on Teaching*, and 4f: *Showing Professionalism*.

Table 3

Distribution of Component Ratings for Domains 1 and 2

Components	1a	1b	1c	1d	1e	1f	2a	2b	2c	2d	2e
Unsatisfactory	0	0	0	0	0	0	0	0	0	0	0
Basic	1	1	3	3	3	8	2	5	2	4	1
Proficient	137	109	99	153	103	168	57	104	117	79	151
Distinguished	62	90	98	44	94	24	141	91	81	117	48
N	200	200	200	200	200	200	200	200	200	200	200

Note. Independent variables: 1a = demonstrating knowledge of content and pedagogy; 1b = demonstrating knowledge of students; 1c = setting instructional outcomes; 1d = demonstrating knowledge of resources; 1e = designing coherent instruction; 1f = designing student assessments; 2a = creating an environment of respect and rapport; 2b = establishing a culture for learning; 2c = managing classroom procedures; 2d = managing student behavior; and 2e = organizing physical space.

Table 4
Distribution of Component Ratings for Domains 3 and 4

Components	3a	3b	3c	3d	3e	4a	4b	4c	4d	4e	4f
Unsatisfactory	0	0	0	0	0	0	0	0	0	0	0
Basic	5	4	6	5	1	2	1	1	4	2	1
Proficient	82	132	109	146	139	75	150	156	114	134	89
Distinguished	113	64	85	49	60	123	49	43	82	64	110
N	200	200	200	200	200	200	200	200	200	200	200

Note. Independent variables: 3a = communicating with students, 3b = using questioning and discussion techniques; 3c = engaging students in learning; 3d = using assessment in instruction; 3e = demonstrating flexibility and responsiveness; 4a = reflecting on teaching; 4b = maintaining accurate records; 4c = communicating with families; 4d = participating in the professional community; 4e = growing and developing professionally; 4f = showing professionalism.

Table 5 shows the mean percentage and standard deviation of the percentage of teachers' students attaining projected growth scores in reading, ranging from 15.4% to 85.7%. The percentages were normally distributed. The mean indicates the sum of the percentages of students attaining projected growth scores for all 200 teachers' divided by 200. The standard deviation describes the variability in the percentage of teachers' students attaining projected growth scores in reading and indicates that approximately 68.2% of the 200 teachers had 58.7471% ± 13.42370% of students attain projected growth scores.

Table 5

Percentage of Students Attaining Projected Growth Scores

Mean	Standard deviation	*N* (Teachers)
58.7471%	13.42370%	200

Tests of Assumptions

Before conducting the multiple correlation analysis, I tested and confirmed six assumptions regarding the variable data (Laerd Statistics, 2015; Rovai et al., 2014). I tested and confirmed the assumption of linearity between the dependent variable data and the collective independent variable data by visual inspection of a scatter plot of studentized residuals and unstandardized predicted values (see Appendix B). I tested and confirmed the assumption of independence of observations with a Durbin–Watson statistic of 2.043. I tested and confirmed the assumption of homoscedasticity by visual inspection of the plot of studentized residuals and unstandardized predicted values (see Appendix B). I tested and confirmed the assumption that no multicollinearity existed among independent variables by tolerance test values greater than 0.1. I tested and confirmed the assumption that no significant outliers emerged. No studentized deleted residuals greater than ± 3 standard deviations emerged. The leveraged values for 195 of the 200 cases were safe (< .2), five cases were risky (.2 to .5) with leveraged values among those five cases ranging from .20057 to .21022, and no leveraged values high (> .5). No Cook's distance values emerged above 1.0. I confirmed the assumption of normality by visual inspection of a histogram (see Appendix C). After testing and confirming the six assumptions, I conducted the multiple correlation analysis.

Finding 1

Research question 1. Do teachers' performance ratings on all 22 FFT

components explain a significant proportion of the variance in the percentage of teachers' students attaining projected growth scores in reading?

Null hypothesis 1. H_0: FFT component ratings do not explain a significant proportion of the variance in the percentage of teachers' students attaining projected growth scores in reading.

The multiple correlation analysis determined the coefficient of multiple correlation ($\underline{R}^2$) for the overall model was 11.5%, with an adjusted R^2 of .005 (see Table 6). Regarding Research Question 1, the model based on all 22 FFT components found the variance in the percentage of teachers' students attaining projected growth scores in reading between teachers rated basic or proficient and teachers rated distinguished was not statistically significant, $F(22, 177) = 1.041$, $p = .417$, adj. $R^2 = .5\%$. The *F*-statistic is the ratio of the between group variance and the within-group variance. When the null hypothesis is true, an *F* value close to 1 is expected (see Table 7). The numbers 22 and 177 are the degrees of freedom in the regression model and residuals, respectively. The *p* value indicates the probability of obtaining the observed *F*-value if the null hypothesis is true. The adjusted R^2 is the percentage of variance in the percentage of teachers' students attaining projected growth scores explained by the teachers' ratings on all 22 FFT components, adjusted for the positive bias of the sample. As the *p* value of .417 is greater than .05 ($p > .05$), Null Hypothesis 1 was not rejected. Table 6 summarizes the model. Table 7 summarizes the ANOVA.

Table 6

Model Summary

Model	*R*	R^2	Adjusted R^2	Standard error of the estimate	Durbin–Watson
1	.339[a]	.115	.005	13.39305%	2.043

Note. Independent variables: (constant), 1a = demonstrating knowledge of content and pedagogy; 1b = demonstrating knowledge of students; 1c = setting instructional outcomes; 1d = demonstrating knowledge of resources;1e = designing coherent instruction; 1f = designing student assessments; 2a = creating an environment of respect and rapport; 2b = establishing a culture for learning; 2c = managing classroom procedures; 2d = managing student behavior; 2e = organizing physical space; 3a = communicating with students; 3b = using questioning and discussion techniques; 3c = engaging students in learning; 3d = using assessment in instruction; 3e = demonstrating flexibility and responsiveness; 4a = Reflecting on Teaching; 4b = maintaining accurate records; 4c = communicating with families; 4d = participating in the professional community; 4e = growing and developing professionally; 4f = showing professionalism.
[a]Percentage of students attaining projected growth scores in reading.

Table 7

ANOVA

	Model	Sum of squares	*df*	mean square	*F*	Sig.
1	Regression	4109.813	22	186.810	1.041	.417
	Residual	31749.144	177	179.374		
	Total	35858.957	199			

Dependent variable: percentage of students attaining projected growth targets. Independent variables: 1a = demonstrating knowledge of content and pedagogy; 1b = demonstrating knowledge of students; 1c = setting instructional outcomes; 1d = demonstrating knowledge of resources;1e = designing coherent instruction; 1f = designing student assessments; 2a = creating an environment of respect and rapport; 2b = establishing a culture for learning; 2c = managing classroom procedures; 2d = managing student behavior; 2e = organizing physical space; 3a = communicating with students; 3b = using questioning and discussion techniques; 3c = engaging students in learning; 3d = using assessment in instruction; 3e = demonstrating flexibility and responsiveness; 4a = reflecting on teaching; 4b = maintaining accurate records; 4c = communicating with families; 4d = participating in the professional community; 4e = growing and developing professionally; 4f = showing professionalism.

Finding 2

Research question 2. Do teachers' performance ratings on any individual FFT component explain a significant proportion of the variance in the percentage of teachers' students attaining projected growth scores in reading?

Null hypothesis 2. H_0: No FFT component rating explains a significant proportion of variance in the percentage of teachers' students attaining projected growth scores in reading.

Regarding Research Question 2, the mean percentage of students attaining projected growth scores in reading was 8.139% higher for teachers rated distinguished than teachers rated basic or proficient on component *3d: Using Assessment in Instruction.* The multiple correlation analysis determined the unstandardized regression coefficient for FFT component *3d: Using Assessment in Instruction* is 8.139 ($B = 8.139$) and that it was statistically significant ($p < .05$); therefore, Null Hypothesis 2 was rejected. The unstandardized regression coefficients for the other 21 FFT components were not statistically significant. The teachers' performance ratings on 10 of the other 21 components had positive correlations and 11 had negative correlations with the percentage of students attaining projected growth scores in reading. The unstandardized coefficients for all 22 individual components ranged from -5.293 to 8.139.

Tables 8 and 9 provide a summary of the coefficients for each FFT component. The unstandardized coefficient (B) indicates the average difference in the percentage of students attaining projected growth scores in reading between teachers rated basic or proficient and teachers rated distinguished on the associated component when the ratings on all other components are held constant. Positive numbers indicate an increase and negative numbers indicate a decrease in the percentage of students attaining projected growth scores in reading when teachers are rated distinguished on the associated component. The standard error of the coefficient (SE_B) is a measure of the precision of the estimate of the coefficient; the smaller the standard error, the more precise the estimate of the coefficient. The standardized coefficient provides the change in the average difference in the percentage of students attaining projected growth scores in reading when teachers' component ratings change by one standard deviation. The significance column provides the p-value or statistical significance of the coefficient.

Unstandardized coefficients are considered statistically significant if the *p*-value is less than .05.

Table 8 shows that teachers rated distinguished on four FFT components had higher percentages of students attaining projected growth scores than teachers rated basic or proficient with unstandardized correlation coefficients ranging from .126 to 3.014, but the coefficients were not statistically significant ($p > .05$).

Table 8

Multiple Correlation Analysis: Domains 1 and 2 Components

	Variables[a]	*B*	SE_B	β	Significance
1	(Constant)	55.068	2.019		.000
	1a	-2.850	2.943	-.098	.334
	1b	2.812	2.820	.104	.320
	1c	-.897	2.481	-.033	.718
	1d	.126	2.781	.004	.964
	1e	1.152	2.693	.043	.669
	1f	-5.153	3.511	-.125	.144
	2a	3.014	2.702	.103	.266
	2b	-5.293	2.877	-.197	.067
	2c	-.200	2.631	-.007	.939
	2d	-.058	2.842	-.002	.984
	2e	-.442	2.702	-.014	.870

Note. *B* = unstandardized regression coefficient; SE_B = Standard error of the coefficient; β = standardized coefficient.

[a]Variables: 1a = demonstrating knowledge of content and pedagogy; 1b = demonstrating knowledge of students; 1c = setting instructional outcomes; 1d = demonstrating knowledge of resources; 1e = designing coherent instruction; 1f = designing student assessments; 2a = creating an environment of respect and rapport; 2b = establishing a culture for learning; 2c = managing classroom procedures; 2d = managing student behavior; 2e = organizing physical space.

Table 9 shows that teachers rated distinguished on seven components had higher percentages of students attaining projected growth scores in reading than teachers rated basic or proficient, with unstandardized correlation coefficients ranging from .718 to 8.139. Component 3d: *Using Assessment in Instruction* was the only component with a statistically significant unstandardized coefficient ($p < .05$).

Table 9

Multiple Correlation Analysis: Domains 3 and 4 Components

	Variables[a]	*B*	SE_B	β	Significance
1	(Constant)	55.068	2.019		.000
	3a	-1.998	2.901	-.074	.492
	3b	.718	2.555	.025	.779
	3c	1.986	2.801	.073	.479
	3d	8.139	2.881	.261	.005
	3e	2.453	2.737	.084	.371
	4a	4.723	2.501	.172	.061
	4b	1.694	2.642	.054	.522
	4c	-3.701	2.713	-.113	.174
	4d	-2.511	2.333	-.092	.283
	4e	-1.879	2.585	-.065	.468
	4f	1.295	2.619	.048	.622

Note. B = unstandardized regression coefficient; SE_B = Standard error of the coefficient; β = standardized coefficient.

[a]Variables: 3a = communicating with students; 3b = using questioning and discussion techniques; 3c = engaging students in learning; 3d = using assessment in instruction; 3e = demonstrating flexibility and responsiveness; 4a = reflecting on teaching; 4b = maintaining accurate records; 4c = communicating with families; 4d = participating in the professional community; 4e = growing and developing professionally; 3d = using assessment in instruction; 4f = showing professionalism.

Summary

The purpose of this study was to examine the correlation between teachers' evaluation ratings and their students' attainment of projected growth scores in reading. The independent variables were the Danielson (2007) FFT component ratings assigned to teachers of third-, fourth-, and fifth-grade students in a western Maryland school district during the 2013–2014 and 2014–2015 school years. The dependent variable was the percentage of each evaluated teachers' students attaining projected growth scores in reading, measured by the MAP reading assessment.

A multiple correlation analysis determined that the collective component ratings explained 11.5% (R^2 = .115) of the variance in the percentage of teachers' students attaining projected growth scores. The adjusted R^2 was .5% (adj. R^2 = .005) and was not

statistically significant ($p > .05$). The unstandardized coefficient for one component, *3d: Using Assessment in Instruction*, was 8.139 and was statistically significant ($p < .05$), meaning teachers rated distinguished had an 8.139% higher mean percentage of students attain projected growth scores in reading than teachers rated basic or proficient. Teachers' ratings on 10 other FFT components had positive correlations to students' attainment of growth projections in reading; however, the correlations were not statistically significant ($p > .05$). The unstandardized coefficients for all 22 individual components ranged from -5.293 to 8.139. Chapter 5 provides conclusions, implications, and recommendations for future research.

Chapter 5 – Conclusions and Implications

Introduction

The framework for this study was predicated on three related concepts. Teacher evaluations should provide accurate and relevant information that can be used to improve teacher performance, make personnel decisions, and ultimately improve student learning (Danielson & McGreal, 2000; Gullickson, 2009; JCSEE, 1988). The accuracy and relevance of the information provided by teacher evaluations rests on the strength of the association between the evaluation criteria and student learning (Kane, 2012; Marzano, 2012; Milanowski, 2004; Milanowski et al., 2004). The strength of the association between evaluation criteria and student learning can be determined by measuring the strength of the correlation between changes in the level of performance of the evaluation criteria and changes in the level of student learning (Kane, 2012; Kane & Cantrell, 2013; Marzano, 2012; Milanowski et al., 2004). Previous researchers discerned a weak to moderate correlation between teacher ratings on a subset of FFT components and a variety of value-added achievement data (Borman, & Kimball, 2005; Holtzapple, 2003; Kane & Staiger, 2012; Milanowski et al., 2004; Sartain et al., 2011). An examination of the correlation between all 22 FFT components and a nationally normed measure of student learning would broaden understanding of the strength of the relationship between the FFT and student learning. The purpose of this study was to examine the correlation between teachers' evaluation ratings and their students' attainment of projected growth scores in reading. The conclusion and implications of this study are discussed in this chapter.

Conclusions

Research question 1. Do teachers' performance ratings on all 22 FFT

components explain a significant proportion of the variance in the percentage of teachers' students attaining projected growth scores in reading?

Collectively, teachers' ratings on FFT components did not have a significant correlation to their students' attainment of projected growth scores in reading. A multiple correlation analysis determined teachers' collective component ratings explained 11.5% (R^2 = .115) of the variance in the percentage of their students' attaining projected growth scores. The adjusted R^2 was .5% (adj. R^2 = .005) and was not statistically significant ($p > .05$). Based on this finding, the FFT does not provide accurate and relevant information that can be used to improve teacher performance, make personnel decisions, and ultimately improve student learning.

Research question 2. Do teachers' performance ratings on any individual FFT component explain a significant proportion of the variance in the percentage of teachers' students attaining projected growth scores in reading?

With one exception, teachers' ratings on individual FFT components did not have a significant correlation to their students' attainment of projected growth scores in reading. A multiple correlation analysis determined the unstandardized coefficient for one component, *3d: Using Assessment in Instruction*, was 8.139 and was statistically significant ($p < .05$), meaning teachers rated distinguished had an 8.139% higher mean percentage of students attain projected growth scores in reading than teachers rated basic or proficient. The unstandardized coefficients for the other 21 individual FFT components ranged from -5.293 to 4.723; however, the correlations were not statistically significant ($p > .05$). Based on this finding, only one FFT component provides accurate and relevant information that can be used to improve teacher performance, make personnel decisions, and ultimately improve student learning.

Findings from this study concerning the correlation between evaluation ratings and student learning are consistent with earlier research that found weak to moderate relationships between teachers' evaluation ratings and student learning (Borman & Kimball, 2005; Holtzapple, 2003; Kane & Staiger, 2012; Milanowski et al., 2004; Sartain et al., 2011). Additionally, descriptive statistics pertaining to the data used in this study are consistent with earlier research that showed teacher evaluations do not differentiate levels of performance (Kraft & Gilmour, 2016; Weisberg et al., 2009). This study examined 4,400 ratings assigned to 200 teachers for the performance of 22 FFT components over a period of 2 years. During that time, teachers' performance was never rated unsatisfactory, was rated basic only 68 times, was rated proficient 2,700 times, and was rated distinguished 1,732 times, suggesting the FFT does not frequently identify subproficient levels of performance.

Implications

Personnel decisions. This study has implications for the use of teachers' FFT ratings to make personnel decisions, especially decisions regarding the awarding of tenure, promotions, or terminations. The findings from this study suggest that teachers performing most of the teaching practices assessed by the FFT at the distinguished level are not necessarily highly effective teachers, meaning those teachers' students are not necessarily learning at a higher rate or level than the students of teachers rated basic or proficient. Similarly, teachers performing at the basic or proficient level are not necessarily less effective teachers than teachers rated distinguished. As discussed in Chapter 1, the problem with using teacher evaluations that do not accurately differentiate levels of teacher effectiveness is that school districts cannot reward highly effective teachers, justify personnel decisions related to competency, or provide targeted

professional development to improve teacher effectiveness (Weisberg et al., 2009) based on those evaluations. Personnel decisions based on FFT component ratings could be contested based on a lack of relevance to student learning. Schools and school districts should consider using additional data sources to inform and justify personnel decisions, including a variety of direct measures of student learning, student portfolios, and student surveys (Kane & Cantrell, 2013; Strunk et al., 2014).

Professional-development decisions. This study also has implications for the use of the FFT to target professional development. Strunk et al. (2014) recommended that school districts identify the teaching practices in their observation instruments that are most closely associated with student achievement and focus their professional development on those teaching practices.

This study found one FFT component closely aligned with student achievement in reading. Component 3d: *Using Assessment in Instruction* explained a significant proportion (8.139%) of the variance in the percentage of teachers' students attaining growth projections. The component includes four elements: assessment criteria, monitoring of student learning, feedback to students, and student self-assessment and monitoring of progress (Danielson, 2007).

The findings from this study suggest providing professional development aimed at improving teacher performance of each element of component 3d: *Using Assessment in Instruction* has value. Performance-level descriptors that separate proficient from distinguished levels of performance could inform the objectives and assessments of professional development. The distinguished level of performance of the *assessment criteria* element requires students to participate in the creation of the assessment criteria (Danielson, 2007). The distinguished level of performance of the *monitoring of student*

learning element requires teachers to monitor the progress of students individually (Danielson, 2007). The distinguished level of performance of the *feedback to students* element requires students to make use of the feedback (Danielson, 2007). The distinguished level of performance of the *student self-assessment and monitoring of progress* element requires students to make use of the information to further learning (Danielson, 2007). Findings from this study suggest that professional development that raises teacher performance of each element of component 3d: *Using Assessment in Instruction* to the distinguished level would have a positive impact on student learning.

Another implication for the use of the FFT to target professional development is more tenuous, but should be considered. Although the corresponding correlation coefficients were not statistically significant, the multiple correlation analysis found three components that each explained a notable percentage of the variance in the percentage of students attaining projected growth scores in reading. Component 1b: *Demonstrating Knowledge of Students* (Danielson, 2007) explained 2.812% of the variance; component 2a: *Creating and Environment of Respect and Rapport* (Danielson, 2007) explained 3.014% of the variance, and component 4a: *Reflecting on Teaching* (Danielson, 2007) explained 4.723% of the variance. Providing professional development to improve the performance of teachers identified as less than distinguished on the aforementioned components may have a positive impact on student learning. Findings from this study suggest providing professional development to improve teacher performance for 18 of the 22 FFT components has no value.

Designing teacher evaluations. This study also has implications for the design of teacher evaluations. Evaluation systems designed to inform personnel decisions and professional-development decisions based on the performance of fixed, standards-based

criteria may not serve either purpose as well as two separate systems. Educators should consider eliminating the need to balance the two purposes. Marzano (2012) asserted that in balancing the two purposes of evaluations, measuring performance and improving performance, evaluations should lean toward improving performance. Danielson (2007) asserted that professional growth is the primary focus of the FFT. A nonevaluative system of observations and feedback could be used to direct and support professional development. Teacher leaders with no supervisory authority could conduct the nonevaluative observations. Teachers could then decide to use the feedback to improve their performance of self-selected, research-based teaching practices (Brophy & Good, 1986; Nye et al., 2004). Professional development could focus on personal strengths and weaknesses, the needs of students, and the specific curriculum being taught, rather than adherence to a fixed set of general criteria.

The evaluations used to inform personnel decisions could be conducted less frequently and focus on new and struggling teachers. With fewer evaluations to conduct, administrators could evaluate teacher effectiveness based on a variety of evidence of student learning. Evaluators could consider the results of standardized assessments, teacher-created assessments, and portfolios of student work collected over time.

Recommendations for Future Research

Findings from this study warrant a recommendation for replication studies to determine if the findings would be similar with teachers of different grade levels and different subject areas. Additionally, findings from this study suggest several areas for further research, including education reform and accountability policies, redesigning teacher-evaluation systems, and the identification of effective teaching practices.

Recent policies that influenced the use of standards-based and student-growth

measures in teacher evaluations have not significantly improved the ability of the evaluations to identify variations in teacher performance (Kraft & Gilmour, 2016). The generally weak correlations between an individual teacher's evaluation ratings and the academic achievement of the teacher's students raises questions regarding the utility of standards-based criteria as a strategy to improve the academic achievement of students.

Research questions regarding the manner in which the results of evaluations have been used in personnel and professional-development decisions should be investigated. Researchers should investigate the extent to which individualized professional development has been provided to teachers based on the results of evaluations and the frequency with which teachers are awarded or denied tenure, terminated, or promoted based on the results of evaluations. Additionally, questions regarding the validity and use of standardized measures of student growth to identify teachers' strengths and weaknesses should be investigated (Darling-Hammond et al., 2012; Ehlert et al., 2013). Such research could provide policymakers with information that could improve current teacher-accountability policies.

Research is also needed to improve the criteria and processes used for teacher evaluations. Teacher-evaluation systems currently use similar research to select criteria and processes to collect evidence and rate teacher performance (Danielson, 2007; Marshall, 2013; Marzano, 2007; Morgan et al., 2014; Steinberg & Donaldson, 2014; Stronge, 2012). Researchers should identify the evaluation criteria teachers believe are most effective for facilitating student learning and the teaching practices students believe are most important to facilitating student learning. Researchers should investigate the validity and reliability of using teacher observations to collect evidence of teacher effectiveness as well as the value of alternative methods of collecting evidence of teacher

effectiveness. Researchers should investigate the value of separating the accountability and professional-growth functions of teacher evaluations. Research on the criteria and processes used in teacher evaluations could help improve the utility of teacher evaluations.

Researchers should also focus on identifying the teaching practices that most effectively facilitate student learning by measuring the strength of the correlation between changes in the level of performance of teaching practices and changes in the level of student learning (Kane, 2012; Kane & Cantrell, 2013; Marzano, 2012; Milanowski et al., 2004). Although considerable research has been conducted in this area (Brophy & Good, 1986; Nye et al., 2004), additional research is needed, especially quantitative research on assignment-specific teaching methods. Researchers should aim to identify the teaching practices that are most effective for teaching specific grade levels and subjects and teaching students with specific learning disabilities. Identifying effective strategies for teaching all students to meet grade-level subject standards would allow school districts to provide professional development based on teachers' current teaching assignments rather than the more general criteria found in standards-based evaluations.

Research is also needed to identify measures of students' academic growth to support teacher accountability. Researchers should address the validity of a variety of student assessments for all subjects and grade levels and help identify methods to control for categorical differences among students (Darling-Hammond et al., 2012; Ehlert et al., 2013).

Limitations

The scope of the data collected and analyzed in this study may limit the extent to which the findings can be applied to other subjects, grade levels, and school districts. In

this study, I used growth in reading as the only measure of student learning, which may not be generalizable to other curricular content or subjects. I examined the performance ratings of teachers of third-, fourth-, and fifth-grade students; thus, results may not be generalizable to teachers of students in lower or higher grades. Data for this study were collected from one school district and may not be applicable to school districts that use different curricula or resources.

Summary

The purpose of this study was to examine the correlation between teachers' evaluation ratings and their students' attainment of projected growth scores in reading. A multiple correlation analysis concluded that the 22 FFT components collectively explained 11.5% of the variance in the percentage of teachers' students attaining projected growth scores, but was not statistically significant. Component *3d: Using Assessment in Instruction* FFT explained 8.139% of the variance in the percentage of teachers' students attaining projected growth scores in reading and was statistically significantly ($p < .05$). Teachers' ratings on 10 other the FFT components had positive correlations to students' attainment of growth projections in reading, but the correlations were not statistically significant.

Findings from this study suggest several areas for further research, including education reform and accountability policies, teacher-evaluation-systems design, identification of effective teaching practices, and identifying valid measures of student growth to support teacher-accountability efforts. This study has two implications regarding the utility of the FFT as a measure of teacher effectiveness. Teachers' FFT performance ratings should not be the sole source of information for making personnel decisions, especially decisions regarding tenure, promotions, and terminations. Teachers'

FFT performance ratings have limited value for targeting professional development toward teaching practices that impact student learning.

References

American Recovery and Reinvestment Act, 16 U.S.C. §§ 1601 *et seq*. (2009).

Andren, K. J. (2010). *An analysis of the concurrent and predictive validity of curriculum based measures, the measures of academic progress, and the New England common assessment program for reading* (Doctoral dissertation, University of Southern Maine, Portland). Retrieved from https://usm.maine.edu /sites/default/files/School%20Psychology/Andren.pdf

Armour-Garb, A. (2009). Should "value-added" models be used to evaluate teachers? *Journal of Policy Analysis and Management, 28,* 692–712. doi:10.1002/pam .20462

Balch, R., & Koedel, C. (2014). Anticipating and incorporating stakeholder feedback when developing value-added models. *Education Policy Analysis Archives, 22*(97), 1–17. doi.org/10.14507/epaa.v22.1701

Ballou, D., & Springer, M. (2015) Using student test scores to measure teacher performance: Some problems in the design and implementation of evaluation systems. *Educational Researcher, 44,* 77–86. doi:10.3102/0013189X15574904

Betebenner, D. W. (2008). *Norm-and criterion-referenced student growth*. Retrieved from the National Center for the Improvement of Educational Assessment website: http://www.nciea.org/publications/normative_criterion_growth_DB08 .pdf

Borman, G., & Kimball, S. (2005). Teacher quality and educational equality: Do teachers with higher standards-based evaluation ratings close student achievement gaps? *The Elementary School Journal. 106,* 2–19. doi:10.1086/496904

Brophy, J., & Good, T. (1986). *Teacher behavior and student achievement.* (Occasional Paper No. 73). The Institute for Research on Teaching. Retrieved from http://education.msu.edu/irt/PDFs/OccasionalPapers/op073.pdf

Collins, C. (2014). Houston, we have a problem: Teachers find no value in the SAS education value-added assessment system. *Education Policy Analysis Archives, 22*(98), 1–42. doi:10.14507/epaa.v22.1594

Cordray, D., Pion, G., Brandt, C., Molefe, A., & Toby, M. (2012). *The impact of the measures of academic progress program on student reading achievement.* Washington, DC: Government Printing Office. Retrieved from http://ies.ed.gov/ncee/edlabs/regions/midwest/pdf/REL_20134000.pdf

Council of Chief State School Officers. (2011). *Interstate teacher assessment and support consortium model core teaching standards: A resource for state dialogue.* Retrieved from http://ccsso.org/Documents/2011/InTASC_Model_Core_Teaching_Standards_2011.pdf

Danielson, C. (2007). *Enhancing professional practice: A framework for teaching* (2nd ed.). Alexandria, VA: Association for Supervision and Curriculum Development.

Danielson, C. (2009). *Implementing the framework for teaching in enhancing professional practice.* Alexandria, VA: Association of Supervision and Curriculum Development.

Danielson Group. (2011). *Framework for teaching sees record growth.* Retrieved from https://www.danielsongroup.org/press-item/framework-for-teaching-sees-record-growth/

Danielson, C., & McGreal, T. (2000). *Teacher evaluation to enhance professional practice*. Alexandria, VA: Association for Supervision and Curriculum Development.

Darling-Hammond, L., Amrein-Beardsley, A., Haertel, E., & Rothstein, J. (2012). Evaluating teacher evaluation. *Phi Delta Kappan, 93*(6), 8–15. doi:10.1177/003172171209300603

Dillon, N. (2008). The merit scale. *American School Board Journal, 195*(4), 28–30. Retrieved from http://www.asbj.com

Dwyer, C. (1998). Psychometrics of Praxis III: Classroom performance assessments. *Journal of Personnel Evaluation in Education, 12,* 163–87. doi:10.1023/A:1008033111392

Eckert, J., & Dabrowski, J. (2010). Should value-added measures be used for performance pay? *Phi Delta Kappan, 91*(8), 88–92. doi:10.1177/003172171009100821

Ehlert, M., Koedel, C., Parsons, E., & Podgursky, P. (2013). *Selecting growth measures for school and teacher evaluations: Should proportionality matter?* (Working paper). Retrieved from http://www.caldercenter.org/sites/default/files/wp-80-updated-v3.pdf

Elementary and Secondary Education Act of 1965, 20 U.S.C. §§ 6301 *et seq.* (1965)

Every Student Succeeds Act, 20 U.S.C. §§ 6301 *et seq.* (2015).

Georgia Department of Education. (2014). *Overview/executive summary of the 2012-2013 TKES and LKES evaluation report.* Retrieved from http://www.gadoe.org/School-Improvement/Teacher-and-Leader-Effectiveness/Documents/FINAL%20Year%203%20Report%20_2-21-2014_FORMATTED%202-23-2014.pdf

Gilbert, A. (2014). *Teacher evaluation tools: An examination of Wyoming evaluation models as a predictor of student achievement* (Doctoral dissertation, University of Wyoming, Laramie). Retrieved from https://books.google.com/books /about/Teacher_Evaluation_Tools.html?id=6-zDrQEACAAJ

Goldring, E., Grissom, J., Rubin, M., Neumerski, C., Cannata, M., Drake, T., & Schuermann, P. (2015). Make room value added: Principals' human capital decisions and the emergence of teacher observation data. *Educational Researcher, 44*, 96–104. doi:10.3102/0013189X15575031

Good, T., & Lavigne, A. (2014). Issues of teacher performance stability are not new: Limitations and possibilities. *Education Policy Analysis Archives, 23*(2), 2–16. doi:10.14507/epaa.v23.1916

Grossman, P., Loeb, S., Cohen, J., & Wyckoff, J. (2013). Measure for measure: The relationship between measures of instructional practice in middle school English language arts and teachers' value-added scores. *American Journal of Education, 119*, 445–470. doi:10.1086/669901

Gullickson, A. (2009). *The personnel evaluation standards*: *How to assess systems for evaluating educators.* Thousand Oaks, CA: Corwin.

Hanushek, E. A. (1971). Teacher characteristics and gains in student achievement: Estimation using micro data. *The American Economic Review, 61,* 280–288. Retrieved from https://www.aeaweb.org/journals/aer

Ho, A., Lewis, D., & MacGregor Farris, J. (2009). The dependence of growth-model results on proficiency cut scores. *Educational Measurement: Issues and Practice, 28*(4), 15–26. doi:10.1111/j.1745-3992.2009.00159.x

Holtzapple, E. (2003). Criterion-related validity evidence for a standards-based teacher evaluation system. *Journal of Personnel Evaluation in Education, 17*, 207–219. doi.org/10.1007/s11092-005-2980-z

Hull, J. (2013). *Trends in teacher evaluation: How states are measuring teacher performance*. Retrieved from Center for Public Education website: http://www.centerforpubliceducation.org/Main-Menu/Evaluating-performance/Trends-in-Teacher-Evaluation-At-A-Glance/Trends-in-Teacher-Evaluation-Full-Report-PDF.pdf

Improving America's Schools Act of 1994, 20 U.S.C. §§ 6301 *et seq.* (1994).

Jensen, N. (2013). *Using the percentage of students meeting or exceeding their growth projections as an evaluation tool* [Web log]. Retrieved from https://www.nwea.org/blog/2013/using-percentage-students-meeting-exceeding-growth-projections-evaluration-tool/

Joint Committee on Standards for Educational Evaluation. (1988). *Personnel evaluation standards: Summary of the standards*. Retrieved from http://www.jcsee.org/personnel-evaluation-standards

Kane, T. (2012). Capturing the dimensions of effective teaching. *Education Next, 12*(4), 34–41. Retrieved from http://educationnext.org

Kane, T., & Cantrell, S. (2013). *Ensuring fair and reliable measures of effective teaching: Culminating findings from the MET project's three-year study* [Policy brief]. Retrieved from http://www.edweek.org/media/17teach-met1.pdf

Kane, T., McCaffrey, D., Miller, T., & Staiger, D. (2013). *Have we identified effective teachers? Validating measures of effective teaching using random assignment*. Retrieved from http://www.hec.ca/iea/seminaires/140401_staiger_douglas.pdf

Kane, T., & Staiger, D. (2012). *Gathering feedback for teaching: combining high-quality observations with student surveys and achievement gains* [Policy brief]. Retrieved from http://files.eric.ed.gov/fulltext/ED540960.pdf

Kane, T., Taylor, E., Tyler, J., & Wooten, A. (2011). Identifying effective classroom practices using student achievement data. *Journal of Human Resources, 46,* 587–613. doi.org/10.1353/jhr.2011.0010

Kimball, S., & Milanowski, A. (2009). Examining teacher evaluation validity and leadership decision making within a standards-based evaluation system. *Educational Administration Quarterly, 45,* 34–70. doi:10.1177 /0013161X08327549

Kraft, M., & Gilmour, A. (2016). *Revisiting the widget effect: Teacher evaluation reforms and the distribution of teacher effectiveness* (Working paper). Retrieved from http://scholar.harvard.edu/mkraft/publications/revisiting-widget-effect -teacher-evaluation-reforms-and-distribution-teacher

Lachlan-Haché, L., Cushing, E., & Bivona, L. (2012). *Student learning objectives as measures of educator effectiveness: The basics.* Retrieved from American Institutes for Research website: http://educatortalent.org/inc/docs/SLOs_Measures _of_Educator_Effectiveness.pdf

Ladd, H., & Lauen, D. (2010). Status versus growth: The distributional effects of school accountability policies. *Journal of Policy Analysis and Management, 29,* 426–450. doi:10.1002/pam.20504

Laerd Statistics. (2015). *Multiple correlation using SPSS Statistics, statistical tutorials and software guides.* Retrieved from https://statistics.laerd.com/

Marshall, K. (2009). *Rethinking teacher supervision and evaluation: How to work smart, build collaboration, and close the achievement gap.* San Francisco, CA: Jossey-Bass.

Martin, L. (2014). *Identifying the relationship between the MAP and WJ-III reading tests to make instructional decisions within a RTI framework* (Unpublished doctoral dissertation) Western Kentucky University, Bowling Green. Retrieved from http://digitalcommons.wku.edu/cgi/viewcontent.cgi?article=1060&context=diss

Maryland State Department of Education. (2012). *Maryland teacher and principal evaluation guidebook.* Retrieved from http://archives.marylandpublicschools.org/NR/rdonlyres/E479C243-AF58-4BD0-B4E8-46677F7757A0/33345/MDTeacherPrincipalReport_041212_rev0912_.pdf

Maryland State Department of Education. (2013). *Maryland's plan for implementing teacher and principal evaluation.* Retrieved from http://archives.marylandpublicschools.org/MSDE/programs/tpe/mdplan_tp_eval.html

Maryland State Department of Education. (2014a). *Maryland state model teacher evaluation model and component measure elements.* Retrieved from http://marylandpublicschools.org/msde/programs/tpe/docs/State_LEA_TPE_Models.pdf

Maryland State Department of Education. (2014b). *Spring 2014 teacher and principal evaluation ratings report to the Maryland state board of education.* Retrieved from http://archives.marylandpublicschools.org/MSDE/programs/tpe/docs/Spring2014TeacherPrincipalAnalysis.pdf

Marzano, R. (2003). *What works in school: Translating research into action*. Alexandria, VA: Association for Supervision and Curriculum Development.

Marzano, R. (2007). *The art and science of teaching: A comprehensive framework for effective instruction*. Alexandria, VA: Association of Supervision and Curriculum Development.

Marzano, R. (2012). The two purposes of teacher evaluation. *Educational Leadership, 70*(3), 14–19. Retrieved from http://www.ascd.org/publications/educational-leadership.aspx

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

McGreal, T. (1995). Characteristics of successful teacher evaluation. In A. Shinkfield, D. Stufflebeam, & N. Madaus (Eds.), *Teacher evaluation: A guide to effective practice* (pp. 208–230). Norwell, MA: Kluwer Academic.

McMillan, J., & Schumacher, S. (2010). *Research in education.* Upper Saddle River, NJ: Pearson Education.

Michigan Department of Education (2014). *Educator evaluations & effectiveness in Michigan.* Retrieved from http://www.michigan.gov/documents/mde/Educator_Evaluations_and_Effectiveness_Report_455793_7.pdf

Milanowski, A. (2004). The relationship between teacher performance evaluation scores and student achievement: Evidence from Cincinnati. *Peabody Journal of Education, 79,* 33–53. doi:10.1207/s15327930pje7904_3

Milanowski, A., Kimball, S., & White, B. (2004). *The relationship between standards-based teacher evaluation scores and student achievement: Replication and extensions at three sites* [Policy paper]. Retrieved from http://www.cpre.wceruw.org/papers/3site_long_TE_SA_AERA04TE.pdf

Morgan, G., Hodge, K., Trepinksi, T., & Anderson, L. (2014). The stability of teacher performance and effectiveness: Implications for policies concerning teacher evaluation. *Education Policy Analysis Archives, 22*(95), 1–18. doi:10.14507/epaa.v22n95.2014

Mueller, L. (2011). How I-O can contribute to the teacher evaluation debate: A response to Lefkowitz. *The Industrial-Organizational Psychologist, 49*(1), 17–21. Retrieved from http://www.siop.org/tip/july11/04mueller.aspx

National Board for Professional Teaching Standards. (2016). *The five core propositions.* Retrieved from http://www.boardcertifiedteachers.org/about-certification/five-core-propositions

Newton, X., Darling-Hammond, L., Haertel, E., & Thomas, E. (2010). Value-added modeling of teacher effectiveness: An exploration of stability across models and contexts. *Educational Policy Analysis Archives, 18*(23), 2–27. doi:10.14507/epaa.v18n23.2010

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

Northwest Evaluation Association. (2012). *Achievement status and growth summary report*. Retrieved from https://legacysupport.nwea.org/sites/www.nwea.org/files/resources/AnnotatedReports-MAP_0.pdf

Nye, B., Konstantopoulos, S., & Hedges, L. (2004). How large are teacher effects? *Educational Evaluation and Policy Analysis, 26,* 237–257. doi:10.3102/01623737026003237

Papay, J. (2011). Different tests, different answers: The stability of teacher value-added estimates across outcome measures. *American Educational Research Journal, 48,* 163–193. doi.org/10.3102/0002831210362589

Rockoff, J. (2004). The impact of individual teachers on student achievement: Evidence from panel data. *American Economic Review, 94,* 247–252. doi:10.1257/0002828041302244

Rovai, A., Baker, J., & Ponton, M. (2014). *Social science research and statistics.* Chesapeake, VA: Watertree.

Ryser, G., & Rambo-Hernandez, K. (2014). Using growth models to measure school performance: Implications for gifted learners. *Gifted Child Today, 37*(1), 17–23. doi:10.1177/1076217513509617

Sanders, W., & Horn, S. (1998). Research findings from the Tennessee value-added assessment system database: Implications for educational evaluation and research. *Journal of Personnel Evaluation in Education, 12,* 247–256. doi:10.1023/A:1008067210518

Sartain, L., Stoelinga, S., & Brown, E. (2011). *Rethinking teacher evaluation in Chicago*. Chicago, IL: University of Chicago. Retrieved from http://consortium.uchicago.edu/sites/default/files/publications/Teacher%20Eval%20Report%20FINAL.pdf

Scriven, M. (1988). Duty-based teacher evaluation. *Journal of Personnel Evaluation in Education, 1,* 319–334. doi:10.1007/BF00124098

Shinkfield, A., & Stufflebeam, D. (1995). *Teacher evaluation: A guide to effective practice*. Norwell, MA: Kluwer Academic.

Spring, J. (2011). *The American school: A global context from the puritans to the Obama era*. New York, NY: McGraw-Hill.

Spring, J. (2012). *American education*. New York, NY: McGraw-Hill.

Steinberg, M., & Donaldson, M. (2014) *The new educational accountability: Understanding the landscape of teacher evaluation in the post NCLB era* [Policy brief]. Retrieved from http://cepa.uconn.edu/wp-content/uploads/sites/399/2014/02/The-New-Educational-Accountability_policy-brief_8-19-14.pdf

Stronge, J. (2012). *Teacher effectiveness performance evaluation system handbook*. Williamsburg, VA: Stronge and Associates Educational Consulting.

Strunk, K., Weinstein, T., & Makkonen, R. (2014). Sorting out the signal: Do multiple measures of teachers' effectiveness provide consistent information to teachers and principals? *Education Policy Analysis Archives, 22*(100), 1–41. doi.org/10.14507/epaa.v22.1590

Stufflebeam, D., & Shinkfield, A. (2007). *Evaluation theory, models, & applications*. San Francisco, CA: Jossey-Bass.

Tennessee Department of Education. (2015). *Teacher evaluation in Tennessee: A report on year 3 implementation*. Retrieved from http://tn.gov/assets/entities/education/attachments/rpt_teacher_evaluation_year_3.pdf

The New Teacher Project. (2013). *Fixing classroom observations: How common core will change the way we look at teaching*. Retrieved from http://tntp.org/assets/documents/TNTP_FixingClassroomObservations_2013.pdf

Tyler, J., Taylor, E., Kane, T., & Wooten, A. (2010). Using student performance data to identify effective classroom practices. *American Economic Review, 100* 256–260. doi:10.1257/aer.100.2.256

U.S. Department of Education. (n.d.). *Targeting growth using student learning objectives as a measure of educator effectiveness*. Retrieved from http://www2.ed.gov/about/inits/ed/implementation-support-unit/tech-assist/targeting-growth.pdf

U.S. Department of Education. (2009). *Race to the top executive summary*. Retrieved from http://www2.ed.gov/programs/racetothetop/executive-summary.pdf

U.S. Department of Education. (2015). *Dear colleagues letter*. Retrieved from http://www2.ed.gov/policy/elsec/leg/essa/transition-dcl.pdf

U.S. Department of Education. (2016). *Implementing accountability and supports under ESEA flexibility.* Retrieved from https://www2.ed.gov/rschstat/eval/title-i /implementing-accountability-esea-flexibility/report.pdf

Wang, S., McCall, M., Hong, J., & Harris, G. (2013). Construct validity and measurement invariance of computerized adaptive testing: Applications to measures of academic progress using confirmatory factor analysis. *Journal of Educational and Developmental Psychology, 3*(1), 88–100. doi:10.5539/jedp .v3n1p88

Washington, A. (2011). *Formal evaluation of teachers: An examination of the relationship between teacher performance and student achievement* (Unpublished doctoral dissertation). University of South Carolina, Columbia. Retrieved from http://scholarcommons.sc.edu/etd/1016

Weisberg, D., Sexton, S., Mulhern, J., & Keeling, D. (2009). *The widget effect: Our national failure to acknowledge and act on differences in teacher effectiveness*. Retrieved from http://widgeteffect.org/downloads/TheWidgetEffect.pdf

Wright, S. P., Horn, S. P., & Sanders, W. L. (1997). Teacher and classroom context effects on student achievement: Implications for teacher evaluation. *Journal of Personnel Evaluation in Education*, 11, 57–67. doi:10.1007/s11092-016-9252-y

Appendix A – The Framework for Teaching

DOMAIN 1: Planning and Preparation

- 1a. Demonstrating Knowledge of Content and Pedagogy
- 1b. Demonstrating Knowledge of Students
- 1c. Setting Instructional Outcomes
- 1d. Demonstrating Knowledge of Resources
- 1e. Designing Coherent Instruction
- 1f. Designing Student Assessments

Domain 2: The Classroom Environment

- 2a. Creating an Environment of Respect and Rapport
- 2b. Establishing a Culture for Learning
- 2c. Managing Classroom Procedures
- 2d. Managing Student Behavior
- 2e. Organizing Physical Space

DOMAIN 3: Instruction

- 3a. Communicating With Students
- 3b. Using Questioning and Discussion Techniques
- 3c. Engaging Students in Learning
- 3d. Using Assessment in Instruction
- 3e. Demonstrating Flexibility and Responsiveness

DOMAIN 4: Professional Responsibilities

- 4a. Reflecting on Teaching
- 4b. Maintaining Accurate Records
- 4c. Communicating with Families

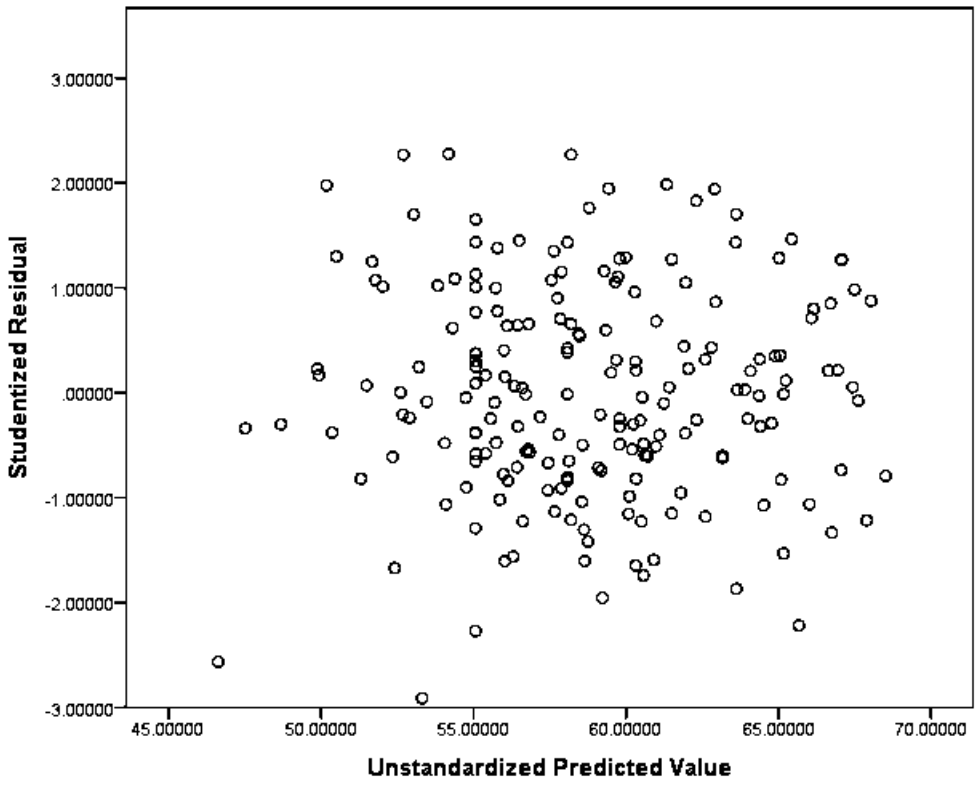
4d. Participating in a Professional Community

4e. Growing and Developing Professionally

4f. Showing Professionalism

Note. From *Enhancing professional practice: A framework for teaching* (pp. 3-4), by C. Danielson, 2007, Alexandria, VA: Association for Supervision and Curriculum Development. Copyright 2007 by the Association for Supervision and Curriculum Development. Reprinted with permission.

Appendix B – Studentized Residuals and Unstandardized Predicted Values



Appendix C – Distribution of Residuals Histogram