January 2013

Value-Added and Observational Measures Used in the Teacher Evaluation Process: A Validation Study

Claudia Güerere
University of South Florida, cguerere@mail.usf.edu

Value-Added and Observational Measures Used in the Teacher Evaluation Process: A Validation Study

by

Claudia Güerere

A dissertation submitted in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
Department of Educational Measurement and Research
College of Education
University of South Florida

Major Professor: Robert F. Dedrick, Ph.D.
Jeffrey D. Kromrey, Ph.D.
Liliana Rodríguez-Campos, Ph.D.
Waynne B. James, Ed.D.

Date of Approval:
May 28, 2013

Keywords: Validity, Educational Accountability, Value-Added Scores, Structural Equation Modeling, Teaching Quality, Teaching Effectiveness

Dedication

A mi mami y mi papi.

Por haberme apoyado en todo momento, por sus consejos, sus valores, por la motivación constante que me ha permitido convertirme en quien soy hoy dia. En gran parte gracias a ustedes, hoy puedo ver alcanzada mi meta, ya que siempre estuvieron impulsándome en los momentos más difíciles. Gracias por haber fomentado en mí el deseo de superación, aprendizaje y el anhelo de triunfo en la vida. Mil palabras no bastarían para agradecerles su apoyo, su comprensión y mas que todo, su amor.

For my mom and my dad.

For supporting me at all times, for your advice, your values, and the constant motivation that have allowed me to become who I am today. Largely thanks to you, my goals and dreams are coming to fruition, as you were always pushing me in the most difficult moments. Thank you for having encouraged in me the desire for self-improvement, learning and the desire to achieve in life. One thousand words would not be enough to thank you for your support, understanding and most of all, your love.

Acknowledgments

I thank my professors throughout my educational career and my committee members for taking the time to mentor me, guide me, and understand me. I especially thank Dr. Dedrick for fermenting in me the love for measurement and testing, Dr. Rodríguez-Campos for her solid advice and unwavering support, Dr. Kromrey for providing me calmness and laughter, and Dr. James for making me take that statistics class I did not want to take. I thank all four of you for believing in me. For your patience and dedication to my project I am forever indebted.

To all those who sat next to me while I worked:

Long hours seemed short with you by my side.

To all those who listened, supported and encouraged:

Thank you.

Table of Contents

List of Tables

List of Figures

List of Equations

Abstract

Scores from value-added models (VAMs), as used for educational accountability, represent the educational effect teachers have on their students. The use of these scores in teacher evaluations for high-stakes decision making is new for the State of Florida. Validity evidence that supports or questions the use of these scores is critically needed. This research, using data from 2385 teachers from 104 schools in one school district in Florida, examined the validity of the value-added scores by correlating these scores with scores from an observational rubric used in the teacher evaluation process. The VAM scores also were examined in relation to several variables that the literature had identified as correlates of quality teaching as well as variables that were theoretically independent of teacher performance.

The observational rubric used in the validation process was based on Marzano's and Danielson's framework and consisted of 34 items and five factors (Ability to Assess Instructional Needs, Plans and Delivers Instruction, Maintains a Student-Centered Learning Environment, Performs Professional Responsibilities, Engages in Continuous Improvement for Self and School). Analyses of the psychometric properties of the observational rubric using confirmatory factor analysis supported the fit of the five-factor structure underlying the rubric. Internal consistency reliabilities for the five observational scales and total score ranged from .81 to .96.

The relationships between the observational rubric scores and VAM scores (with and without the standard error of measurement (SE) applied to the VAM score) were generally weak for the overall sample (range of correlations = .05 to .09 for the five observational scales and VAM with SE; .14 to .18 for the five observational scales and VAM without SE). Inspection of the relationship between the VAM and total observational scores within each of the 104 schools revealed that while some schools had a strong relationship, the majority of the schools revealed little to no relationship between the two measures that represent a quality/effective teacher.

The last part of this research investigated the relationship of the VAM scores and scores from the observational rubric with variables that had been identified in the literature as correlates of quality teaching. In addition, relationships between variables that the literature had shown to be independent of quality teaching were also examined. Results indicated that VAM scores were not significantly related to any of the predictor variables (e.g., National Board Certification, years of experience, gender, etc.). The observational rubric, on the other hand, had significant relations with National Board Certification, years of experience, and gender.

The validity evidence provided in this research calls for caution when using VAM scores in teacher evaluations for high-stakes decision making. The weak relations between the observational scores of teachers' performance and teachers' value-added scores suggest that these measures are representing different dimensions of the multidimensional construct of teaching quality. Ongoing research is needed to better understand the strengths and limitations of both the observational and VAM measures and the reasons why these measures do not often converge. In addition, teacher factors (e.g., grade level) that can account for variation in both the VAM and observational scores need to be identified.

Chapter One: Introduction

Research has demonstrated that the quality of a teacher has a very strong influence on student achievement (Ferguson, 1998; Hanushek, Kain, & Rivkin, 1999; Hanushek, 1992; Kyriakides & Creemers, 2008; Rivkin, Hanushek, & Kain, 2005; Rowan, Correnti, & Miller, 2002; Sanders, 1998; Wright, Horn & Sanders, 1997). For this reason several attempts have recently been made to create more accountability for teachers in the classroom. For example, in an effort to focus on teacher accountability, President Obama signed a law in February 2009 that provided money to the Race to the Top Fund (RTTT). The goal of this fund was to provide incentives for states to adopt pay for performance standards and implement ways to tie teachers' pay to how well their students were doing in the classroom (Race to the Top Fund, 2011).

Individual states have also begun passing laws that ask for more accountability for teachers in the educational system. This accountability requirement is fulfilled, in part, by mandating that teachers be paid for their performance rather than by years of service and the qualifications obtained (Koedel & Betts, 2011), criteria that historically have been used in compensation formulas. A specific example is the State of Florida. Early in 2011 the State of Florida passed Senate Bill 736 (SB736), which stipulated that all teachers be paid for their performance in part by measures of their students' success (Senate Bill 0736, n.d.). This Bill further provided greater accountability for the educational system as a whole by including teachers in the measurement process.

Measures of teacher accountability are also present at the district level. One common measure that is part of the teacher accountability process involves the use of observational rubrics. Using these observational rubrics, administrators decide if the teachers are doing a good job in their teaching efforts and reward them accordingly. Although these observational measures are grounded in many years of empirical research (Danielson, 2011; Marzano, 2007) and have many benefits (e.g., observing what occurs in a classroom), as with all measurement approaches, this method also contains some limitations (Jacob & Lefgren, 2008; Murnane et al., 1991), which include potential bias from the observer/evaluator (e.g., initial impressions or personal opinions) (Strong, 2011). Further, the observer/evaluator may not be an expert in the topic or grade level being taught, thus limiting the understanding of what is being observed.

A benefit of using multiple measurement approaches is that usually not all methods have the same weaknesses. Because of the imperfections of an observational method of teacher evaluation, a push has developed to add new approaches to the evaluations of teachers. This new type of evaluation system falls under the label of Value-Added Modeling (VAM). Value-added models represent a variety of mathematical models that can differ in terms of the components of the model (e.g., presence or absence of covariates or control variables) or the assumptions and interpretations (e.g., the persistence of prior teacher effects on future outcomes) that can be made from them (Tekwe et al., 2004). These models use the results of students' test scores to mathematically estimate the effect a teacher has on the academic achievement of the teacher's students keeping in mind that different effects can be found using different subject areas (reading or mathematics). A VAM score for a teacher represents how much that teacher was able to add to students' knowledge while he or she instructed them. With the use of these scores, teachers can be ranked by how effective they were in producing student test scores that were higher than were predicted for them.

There is a strong momentum to add VAM scores to teachers' yearly reviews because some policy makers argue that rewarding teachers on their results will incentivize better performance (Hanushek, 2007; Schacter & Thum, 2004). Further, there is strong momentum to accurately understand the effect teachers have on their students. The proposal to add VAM scores into the evaluation process takes away some of the idiosyncrasies of principal administered observations by focusing the evaluation on measureable constructs.

The use of VAM in teacher evaluations seems to hold an advantage over observational methods of evaluation. A reason for the advantage is that VAMs tend to be an equalizer of several factors that may affect teachers that are out of their control. Examples of factors that could be equalized include any special needs of a student, or whether English is the student's native language. The goal of VAM is to avoid unfairly penalizing or rewarding teachers in their evaluations because of the characteristics of the students in their classroom. Equalization of these factors is done statistically and not through the interpretation of an administrator.

But, like other measures of accountability, VAM is not free of flaws. The most troubling is that research has found the reliability of the scores derived from the models to be less than ideal, possibly indicating that there is much error in the teacher VAM scores (Koedel & Betts, 2007; Lockwood, Louis, & McCaffrey, 2002; McCaffrey, Sass, Lockwood, & Mihaly, 2009). This puts in question the ability to replicate the scores and to have trust that the score produced are an accurate representation of the effectiveness of a teacher.

Because of their imperfections, teacher evaluations as accountability systems need to be evaluated as well as the individual pieces (VAM scores and scores from observational rubrics) to understand if the results produced are accurate representations of how teachers are performing. Since the purpose of a teacher accountability system is to be able to evaluate the performance of a teacher, if this system is not working properly, the results obtained from it may not be valid. Teacher evaluations are high stakes in the State of Florida (teachers will be retained or let go), and therefore the evaluations need to be an accurate reflection of teacher quality (Senate Bill 0736, n.d.).

The addition of VAM scores in teacher evaluations is new to the State of Florida and to date no validity evidence has been provided for them. The current research aimed to provide validity evidence of VAM scores of teachers in a Florida southeastern district by examining the relation of VAM scores to scores obtained from an observational method. In addition, this study aimed to examine how each of these measures of teacher quality (i.e., VAM scores and observational scores) related to other variables that were hypothesized to be related to quality teaching. Currently there is no "gold standard" for the evaluation of quality teaching, or even a clear definition of traits a quality teacher might possess. Since there is no perfect, or even universally accepted method for identification of quality teaching, inspection of the psychometric qualities of both the

VAM scores and the observational rubric scores is needed. Without inspection of both, even if a relationship is found, there would be no way to discern how meaningful this relation is because either or both measurement approaches could be flawed.

The southeastern district in the U.S. that was used in this study developed the teacher observational rubric to be administered by principals and assistant principals based on suggestions by industry standards (Danielson, 2006; Marzano, 2007). The rubric, which measures five constructs (Ability to Assess Instructional Needs, Plans and Delivers Instruction, Maintains a Student-Centered Learning Environment, Performs Professional Responsibilities, Engages in Continuous Improvement for Self and School), is based on teacher practices that have been empirically documented to enhance student learning. The rubric covers the areas of teacher planning, the environment in the classroom, the actual instruction, and other professional responsibilities a teacher may have (Danielson, 2007).

The value-added scores used in this study are considered by the State of Florida to be measures of students' academic achievement gains. The state contracted with an external company, the American Institute for Research (AIR), to develop the value-added model that produced the teacher scores derived from student achievement that were used in the present study. The model that was chosen, now called the Florida model, contains covariates and uses individual data, classroom data, and students' scores on the Florida Comprehensive Assessment Test (FCAT).

The Florida VAM scores were derived from an error-in-variable (i.e., x=t+e where a student's score is comprised of a true score and error) covariate adjustment model with 10 predictor variables (Value-Added Model White Paper, n.d.). The variables that were included in the model per the Value-Added Model White Paper (n.d.) can be seen in Table 1.

Because of the high-stakes decisions that are made from the use of the VAM scores and teacher evaluations as a whole, evidence to support the validity of the model and the scores derived from it is imperative. As stated in the *Standards for Educational and Psychological Testing*, the term validity “refers to the degree to which evidence and theory support the interpretations of the test” (American Educational Research Association, American Psychological Association, and National Council on Measurement in Education, 1999, p. 9). Further, they state that a “sound validity argument integrates various strands of evidence into a coherent account of the degree to which existing evidence and theory support the intended interpretation of test scores for a specific use” (American Educational Research Association, American Psychological Association, and National Council on Measurement in Education, 1999, p. 17). Validity evidence of VAM scores could reveal the appropriateness of their use for high-stakes decisions.

There are several types of validity evidence that can be gathered to support the meaningfulness of VAM scores. This evidence includes correlations with other measures of teacher quality, such as those based on observational rubrics (i.e., convergent validity) and correlations with other relevant variables, as defined by a nomological network of teacher quality. The *Standards for Educational and Psychological Testing* (1999) served as this benchmark throughout the study (American Educational Research Association, American Psychological Association, and National Council on Measurement in

Table 1

List of Covariates in Florida Value-Added Formula

Name of Covariate
• The number of subject-relevant courses in which the student is enrolled
• Two prior years of achievement scores
• Students with Disabilities (SWD) status
• English language learner (ELL) status
• Gifted status
• Attendance
• Mobility (number of transitions)
• Difference from modal age in grade (as an indicator of retention)
• Class size
• Homogeneity of entering test scores in the class

Education). Based on the standards, test scores used for a new purpose must be validated (Standard 1.4); evidence of the internal structure of the test must be explored (Standard 1.11); reliability and standard errors should be presented for every score and subscore (Standard 2.1); and if subjective judgment is present in the scoring, evidence of inter-rater reliability needs to be provided and sources of error (Standard 2.10 and Standard 14.5) need to be examined.

Problem Statement

Though much research has been conducted on value-added models and how well they function, currently, there is scarce research providing validity evidence of VAM scores in relation to other variables, including scores from an observational rubric. Research designed to examine the relationship between VAM scores and the ratings given by the teachers' principals is in high demand (Amrein-Beardsley, 2008; Braun, 2004; Harris & Hill, 2009; Hill, Kapitula, & Umland, 2011; Kupermintz, 2003; McCaffrey et al., 2004a; Meyer, 1997; Rubin, Stuart, & Zanutto, 2004). Part of the demand arises out of the perceived lack of connection between theory and empirical evidence (Harris & Rutledge, 2010) and another part from the need for demonstrated validity evidence prior to using VAM scores for high-stakes decision-making (Kupermintz, 2003). Research on how value-added scores relate to accepted empirical evidence of effective teaching is needed to provide evidence to support or question the use of value-added scores in teacher evaluations, especially for high-stakes decisions.

Purpose of the Study

The purpose of this study was to examine how value-added scores relate to accepted empirical evidence of effective teaching in order to provide evidence to support or question the use of value-added scores in teacher evaluations. This study examined the validity of the Florida VAM scores and how they relate to the district's observational rubric. In addition, this study examined how VAM scores and scores from the observational rubric related to other established measures of teacher quality. Some of the measures of teacher quality that have been found in the literature to impact student performance include possession of a National Board Certification and years of experience (Murnane & Phillips, 1981b; Rockoff, 2004; Strong, 2011). Since research demonstrates that the impact of years of experience may peak somewhere between three and 10 years, linear and nonlinear (i.e., quadratic) relations between teachers' years of experience and VAM and observational scores were examined (Murnane & Phillips, 1981b; Rockoff, 2004; Strong, 2011). Chapter Two summarizes some of the literature for these variables and their hypothesized relationship to student achievement.

Research Questions

The following research questions were examined:

All of these questions are answered with a sample of teachers from a large southeastern school district.

1a) To what extent are the observational data used to evaluate teachers during the 2011-2012 school year consistent with the five-factor measurement model (Ability to Assess Instructional Needs, Plans and Delivers Instruction, Maintains a Student-Centered Learning Environment, Performs Professional Responsibilities, Engages in Continuous Improvement for Self and School) underlying the observational rubric?

1b) For the observational rubric, what is the estimated internal consistency reliability of the scores for the five factors (Ability to Assess Instructional Needs, Plans and Delivers Instruction, Maintains a Student-Centered Learning Environment, Performs Professional Responsibilities, Engages in Continuous Improvement for Self and School) collected through observations obtained during the 2011-2012 school year?

2) Do administrators' observational ratings of teachers based on the rubric correlate with teachers' value-added scores from the Florida VAM within the 2011-2012 school-year?

3) Do the teachers' VAM scores for the 2011-2012 school year and the scores from the observational rubric relate to other theoretically relevant teacher variables (e.g., National Board Certification, years of experience) and not to theoretically unrelated variables (e.g., gender, race and ethnicity)?

Significance of the Study

This study provided several sources of evidence of validity for VAM scores. These sources of evidence included comparing VAM scores to the teacher observational rubric meant to explicate quality teachers, and variables that are correlates of quality teaching. The results provided initial evidence of the relationship between VAM scores and the aforementioned variables. In addition, this study provided evidence of the factorial validity of the five-factor measurement model underlying the observational rubric (Ability to Assess Instructional Needs, Plans and Delivers Instruction, Maintains a Student-Centered Learning Environment, Performs Professional Responsibilities, Engages in Continuous Improvement for Self and School) used in the validation process for the VAM scores.

Limitations of the Study

This study was based on a teacher sample from one school district only in Florida. Because of the nature of VAM scores being calculated at the State level (not district level) and the fact that each district has the ability to choose the components that make up the observational rubric, the results would not be generalizable to different districts with different observational methods.

Further, this study was limited to the VAM model already in place in the State of Florida and does not provide evidence of the appropriateness of the model that was developed or the predictor variables that were chosen to be a part of the model. Validity evidence provided in this study relies solely on the scores as they were delivered to the large southeastern school district in Florida, without any modifications to the scores.

Lastly, this study relied on the teacher VAM scores from the Florida model as developed by AIR for the 2011-2012 school year. Any future modifications to the model itself may not create the same scores and may also change the score value each individual teacher receives. A change in value-added scores from year to year or through the use of a different value-added model might reveal different results of validity evidence

Definition of Terms

Confirmatory Factor Analysis: inspects the correlations among a set of variables using a relatively small number of underlying factors with the factor structures specified in advance (Brennan, 2006).

Nomological Network: can be viewed as an "interlocking system of laws which constitute a theory" (Cronbach & Meehl, 1955, p. 290). The nomological network aims to look at the relationships between constructs as specified by some theory.

Observational Rubrics: a common evaluation measure where administrators use a set of indicators to rate teacher classroom performance.

Structural Equation Modeling: a statistical method to inspect the relationships of constructs that are part of a conceptual or theoretical framework (Benson, 1998; Benson & Hagtvet, 1996; Brennan, 2006; Graham, 2008; McDonald, 1999).

Validity: According to the *Standards for Educational and Psychological Testing*, the term validity "refers to the degree to which evidence and theory support the interpretations of the test" (American Educational Research Association, American Psychological Association, and National Council on Measurement in Education, 1999, p. 9).

Value-Added Models for Teachers: statistical models for the evaluation of teachers representing the contribution in a given year teachers make on their students by comparing current school year test scores of their students to the scores of those same students in the previous school year, as well as to the scores of other students in the same grade.

Chapter Two: Literature Review

The purpose of this study is to examine how value-added scores relate to accepted empirical evidence of effective teaching, in order to provide validity evidence to support or question the use of value-added scores in teacher evaluations. This review of literature addresses teacher quality including definitions and the difference between quality and effectiveness. A review of predictors of teacher quality and research findings regarding the effect of teacher quality on student achievement is provided. The statistical foundation underlying value-added models along with the history, types of models, the Florida model, and the problems and benefits of these models are discussed. Teacher observational methods and their role in the teacher evaluation process are discussed. Lastly, the *Standards for Educational and Psychological Testing* is used as a framework for examining the measurement issues that underlie the teacher observational and value-added scores.

Teacher Quality

A substantial body of research has established that teachers are a valuable component to student success, and better teachers produce better results from their students (Aaronson, Barrow, & Sander, 2007; Goldhaber & Anthony, 2003; Goldhaber, Brewer, & Anderson, 1999; Goldhaber, & Theobald, 2011; Rivkin, Hanushek, & Kain, 2005; Rockoff, 2004). This means that teachers who are better at their job will have better outcomes from the students that they teach. Because of this knowledge, finding out what makes an effective teacher is crucial to the development of the profession.

The key is identifying what qualities make a teacher better. At this time there is no clear definition, or gold standard, for the qualities a teacher must have to make them quality teachers. There are, though, many assumptions and research on characteristics that may make teachers better in their profession. The initial step in identifying these characteristics includes defining the difference between quality and effectiveness.

Quality/Effectiveness. The terms quality and effectiveness are casually used in the description of a teacher. General understanding, though ambiguous, is that quality and effectiveness are both desired from a teacher. The terms are made even more ambiguous by being described by different terms such as expert teacher, highly qualified teacher, or even a master teacher.

In the literature, quality can be described, depending on the authors' point of view, as characteristics teachers may possess, qualifications they have earned, methods of teaching, or even the results obtained from students (Berliner, 2005; Competencies for Teachers, n.d.; Darling-Hammond, 1997; Kelly, 2012; Strong, 2011). Effectiveness is a part of quality teaching, but it relates to the outcomes achieved by students (Berliner, 1987; Strong, 2011). The understanding of this difference is crucial because value-added models are examples of measures of teacher effectiveness that are based on student outcomes, which in turn are also a part of quality teaching. This review will cover aspects that represent quality teaching, including teacher effectiveness as operationalized using the scores from value-added models.

Research on quality teachers. With the known connection between student achievement and teaching, much research has been conducted on characteristics of teachers and the perceived effect on student outcomes. The following is a review of research on teacher factors that have been examined in relation to student achievement. The variables considered are the most commonly studied.

Teacher education. The educational degree a teacher holds is thought to be a quality trait leading to higher student achievement. It is perceived that if teachers spend time and effort earning a higher degree (e.g., master's degree), they would be more engaged in their profession and in turn, more engaged with their students. Further, it has been common practice for districts to pay teachers more for a higher educational degree.

Research has found that teacher qualifications are weak predictors of student achievement (Berger & Toma, 1994; Borland & Howsen, 1992; Card & Krueger, 1992; Ehrenberg & Brewer, 1994; Goldhaber & Brewer, 2000; Hanushek 1986, 1992, 1997; Harnisch, 1987; Harris & Saas, 2009; Miller, McKenna, & McKenna, 1996; Montmarquette & Mahseredjian, 1989). This variable was found in research to have mixed effects, or insignificant positive or negative effects on student achievement. These inconsistent results have been replicated over the years in numerous studies.

Teacher salary. A variable that is commonly researched for its connection to student achievement is the amount of money teachers are compensated for the work they do. This variable has produced mixed results in research as it relates to student achievement. Many empirical studies found a positive effect of teacher salary on student achievement (Butler & McNertney, 1991; Card & Krueger, 1992; Dolan & Schmidt, 1987; Hanushek, 1997; Sanders, 1993; Stern, 1989). These studies indicated that the higher the teacher's salary, the higher the scores on student assessments. Research for this question was conducted across several geographic areas and over several decades.

The findings of all studies are not homogenous in regard to the effect of teacher salary on student achievement. Other studies have found a negative effect between teachers' salaries and student achievement (Borland & Howsen, 1992; Kurth, 1987). This inverse relationship was explained by the authors of the research as a potential ceiling effect on salary. Regardless of the positive or negative finding of the research studies, all authors mentioned that higher salaries usually imply more years in teaching and thus more experience. The number of years of experience a teacher has is also an important variable that has much research.

Years of experience. The longer a person remains at the same employment, the more time he or she has to master the skills involved. Research studies have found a positive relationship between years of experience of a teacher and student achievement (Bosshardt & Watts, 1990; Card & Krueger, 1992; Ehrenberg & Brewer, 1994; Grimes & Register, 1990; Hanushek, 1992, 1997; Montmarquette & Mahseredjian, 1989; Murnane & Phillips, 1981ab). Because of these positive finding there is reason to believe teachers' years of experience could affect how well they perform their job duties (Harris & Rutledge, 2010). These findings stress the fact that the longer teachers remain as teachers, the more effective they become, and in turn the better the results they obtain from the students in their classroom.

Though positive effects of years of experience on student achievement were found in almost all studies, there also appears to be an indication that there is a learning curve to becoming an effective teacher. This learning period might take several years (Murnane, Willett, & Levy, 1995). This learning curve might further be an indication of the positive relationship between teachers' experience and student achievement, yet this effect tends to attenuate at a certain point in the teacher's career.

Personal characteristics *(Race/Ethnicity and Gender)*. Evidence for or against having a teacher from the same ethnic background as his or her students is limited and the effects may be more indirect in that a student can see a role model, which may then affect student achievement (Strong, 2011). Studies have suggested that having teachers of the same ethnic background as their students can have a positive effect on student achievement, though only in certain subjects (Dee, 2004; Hanushek, 1971). In general results of these studies have been mixed (Ferguson, 1998). Further, these studies only inspected the relationships between White and African American students and teachers, without much inspection of other races.

The role a teacher's gender has on student educational outcomes has also been investigated. Though not much research has been conducted, studies have found a slightly positive to no relationship between the teacher-student match on gender and how successful the student is in completing his or her schooling career (Dee, 2004, 2005; Ehrenberg & Brewer, 1994; Nixon & Robinson, 1999). Overall, these teacher characteristics seem to have little effect on student achievement.

National Board Certification

A certificate can be obtained from the National Board for Professional Teaching Standards that designates a teacher as National Board Certified (NBC). This certification can be acquired as a supplement to state requirements and identifies teachers knowledgeable in their content area, and able teachers in K-12th grades (National Board, 2013). This certification lasts for 10 years at which time renewal of the application is needed.

This certification can be procured through a rigorous process that demonstrates an individual's teaching practice through assessments and portfolios (National Board, 2013). The possession of this designation attests to the teacher's leadership skills and ability to enhance students' education, and results in an increase in the teacher's salary (National Board, 2013).

Much research has been conducted on the relationship between teachers who hold this designation and student achievement. Large studies have found a positive relationship between teachers who are NBC and student achievement. This means that students of teachers who have achieved NBC certification have higher outcomes on standardized assessments than students of other teachers at the elementary levels (Card & Krueger, 1996; Goldhaber & Anthony, 2007; Vandevoort, Amrein-Beardsley, & Berliner, 2004). These achievement level differences were not always statistically significant.

Other studies have looked into what having this designation actually means. Several studies have understood this certification to imply a more effective teacher (Cavalluzzo, 2004; Sato, Chung, & Darling-Hammond, 2008; Smith, Gordon, Colby, &

Wang, 2005; Vandevoort, Amrein-Beardsley, & Berliner, 2004). A more effective teacher is one who can obtain better results from his or her students in regards to achievement.

As demonstrated by these studies, there seems to be an important effect of possessing certification from the National Board and student achievement. This relationship appears to be a positive effect. Other variables, such as value-added scores, which are also meant to measure teacher effectiveness, should have a positive relationship with this certification. Teachers who obtain a NBC should have a higher VAM score than other teachers.

Value-Added Modeling

Growth modeling has become an increasingly popular tool in the educational setting because it aims to predict whether a student has progressed academically with the use of previous years' data. Value-added modeling, specifically, is now used in many districts and states throughout the U.S. as a measure of student growth. The popularity of VAM has arisen from the ability of these measures to look at students' growth over time as opposed to simply seeing a single data point in a student's career (Schaeffer, 2004). VAM informs not just if a student was proficient in a subject, but further provides information about the degree of proficiency. The increase in the amount of information that can be determined by a student's test scores over time has led to advancements of VAM use for teacher accountability models.

Value-added models are normative in nature. The State of Florida uses all the teachers in the state to create these scores. Teachers who teach the courses listed in

Appendix A are included in this pool. Individual districts will have teachers who fall somewhere in the distribution of scores, made up of all teachers in the state.

There are many reasons that the focus has moved toward the use of VAM in teacher accountability models. According to Hanushek and Rivkin (2010), research supports that this measure can quantify the differences in effectiveness of teachers, even of teachers within the same schools. This tool can assist in properly identifying teachers with regard to their ability to have students make learning gains.

This section will provide information regarding the history of value-added models, the different types of models, and the advantages and disadvantages of using value-added models for rating teachers. The last part of this section explains the Florida value-added model and includes how it was developed and the predictor variables in the model.

History. The history of VAM loosely begins in the 1840s in the U.S. when the city of Boston implemented an assessment to rate the academic differences amongst a large group of students, between classrooms and different schools (Resnick, 1982). This preliminary step to modern VAM methods was intended to observe and compare the differences between students in different school settings, thus stressing the importance of measurement to understand students and inform decisions.

In the 1960s, with the Soviet Union's ability to launch a rocket into outer space (Sputnik), the U.S. began several efforts to ensure that students were being held accountable including the beginning of the National Assessment of Educational Progress (NAEP) (Glaser & Silver, 1994). The NAEP assessment allowed for students' progress to be measured at certain intervals in time. This allowed the country to examine and keep track of student growth.

Another initiative implemented as a result of Sputnik was the Equality of Educational Opportunity Survey, which culminated in the Coleman Report (Glaser & Silver, 1994). This report found that there were large variations in achievement levels across the country (Coleman et al., 1966). Because there was now a clear finding that not every student had the same knowledge upon graduation, more actions were taken.

Because there was a belief that something was wrong with the U.S. educational system, a report was initiated to examine the type of education students were receiving (Gardner, Larsen, & Baker, 1983). This report provided the foundations of what courses students in high school needed to take; asked high schools and universities to be more rigorous; and asked for changes in teachers' salaries and work contracts (Gardner, Larsen, & Baker, 1983). All of these changes were meant to bring more accountability to the educational system as a whole, and to the teachers who were a part of this system.

In 1994 Goals 2000, which was made law by President Clinton, attempted to have states develop standards and create assessments to test student knowledge on those standards (Superfine, 2005). This program was not successful for multiple reasons. It was followed by the No Child Left Behind Act (NCLB).

The next notable action that focused attention towards testing was the passing of the No Child Left Behind (NCLB) Act in 2001, and implemented in 2002, which demanded accountability of teachers in the classroom (Public Law 107-110) (U.S. Department of Education). This act refocused the nation's attention towards testing and it further placed emphasis on tying teacher performance expectations to student scores on assessments.

Though for years states have been looking at students' achievement by assessing whether they reach a certain level of proficiency, such as Adequate Yearly Progress (AYP), this method is not ideal as it groups student performance into broad categories (Koretz, 2003). Simply stated, by not keeping all of the information from a particular score a child may have received on an assessment, it is impossible to determine the actual amount of proficiency, and the only thing that can be determined is if proficiency was observed. For this reason, attempts were made to develop measures for use in accountability that would maintain as much information from the test scores as possible.

One method currently in place that can be used for accountability purposes and which uses information of students' scores over time (as opposed to a snapshot in time) involves the use of value-added models. Since research has demonstrated that teachers do in fact have an effect on the students they teach, value-added models have been introduced as a way to estimate the effect a teacher has on academic achievement of a student (Hill, Kapitula, & Umland, 2011). These statistical methods provide individual teachers with a score that takes into account several predictor variables, and which include current and previous test scores of the students in their class. This VAM score can then be used to compare teachers based on their levels of student effectiveness, and be used in pay-for-performance plans.

Different Types. There are several types of value-added models currently in existence. One of the reasons for the several models is that teachers and students change over time, thus a simple hierarchical linear model would not be adequate to understand the effects of a teacher on students (e.g., McCaffrey et al., 2004b; Raudenbush & Bryk, 2002). Thus many attempts have been made to identify the most effective method to measure the effect of teachers on student achievement. VAMs can be different in several ways including the model itself as well as the statistical assumptions underlying the models (Tekwe et al., 2004).

Three main types of value-added models include the covariate adjustment model, the one year gains model, and the cross-classification model (Raudenbush & Bryk, 2002). Briefly, the covariate adjustment model uses scores from previous years and includes covariates (predictor variables); the one year gains model subtracts the current year score from the previous year's score and still includes covariates; the complex cross-classified model uses random effects with the outcome differences being test scores or test score gains (McCaffrey et al., 2004b; Rubin, Stuart, & Zanutto, 2004). These different models are currently in place for several pay-for-performance plans across the United States.

For example, the Tennessee value-added model monitors the gains that students make through time on state assessments but does not include demographic predictor variables (i.e., covariates) (Sanders et al., 2002). On the other hand, the Florida model includes many predictor variables. Each state has the autonomy to decide the model that best suits its needs. Yet, even if states chose the same type of value-added model to use for the calculation of teacher effect on students, each state or district has the liberty to make individual modifications to the model.

Advantages. Growth modeling is now considered a better model for inspecting true differences between teachers and schools than the previously established methods, such as AYP (Linn, 2006; Meyer, 2000; Raudenbush, 2004). Research has demonstrated that value-added modeling can be a meaningful measure of teacher effects on student achievement (Jacob & Lefgren, 2005; Kane & Staiger, 2008). For these reasons, there is an increase in the use of value-added modeling for pay-for performance plans.

One of the main advantages of using value-added modeling is that it tends to be an equalizer of several factors that affect teachers and are out of their control, in turn reducing systematic error (Harris, 2011). For instance, teachers will not be penalized or rewarded unfairly for the individual characteristics of the students they teach (Ballou, 2002; McCaffrey et al., 2004b). For these reasons, scores from value-added models make it possible to compare teachers who have students who differ on demographics, socio-economic status, or abilities.

Growth models further have the ability to take into account the differences that existed prior to the current years test score (Linn, 2008). VAMs rely on several predictor variables that are measured over time, as opposed to a single measure, thus increasing the possibility of identifying a trend (Amrein-Beardsley, 2008). This in turn ensures that the scores measure student gains and make it fairer for teachers and schools.

Disadvantages. A primary disadvantage to using value-added modeling is related to the lack of transparency of the models used for pay-for-performance. Because of proprietary information, the models have generally not been open for peer review (Amrein-Beardsley, 2008; Kupermintz, 2003). Consequentially, it is impossible to obtain the opinions of experts from across the country with regard to the models or for the statistical community to provide suggestions for improvement.

Another disadvantage of using value-added models for pay-for-performance plans is that research on existing models has found causes for concern in using these models. Reliabilities of the scores derived from these models have been modest to low (Koedel & Betts, 2007; Lockwood, Louis, & McCaffrey, 2002; McCaffrey, Sass, Lockwood, & Mihaly, 2009). This fact is not reassuring when the possibility of using these scores for continual employment exists.

Another problematic aspect of scores from a value-added model is that research has found different results depending on the assessment used in the model (Lockwood et al., 2007; Papay, 2011). Since different assessments are used to calculate value-added scores in different states, the same teacher could potentially receive a better score in Florida than in Tennessee, and vice-versa. This is not desirable because the models are supposed to be stable enough to detect teacher effects regardless of external conditions.

Several studies have also compared value-added outcome scores to teacher evaluations completed by principals. The correlations of those scores have been low to moderate (Gallagher, 2004; Kimball, White, & Milanowski, 2004; Milanowski, 2004).

Milanowski (2004) compared VAM scores for teachers in Cincinnati to the Cincinnati teacher evaluation rubric (Teacher *N*=212) for reading, mathematics, and science (teachers were analyzed in multiple categories). A composite score based on four domains from the observational rubric was used in this study. This study used about 66% of the students who qualified for analyses in the computation of VAM scores as extreme student scores, based on the scale score of the state and district assessments, were removed from the sample. Results were presented by grade and by subject with correlations in reading from grades 3 to 8 ranging from .03 to .45, mathematics from .20 to .56, and science from -.01 to .33. Results combined over grade level produced correlations in reading of .32 (95% confidence interval = .18 to .45), mathematics of .43 (95% confidence interval= .29 to .55), and science of .27 (95% confidence interval = .09 to .46) (Milanowski, 2004).

Kimball, White, and Milanowski (2004) inspected the relationship between VAM scores and scores from an observational rubric, based on the work of Charlotte Danielson, in a county in Nevada. Analysis was based on 328 teachers (123 teaching 3rd grade, 87 teaching 4th grade, and 118 teaching fifth grade) (Kimball, White, & Milanowski, 2004). The empirical Bayes estimates resulting from the VAM were then correlated with the observational rubric in the district. The resulting correlations were very weak to weak (3rd grade reading and mathematics, *r*=.10; 4th grade reading, *r*=.28; 4th grade mathematics, *r*=.07; 5th grade reading, *r*=.28; 5th grade mathematics *r*=.37) (Kimball, White, & Milanowski, 2004).

Another study by Gallagher (2004) inspected the relationship between VAM scores and teacher evaluation scores based on an observational rubric. One Los Angeles elementary school was chosen for this research and based on 34 5th grade teachers the correlations between the VAM scores were low to moderate by subject (reading *r*=.50; mathematics *r*=.21; language arts *r*=.18; composite *r*=.36) (Gallagher, 2004). Thus, this study represents another research study that found relatively weak (and one moderate) correlations between the VAM scores and scores from an observational rubric. Sample size was very small for this study.

In general, though there are several positive aspects about value-added modeling there also several drawbacks to using the models. All research inspected suggests caution when using value-added modeling for high-stakes decision-making; several researchers have noted that VAM scores should not be used in isolation but should be one part of a comprehensive evaluation of teachers' performance. These previous results underscore the need for validity studies on these measures.

Florida value-added model. The State of Florida has attempted for many years to pay teachers based on their performance. The first attempts occurred during the 1990s and 2000s but the results of the attempts obtained mixed reviews at best (Hill, Kapitula, & Umland, 2011). Efforts to create a method to pay teachers based on their effects on student achievement were not a top priority for several years given previous results. Race to the Top funds have made the State of Florida again invested in creating a pay for performance plan that can be appropriately implemented.

In the State of Florida, the resulting scores from value-added models are derived in part from student scores on the Florida Comprehensive Achievement Test (FCAT). Since the results of this assessment are a large component of the covariates in the Florida value-added model, an understanding of the standardized statewide test is essential to understanding the Florida model.

FCAT. As stated in the Florida Department of Education website, the FCAT began its implementation in 1998 (Florida Department of Education, n.d. a). The FCAT is a “criterion-referenced test in mathematics, reading, science, and writing, which measure student progress toward meeting the Sunshine State Standards (SSS) benchmarks” (Florida Department of Education, n.d. a). The test was constructed using rigorous industry accepted standards and has been equated, from year to year, taking into account grade level differences.

The FCAT student results are presented in Developmental Scale Scores (DSS). This form of score, which ranges from 0-3000, was developed to “track student progress over time and across grade levels to indicate student ‘growth,’ or ‘learning gains’ (Florida Department of Education, n.d. b, para. 25). The school year 2010-2011 was the last year that the FCAT was used for testing purposes continuously through the tenth grade. The State of Florida is now moving towards end of course exams (EOC’s), which will replace portions of the FCAT (Ash, n.d.) and future years VAM scores will be developed from these measures.

Development of the model. To determine teacher value-added scores, the State of Florida contracted with an external company, The American Institute for Research (AIR). Because of proprietary reasons, there is only limited information on the actual model this company has created. Though there is insufficient information regarding the details of the model, there is a plethora of information regarding how the model was constructed.

The American Institute for Research cooperated with a committee made up of community stakeholders to design and implement the model for the State of Florida. The committee, called the Student Growth Implementation Committee (SGIC), working closely with AIR, made a recommendation for a covariate adjustment model with eight predictor variables that was accepted by the State of Florida (Value-added model White Paper, n.d.). The covariate model uses scores from the current year test as the outcome variable while prior year test scores and other variables are used as covariates; the model treats teachers and schools as coming from a distribution of random effects (American Institute for Research, n.d.).

The final model is a hierarchical linear model with separate levels for the variation between schools, the variation between teachers within a particular school, and the variation between students in a particular classroom, all computed as orthogonal (uncorrelated) components (American Institute for Research, n.d.). Calculations are done using data from the entire state, not district by district, and therefore, differentiations between the statewide expectation and specific school differentiations (which could be explained by better leadership or assignment of students and teachers) are calculated and become the school component of the equation (American Institute for Research, n.d.). The final score for a teacher is then made up of the particular teacher score adding in half of the school component. The model, in general form, can be found in Equation 1.

$$y_{ti} = \mathbf{X}_i\boldsymbol{\beta} + \sum_{r=1}^{L} y_{t-r,i}\gamma_{t-r} + \sum_{q=1}^{Q} \mathbf{Z}_{qi}\boldsymbol{\theta}_q + e_i \qquad (1)$$

According to the Florida Value-Added Technical Report

> y_{ti} is the observed score at time *t* for student *i*, $\mathbf{X}_i$ is the model matrix for the student and school level demographic variables, $\boldsymbol{\beta}$ is a vector of coefficients capturing the effect of any demographics included in the model, $y_{t-r,i}$ is the observed lag score at time *t-r* ($r \in \{1,2,\dots,L\}$), $\boldsymbol{\gamma}$ is the coefficient vector capturing the effects of lagged scores, $\mathbf{Z}_{qi}$ is a design matrix with one column for

> each unit in q ($q \in \{1,2,\ldots,Q\}$) and one row for each student record in the database. The entries in the matrix indicate the association between the test represented in the row and the unit (e.g., school, teacher) represented in the column. We often concatenate the sub-matrices such that $\mathbf{Z} = \{\mathbf{Z}_1,\ldots,\mathbf{Z}_Q\}$. $\boldsymbol{\theta}_q$ is the vector of effects for the units within a level. For example, it might be the vector of school or teacher effects which may be estimated as random or fixed effects. When the vector of effects is treated as random, then we assume $\boldsymbol{\theta}_q \sim N(0, \sigma^2_{\boldsymbol{\theta}_q})$ for each level of q. (American Institute for Research, n.d., p. 6)

From the formula the teacher effects can be derived, which are the residual variations at the teacher level once the student and school factors are separated. As previously stated, student expectations (how they are predicted to perform) are calculated in relation to a comparison to other students with similar characteristics and prior test scores. The difference between what is expected of the student, and how the student actually performed is called the residual, and those residuals are then aggregated by teacher using empirical Bayes estimation to calculate the teacher effect (American Institute for Research, n.d.). The formula for the aggregate teacher effect estimates ($\tilde{\Theta}_{\boldsymbol{j}}$ = aggregate for teacher j) can be seen in equation two, "where σ_t^2 is the teacher level variance, σ_s^2 is the school level variance, σ_e^2 is the residual variance, N_j denotes the number of students in class *j* and the notation *(j)i* is used to mean that student *i* in class *j*" (American Institute for Research, n.d., p. 7).

$$\tilde{\Theta}_{\boldsymbol{j}} = \frac{N_j\sigma_t^2}{N_j(\sigma_s^2+\sigma_t^2)+\sigma_e^2}\frac{\sum_{i=1}^{N_j} r_{(j)i}}{N_j} \qquad (2)$$

The variables that were included in the model according to the Value-Added Model White Paper (n.d.) are: the number of subject-relevant courses in which the student is enrolled; two prior years of achievement scores; Students with Disabilities (SWD) status; English language learner (ELL) status; gifted status; attendance; mobility (number of transitions); difference from modal age in grade (as an indicator of retention); class size; and homogeneity of entering test scores in the class. According to SB736, the use of gender, race/ethnicity, and socioeconomic status could not be used as covariates in the value-added model (Senate Bill 0736, n.d.). A table with explanations of these covariates can be seen in Appendix B. These variables were considered to be the most important aspects of teaching in need of statistical control.

The SGIC not only decided what covariates to include, but also made business rules to be used while processing the data. "Business rules consist of decisions about student attribution to teachers, how duplicate or missing data are managed, how growth expectations for students taking multiple courses or having multiple teachers are determined, etc." (Value-added model White Paper, n.d., p. 5). The same document also states that more specific details for these business rules would be provided in the Technical Report, however, review of said report (American Institute for Research, n.d.) revealed that it does not address the business rules.

The final model is considered an error-in-variable (i.e., x=t+e where a student's score is comprised of a true score and error) covariate model (McCaffrey et al., 2004b). In order to account for higher errors at the extremes of the conditional standard errors of measurement (CSEM), and because there is heteroscedasticity in the error term, the error-in-variable regression model was chosen by the committee as the most appropriate way to derive the VAM scores using empirical Bayes estimation (American Institute for Research, n.d.).

Ultimately, a "teacher's value-added score reflects the average amount of learning growth of the teacher's students above or below the expected learning growth of similar students in the state, using the variables accounted for in the model" (Value-Added Model White Paper, n.d., p. 2). This model further includes past test scores of students in order to properly calculate their expected gains. The resulting scores can then be used to compare teachers to one another.

Though the Value-Added Model White Paper (n.d.) states that the technical manual will include all information necessary to replicate the model, the presenters of the model at the state conference held in Orlando on August 1 and 2, 2011, constantly reminded the public that replication was impossible at the district level because they had used the entire state data to calculate the VAM scores (Webcast, 2011). Scores could be replicated if scores from every district in the state were available and AIR explained that any change in an individual teacher's population of students would create a change in every teachers' scores. Insufficient time has passed for research and reports to be available on the Florida VAM. For this reason it is imperative that a validity analysis be conducted to better understand the scores that come from this model.

Observational Methods

Currently there is extensive research and literature on methods to evaluate teachers through observation. This literature can be divided into two categories:

administrator decisions and observational rubrics based on specific standards. Specifically, this research is based on practices that effective teachers employ in the classroom to increase student achievement.

Research on administrative review of teachers, specifically by principals, has demonstrated the benefits and flaws of this type of evaluation and how scores from these observations relate to student educational achievement (Anderson, 1954; Armor et al., 1976; Brookover, 1945; Gotham, 1945; Hill, 1921; Jacob & Lefgren, 2005, 2008; Manatt & Daniels, 1990; Medley & Coker, 1987; Wilkerson et al., 2000). Studies have found that principals are capable of identifying highly effective and highly ineffective teachers, but are not as adept at identifying the average teacher (Jacob & Lefgren, 2008). Further, teachers have complained about their lack of understanding the reasons why principals assign bonuses to some teachers and not others (Murnane et al., 1991).

Another large body of research involved classroom observations utilizing frameworks that are meant to depict actions and activities effective teachers should engage in (Gallagher, 2004; Holtzapple, 2003; Kimball, White, & Milanowski, 2004; Milanowski, 2004; Schacter & Thum, 2004). This can also be referred to as standards based evaluations as the frameworks are composed of standards. Standards have been developed and compiled by organizations such as the National Board for Professional Teaching Standards, The Bill and Melinda Gates Foundation, and the Interstate New Teacher Assessment and Support Consortium, and these standards incorporate classroom evidence into teacher evaluations (Darling-Hammond et al., 2012).

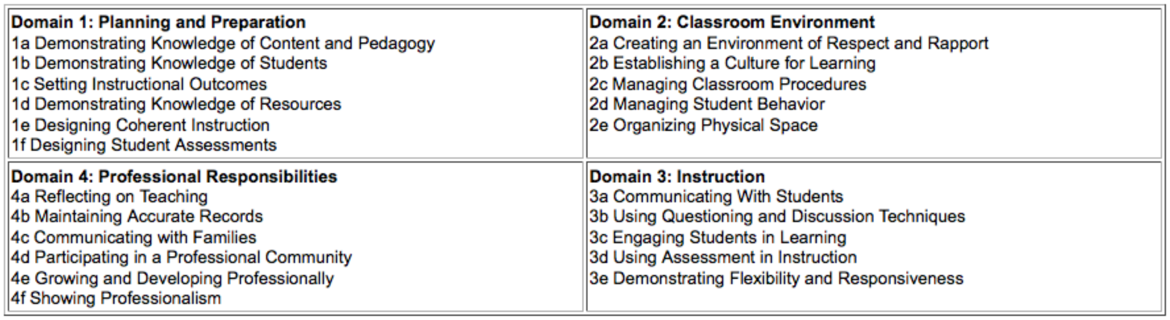
Local school districts have the choice of what framework, or combination of frameworks to use to develop their observation rubrics. Most recently, two experts in the field of education have emerged with frameworks that are gaining popularity amongst local education agencies. Danielson developed a framework for teaching that encompasses aspects such as planning and preparation, demonstrating knowledge of students, designing coherent instruction and managing student behavior (2007). Marzano (2007) has presented a slightly different framework that encompasses aspects such as using effective instructional strategies, using effective management strategies and using effective classroom curriculum design strategies. These are just but a few examples of the types of observational evaluations currently in existence.

Danielson's framework, specifically, has become increasingly integrated into educational systems. Specifically, it is the approved model for Arkansas, Delaware, Idaho, Illinois, New Jersey, New York City, and South Dakota. The framework also has much exposure in the State of Florida as it is being used by a large number of districts (Baker, Bay, Escambia, Hernando, Highlands, Hillsborough, Lee, Levy, Madison, Marion, Monroe, Okaloosa, Pinellas, Polk, Sumter, just to name a few) (Approved District Performance Evaluation Systems, n.d.).

The framework was originally developed and published in 1996 based on research compiled by Educational Testing Service (ETS) for use in a classroom assessment for licensing (called the PRAXIS), and included the skills needed by teachers (Danielson, 2011). The framework's development trajectory has been research-based but the most important recent changes involve research from the Bill and Melinda Gates Measuring Effective Teachers (MET) project, which while not changing the form of the rubric (4 domains and 22 components), did create additional resources aimed at providing clarity to each of the parts of the rubric (Danielson, 2011). The rubric domains and the components can be seen in Figure 1 in *The framework for teaching* (n.d.).

There are several issues that must be addressed concerning potential sources of error for observational rubrics, specifically, the human component. The cognitive load required for observation can reduce the validity and reliability of the data collected. Some of the major sources of systematic error that can occur during an observation caused by the observer(s) include the error of leniency, the error of central tendency, and the halo effect (Gall, Borg, & Gall, 1996). These errors can change the score a person should receive because the observer marks too highly, marks most scores around the middle point, or is influenced by early impressions of an individual's performance (Gall, Borg, & Gall, 1996). For this reason, continuous training of the observers is as important as a well-developed rubric.

The reliability of observational scores also is influenced by the number of times teachers are observed. When there is substantial day-to-day variation in teacher classroom performance there is a need to have more observations to obtain acceptable levels of score reliability. Hill, Charalambos, and Kraft (2012), for example, found that even with two observers on three occasions, the reliability of scores from the Mathematical Quality of Instruction (MQI) observational assessment was only .77, .71, and .81 on the MQI subscales.



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: 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 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	**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

Figure 1. Danielson Framework for teaching (n.d.).

Measurement Issues

Personnel decisions are very high stakes and value-added scores as well as scores from observational rubrics are frequently used for this purpose. For this reason, the use of a framework of standards is appropriate to evaluate if these types of measures are appropriately developed. Frameworks that can be used include *The Standards for Educational and Psychological Testing* that speaks of validity and reliability and provides useful methods for evaluating the appropriate uses of scores for making decisions (American Educational Research Association, American Psychological Association, and National Council on Measurement in Education, 1999). According to the authors, “the intent of the Standards is to promote the sound and ethical use of tests and to provide a basis for evaluating the quality of testing practices”; in addition, the purpose is “to provide criteria for the evaluation of tests, testing practices, and the effects of test use” (American Educational Research Association, American Psychological Association, and National Council on Measurement in Education, 1999, pp. 1-2).

There are several standards that apply for scores derived from value-added models and observational rubrics. Scores from these two measures are being used to make decisions and the standards are designed to promote sound practices. Relevant standards can be seen in Table 2 taken from *The Standards for Educational and Psychological Testing* (American Educational Research Association, American Psychological Association, and National Council on Measurement in Education, 1999). These standards are used as benchmarks to ensure that appropriate procedures for test development and score use are followed, and to provide evidence of validity.

Validity. According to the *Standards for Educational and Psychological Testing*, the term validity "refers to the degree to which evidence and theory support the interpretations of the test" (American Educational Research Association, American Psychological Association, and National Council on Measurement in Education, 1999, p. 9). Validity evidence can be used to judge if a measure is actually measuring what it was intended to measure, and if it is used in the way in which it was intended (Cronbach, 1971; Crocker & Algina, 2006; Messick, 1981, 1993, 1995). Validity is not a property of the test, but rather, of the scores of the test (Messick, 1995). The inspection of validity is important for value-added models because it can provide evidence of the appropriateness of the resulting scores.

Validity evidence can be obtained by gathering information surrounding the measure (Crocker & Algina, 2006; Kane, 2006). For this reason there are several sources that could be used to gather evidence for validity including inspection of the content, the internal structure of the measure (e.g., exploratory and confirmatory factor analysis), and

Table 2

Relevant Standards for Instrument Development and Interpretation of Scores

Standard	Description
Standard 1.3	If validity for some common or likely interpretation has not been investigated, or if the interpretation is inconsistent with available evidence, that fact should be made clear and potential users should be cautioned about making unsupported interpretations.
Standard 1.4	If a test is used in a way that has not been validated, it is incumbent on the user to justify the new use, collecting new evidence if necessary.
Standard 1.11	If the rationale for a test use or interpretation depends on premises about the relationships among parts of the test, evidence concerning the internal structure of the test should be provided.
Standard 2.1	For each total score, subscore, or combination of scores that is to be interpreted, estimates of relevant reliabilities and standard errors of measurement information functions should be reported.
Standard 2.10	When subjective judgment enters into test scoring, evidence should be provided on both inter-rater consistency in scoring and within-examinee consistency over repeated measurements.
Standard 3.24	When scoring is done locally and requires scorer judgment, the test user is responsible for providing adequate training and instruction to the scorers and for examining scorer agreement and accuracy.
Standard 14.5	Individuals conducting and interpreting empirical studies of predictor-criterion relationships should identify contaminants and artifacts that may have influenced study findings, such as error of measurement, range restriction, and the effects of missing data.

relationship of the scores to other variables, to name a few (American Educational Research Association, American Psychological Association, and National Council on Measurement in Education, 1999). Collecting various types of evidence could determine if a measure is in fact performing and being used as intended.

Validation requires several sources of evidence that can be collected in the form of correlations, differentiation between groups, factor analysis, multitrait-multimethod analyses, or other approaches (Campbell & Fiske, 1959; Crocker & Algina, 2006). In theory, a measure should correlate better with an independent measure that measures the same trait versus a different trait (Campbell & Fiske, 1959). Value-added models can be compared to other measures purported to measure the same construct to obtain validity evidence.

Nomological network. One method that can be used to gather evidence of the validity of a measure relies on what Cronbach and Meehl (1955) referred to as the nomological network. According to Cronbach and Meehl, a nomological network can be viewed as an "interlocking system of laws which constitute a theory" (1955, p. 290). The nomological network aims to look at the relationships between constructs as specified by some theory.

Nomological network relationships can be investigated through several statistical methods. Statistical relations can be investigated through simple statistical methods such as the Pearson product moment correlation if the variables allow for it, or through more sophisticated methods such as structural equation modeling (SEM; Benson, 1998; Benson & Hagtvet, 1996; Brennan, 2006; Graham, 2008; McDonald, 1999), hierarchical linear modeling, or factor analyses (Brennan, 2006). Through the use of these statistical methods, the relationships between variables suggested by theory can be examined, thus providing evidence of the validity of the measures used to represent the constructs within the networks.

Summary

Since NCLB was introduced as law in 2002 and the newly passed Florida Senate Bill 736, which ties teachers' salaries to student achievement through their scores on assessments, VAM has grown in popularity as a tool in the accountability process. Providing validity evidence of these VAM scores through inspection of several sources of information is imperative to understanding how well the scores from these models are functioning. Value-added model scores are being used in high stake situations as they influence teacher continued employment, and therefore the need for validity evidence is critical. Validity evidence, as determined by examining if these scores are correlated with variables that are theoretically meaningful, is needed if VAM scores are to be used for making decisions.

Chapter Three: Methods

The purpose of this study was to evaluate the validity of the value-added scores used in teacher evaluations by examining the relation of these scores to widely used indicators of effective and quality teaching. According to the *Standards for Educational and Psychological Measurement*, "[a] sound validity argument integrates various strands of evidence . . ." (American Educational Research Association, American Psychological Association, and National Council on Measurement in Education, 1999, p. 17). Information on the validity of VAM scores is needed if these scores are to be used as indicators of teacher quality.

Participants and Setting

This study focused on a large southeastern school district in Florida. The State of Florida contracted with the American Institute for Research to compute VAM scores for all teachers in this southeastern school district who taught students who took the Reading or Mathematics Florida Comprehensive Assessment Test at the end of the 2011-2012 school year. The computed VAM scores for the teachers were released to Florida districts in October of 2012.

The district employs around 8000 teachers at all levels (elementary, middle, and high school) and 3,687 teachers received a reading or mathematics VAM score from the state for the 2011-2012 school year. Because VAM scores are computed using up to three years of prior data, some teachers who received a VAM score from the state were not observed by an administrator (they may have retired, transferred, changed positions within the district to non-instructional staff, etc.). Table 3 provides descriptive statistics for the demographic variables (e.g., teacher gender) used to answer the research questions, separated by samples (cases with both VAM and observational scores; cases with observational scores; cases with VAM scores).

Value-Added Model (VAM) Scores

According to the Student Success Act (2011), “at least 50% of a [teacher’s] performance evaluation must be based upon data and indicators of student learning growth assessed annually and measured by statewide assessments or, for subjects and grade levels not measured by statewide assessments, by district assessments”. Scores from value-added models, using data from the FCAT, were chosen by the state to meet this need.

The value-added model adopted in the State of Florida estimates the effects of 10 predictors on the current year student score on the FCAT, demonstrating the typical growth for a student as compared to similar students around the state. The model simultaneously estimates the school and teacher effect estimates on student learning as deviation scores from the typical amount of learning in the state (Florida’s Value-Added Technical Assistance Workshop, 2011). The final teacher value-added score, according to Florida’s Value-Added Technical Assistance Workshop (2011) can be seen in Equation 3.

Table 3

Demographics for Teachers in the District, Separated by Types of Scores

Variable		*N*=2385 Teachers with both VAM score and Observational rubric score		*N*=6441 All teachers in the district who received a score on the observational rubric		*N*=3687 Teachers in the district who received a VAM score from the state	
Gender	Male	391 (16.4%)		1357 (21.1%)		680 (18.4%)	
	Female	1994 (83.6%)		5084 (78.9%)		3007 (81.6%)	
Years Teaching Experience (Total)	< 1	19 (0.8%)		32 (0.5%)		176 (4.8%)	
	1-5	653 (27.4%)		1374 (21.3%)		1198 (32.0%)	
	6-10	620 (26.0%)		1429 (22.2%)		917 (25.0%)	
	>10	1093 (45.8%)		3606 (56.0%)		1396 (38.0%)	
Race	Asian	28 (1.2%)		77 (1.2%)		46 (1.2%)	
	Black	201 (8.4%)		510 (7.9%)		345 (9.4%)	
	Hawaiian/Pacific Islander	5 (0.2%)		10 (0.2%)		7 (0.2)	
	American Indian/Alaskan	21 (0.9%)		59 (0.9%)		31 (0.8%)	
	White	2122 (89.0%)		5849 (90.8%)		3294 (89.3%)	
Ethnicity (Marked YES to Hispanic/Latino regardless of Race)	Asian	0		1		1	
	Black	4		15		4	
	Hawaiian/Pacific Islander	2		2		2	
	American Indian/Alaskan	3		13		5	
	White	87		290		133	
	Total	95 (4.0%)		310 (4.8%)		142 (3.9%)	
National Board Certified	Yes	140 (5.9%)		150 (2.3%)		184 (5.0%)	
	No	2245 (94.1%)		6291 (97.7%)		3503 (95.0%)	
Number of schools represented in the sample		104 (Comprised of 16 High schools, 18 Middle schools, 68 elementary schools, and 2 k-8 schools)		126 (Comprised of 18 High schools, 19 middle schools, 73 elementary schools, and 16 Special Schools)		150 (Comprised of 22 high schools, 19 middle schools, 76 elementary schools, and 33 special schools)	
Frequencies of teachers by schools		Frequency of teachers	# of Schools	Frequency of teachers	# of Schools	Frequency of teachers	# of Schools
		120 to <130	0	120 to <130	1	120 to <130	0
		110 to <120	0	110 to <120	5	110 to <120	0
		100 to <110	0	100 to <110	4	100 to <110	0

Table 3 (continued)

Variable	*N*=2385 Sample containing both VAM score and Observational rubric score		*N*=6441 All teachers in the district who received a score on the observational rubric		*N*=3687 Teachers in the district who received a VAM score from the state	
Frequencies of teachers by schools	90 to <100	0	90 to <100	2	90 to <100	0
	80 to <90	0	80 to <90	7	80 to <90	1
	70 to <80	0	70 to <80	5	70 to <80	3
	60 to <70	0	60 to <70	9	60 to <70	6
	50 to <60	0	50 to <60	14	50 to <60	12
	40 to <50	18	40 to <50	41	40 to <50	10
	30 to <40	12	30 to <40	21	30 to <40	6
	20 to <30	14	20 to <30	9	20 to <30	43
	10 to <20	59	10 to <20	2	10 to <20	45
	0 to <10	1	0 to <10	6	0 to <10	24
Grade levels represented in the sample	4th through 10th grades		K-12		4th through 10th grades	
Subject areas represented in the sample	Reading and Math		ALL		Reading and Math	

Note. Numbers represent the number of teachers. Numbers in parentheses are the percent.

$$\text{Teacher Value-Added Score} = \text{Unique Teacher Component} + .50 * \text{Common School Component} \qquad (3)$$

VAM scores use as part of the equation, FCAT reading and mathematics scores from students to account for prior achievement. These scores are calculated by AIR, a contractor of the State of Florida. For this reason, and since the FCAT is taken by students in April, VAM scores are not submitted to each district until October of each year. The files are delivered to the districts through a secure file transfer protocol (FTP) in which only authorized agents in each district are able to access the files provided. Files delivered to the district include teacher VAM score estimates by grade and their standard error of measurement for reading and mathematics scores.

The file also contains a combined score for each teacher, which aggregates the scores per teacher by grade and subject. This aggregation is computed by AIR as a weighted transformation where all VAM scores and estimates are converted into a common metric by dividing by the average years growth and then doing a weighted average of the scores by number of students (Florida's Value-Added Technical Assistance Workshop, 2011). The result is a score where if a teacher only teaches one subject, the VAM score is an aggregation of either reading or mathematics by grade levels, or if the teacher instructs both subjects the calculation is an aggregation of reading and mathematics by grade levels.

This research used the combined scores for teachers. The standard error of measurement (SE) was taken into account and all combined VAM scores were transformed into a new score as presented in equation 4:

$$\text{VAM with SE} = \text{VAM score} + 1.96(\text{SE}) \tag{4}$$

This calculation created a score where all individuals received a score at the highest possible point in their 95% confidence band.

To ensure accuracy in rosters for students assigned to teachers and teachers assigned to courses, the State of Florida followed the statute requiring teachers be allowed to verify their rosters and make corrections for any mistakes (Student Success Act, 2011). The State of Florida, in combination with the Bill and Melinda Gates foundation and their Teacher Student Data Link Project, have provided each district with an online tool (using district survey files) that allows teachers access to verify, and modify the students who are attached to them (State Board of Education Presentation, 2012). This tool was open to teachers for three weeks in the month of May, 2012 for review and amendments. To ensure appropriate addition or deletion of students in the rosters, district rules mandated that any change made by teachers be approved/denied by their administrator, and then checked by the area superintendents. The district did not keep track of the number of changes that were made, approved, or denied.

AIR also had business rules for their calculations which affected the data. This includes only having students who had at least two years of assessment data available for prediction purposes (Webinar Presentation, 2012). This means that thought teachers may have had students correctly placed in their rosters, some students may not have been used in the VAM calculations because of lack of availability of prior year data. Further, any teacher with less than two students did not have a value-added score calculated for them (minimum *n*=2 by the State of Florida).

The Observational Rubric

An important variable that was used to provide evidence of the validity of VAM scores was the observational rubric developed by the large southeastern Florida district.

In order to be able to provide validity evidence for VAM scores with the use of this rubric, as established in *The Standards for Educational and Psychological Testing,* it is first important to establish that the rubric itself provides valid and reliable scores for comparison (American Educational Research Association, American Psychological Association, and National Council on Measurement in Education, 1999). The creation of the observational instrument followed a rigorous process, which included a committee that followed research closely, training of the raters, and pilot testing in 15 schools during the 2010-11 school year prior to district wide implementation in the 2011-2012 school year.

Two examples of the 34 types of indicators in the rubric and their research support can be seen in Table 4. These indicators have been found to be associated with teachers who achieve higher academic results from their students. It is appropriate to compare the results of the observational rubric to VAM scores as they both attempt to measure the same construct (the observational tool is based on teacher practices that have been empirically documented to enhance student learning and VAM scores are meant to measure the effect a teacher has on the academic achievement of a student).

The observational rubric is completed by school administrators during the formal summative observation of teachers of about 30 minutes, occurring towards the end of the school year (May 2012). Administrators also complete at least one formative evaluation of every teacher during the year (though administrators are encouraged to complete more than one) lasting about 10-15 minutes. Formative evaluations require that administrators note the effectiveness of the teacher, and provide them with feedback for improvement.

Table 4

Examples of Indicators Used in the Teacher Observational Rubric

Indicator	Research Base
Does the teacher aid students in guiding and tracking their own educational progress?	Marzano, 2007
Does the teacher take initiative to understand and modify instruction and communication based on the diversity of the students?	Danielson, 2007

Formative evaluations are not centrally gathered by the district and remain in the control of the administrator while summative observations are collected through a web-based system where the administrator is able to enter the score and supporting evidence for the indicator.

All classroom teachers are evaluated on the same observational rubric by their administrators. In order to be able to observe a teacher, an administrator must have passed the district's rigorous training and be considered certified. The rubric indicators are evidence based and the certified administrators are not aware of the individual teacher's VAM score for the current year while observing and gathering data. Observers are instructed to only mark the indicators as successfully met if they can observe the particular evidence during the observation period or through the evidence teachers provide them. During the summative evaluation, administrators may use evidence gathered from the formative evaluations. During the observations administrators are to mark each of the indicators with a score of 0 (unsatisfactory: implementation of the indicator was called for but not exhibited), 1 (Developing/Needs Improvement:

implemented incorrectly or with parts missing), 2 (Effective: executed the majority of the strategy which had a positive effect on the majority of the students), or 3 (Highly Effective: created new strategies, adapted to benefit ALL students).

Following standard 2.1 in *The Standards for Educational and Psychological Testing*, estimates of reliabilities must be observed and reported (American Educational Research Association, American Psychological Association, and National Council on Measurement in Education, 1999). Review of the internal consistency of the results of the pilot test administered in 2010-2011 revealed appropriate Cronbach alpha values per construct. The names of the constructs and alpha values from the pilot test can be seen in Table 5. According to standard 2.10 in the *Standards for Educational and Psychological Testing*, inter-rater consistency should be provided when scoring is done by subjective judgment (American Educational Research Association, American Psychological Association, and National Council on Measurement in Education, 1999). To date there has been no inspection of inter-rater reliability on the scores from the observational rubric.

The approved observational rubric was used during the 2011-2012 school year for all teacher evaluations in all schools in the district. Only elementary, middle, and high schools were used to answer the main research questions in this study (e.g., no charter schools, adult education programs, etc.).

Teacher Quality and Effectiveness Variables

This study focused on several variables that have historically been used to represent teacher quality and effectiveness. The thought behind these measured variables

Table 5

Cronbach Alpha Values for the Pilot Administration by Construct (2010-11 School Year)

Construct	Cronbach α	# of Indicators
1.1 Ability to Assess Instructional Needs	.828	5
1.2 Plans and Delivers Instruction	.904	9
2.1 Maintains a Student-Centered Learning Environment	.708	11
3.1 Performs Professional Responsibilities	.733	2
3.2 Engages in Continuous Improvement for Self and School	.869	7

is that teachers who hold these characteristics, degrees, or certifications are more highly qualified than those who do not. As described by Strong (2011), there is no exact definition of what a quality teacher actually must have or be in order to be designated that, instead, there are many types of characteristics that might make a teacher highly qualified.

This study also used variables that should theoretically have no relationship between the teachers' characteristics and student achievement. Since there is no axiom for what a quality teacher is, several variables should be inspected when attempting, through a validity study, to understand the performance of value-added modeling scores.

Though there is not one accepted understanding of what a quality teacher means, there are assumptions of relationships that should be present between certain variables. In a nomological network, one can determine the connections between variables and then calculate correlations involving these hypothesized relationships. Since there are several theories on what a quality teacher should be, including what the teacher comes to the job with (certifications), how they behave and perform in the classroom (results on their evaluations, amount of time they have been teaching), and the results they achieve from their students (value-added scores), all are inspected in this study (Strong, 2011). The hypothesis was that these variables should correlate to some degree with VAM scores.

Variables that were expected to have no correlation with how effective a teacher may be can also be used for a validity argument since the study should find little to no relationship between these variables and the two measures of quality teaching: the VAM scores and the observational rubric scores. Table 6 depicts the variables used in this study and the hypothesized relationships with the VAM scores, while Table 7 presents the timeline for when the data were collected, by whom, analyses needed, and when the data were received by the district. Though these variables may not be perfect indicators of teacher quality, and some may have received criticism, inspection of the relationships between them will provide validity information as part of the nomological network.

Design

This study used a multi-method quantitative design. Validity evidence for VAM scores was provided using several techniques and methods. The nested structure of this data was taken into consideration in all analyses conducted. Appropriate sample sizes are discussed for each question and method used. Prior to any analysis, preliminary analyses were conducted to include descriptive analyses to look at distributions of variables (e.g., skeweness and kurtosis, outliers), patterns of missing data, demographic characteristics of the sample, inspection of violations of the assumptions (if applicable), etc.

Table 6

Teacher Variables Used for Validity Evidence

Variable	Scale	Original Purpose/Gathered From	Hypothesized correlation based on research
Teacher Value-Added Score	Scores are aggregated for teachers across subjects and grades. Range from - 2 to 2	Provided by the State of Florida to be used as the student data portion of the teacher evaluation per Senate Bill 736	Variable of interest
Administrator Evaluation Score of the teacher on the observational Rubric. Highly Effective to Unsatisfactory (0-3 point scale)	Variable ranges from 0 to 102	From the Evaluation Appraisal instrument developed for 15 pilot schools under TIF grant	
Teacher Years of Experience	The number of years as a classroom teacher. Includes years in all districts teacher has worked in. (in addition to years, years^2 was used as a predictor of VAM scores)	From staff survey for the district	Years of experience have been found to be positively related to student achievement (Bosshardt & Watts, 1990; Card & Krueger, 1992; Ehrenberg & Brewer, 1994; Grimes & Register, 1990; Hanushek, 1992, 1997; Montmarquette & Mahseredjian, 1989; Murnane & Phillips, 1981ab)
National Board Certified Teacher	Certification treated as binary (1=has certification, 0=does not have certification)	Data obtained from the VAM files delivered from the state	Research has demonstrated that possessing a National Board Certification has a positive outcome for student performance (Card & Krueger, 1996; Cavalluzzo, 2004; Goldhaber & Anthony, 2007; Sato, Chung, & Darling-Hammond, 2008; Smith, Gordon, Colby, & Wang, 2005; Vandevoort, Amrein-Beardsley, & Berliner, 2004)

Table 6 (continued)

Variable	Scale	Original Purpose/Gathered From	Hypothesized correlation based on research
Teacher Gender	Originally coded with alphabetical letters (M, F) but was recoded to binary (0=Male, 1=Female)	Mandated staff reporting by the FLDOE	Research has demonstrated little to no correlation between teacher gender and student achievement (Dee, 2005; Ehrenberg & Brewer, 1994; Nixon & Robinson, 1999)
Teacher Race	There are five variables (American Indian or Alaska Native, Asian, Black or African American, Native Hawaiian or Other Pacific Islander, White). These are each coded in alpha character of Y or N and were recoded into binary 0=no and 1=yes	Mandated staff reporting by the FLDOE	Very small positive to no connection found between a teacher's race and student achievement (Dee, 2004; Ferguson, 1998; Hanushek, 1971; Strong, 2011)
Teacher Ethnicity	Alpha character designating Y=the staff is of Hispanic or Latino origin or No. This variable was changed to binary with 0=no and 1=yes to Hispanic origin	Mandated staff reporting by the FLDOE	Very small positive to no connection found between having a teacher of the same ethnicity as their students (Dee, 2004, Ferguson, 1998; Hanushek, 1971; Strong, 2011)

Table 7

Timeline for Variable Collection

Variable Name	Party Responsible for Collection and date	Date(s) analyzed	Date(s) Received by the district
VAM Scores	Students took the FCAT April 16-27, 2012 (Statewide Assessment Schedule, 2012); the state delivers the scores to AIR	AIR analyzed the scores and computed a VAM estimate score for each teacher in the district from the students assigned to them.	These scores were delivered to the district in October 2012
Observational Rubric Scores	Administrators in the district observed all teachers May 2012	The district entered the scores into the main database June 2012	N/A

Table 7 (continued)

Variable Name	Party Responsible for Collection and date	Date(s) analyzed	Date(s) Received by the district
National Board Certification	State mandated staff survey variable. Verified every year by the district and uploaded to the state three times a year in survey 2 (October), survey 3 (February), and survey 5 (August)	N/A	N/A
Years of experience	State mandated staff survey variable. Verified every year by the district and uploaded to the state three times a year in survey 2 (October) survey 3 (February) and survey 5 (August). File used was survey 5 to ensure most accurate data for the school year	N/A	N/A
Race/Ethnicity	State mandated staff survey variable. Verified every year by the district and uploaded to the state three times a year in survey 2 (October), survey 3 (February), and survey 5 (August).	N/A	N/A
Gender	State mandated staff survey variable. Verified every year by the district and uploaded to the state three times a year in survey 2 (October), survey 3 (February), and survey 5 (August).	N/A	N/A

Research question one. In order to provide validity evidence for VAM scores, it is imperative to establish that the scores from the instrument that are used as part of the validity argument are reliable and valid. This is to say that the scores that are produced from this measure, and the measure itself, produce results that are accurate reflections of the teachers' characteristics they intend to evaluate. The study of the observational rubric was done in two parts to examine different psychometric aspects of this measure. These aspects included the dimensionality of the instrument and internal consistency reliability of the scores from the observational rubric.

Dimensionality of the observational rubric. To investigate the dimensionality of the observational system, a confirmatory factor analysis that took into account the nested data structure was conducted using maximum likelihood estimation in the Mplus 5.21 software. Initially, the five-factor model underlying the observational measure was tested for fit. Fit indices were used to assess the model whose sets of variances, covariances and paths fit the data the best. Fit indices measure the discrepancy between the covariance matrix of the sample and the covariance matrix implied by the model (Hancock & Mueller, 2006).

The fit indices that were used in these analyses included the chi-square, the standardized root mean square residual (SRMSR), the root mean square error of approximation (RMSEA) index, and the comparative fit index (CFI). It is important to keep in mind that the cut off points for all of these measures are subjective (Browne & Cudeck, 1993; Steiger, 2000). Regardless, there are some generally accepted standards for cut point values for the fit indices that were used to assess model fit.

The desired outcome of a chi-squared analysis would be to find no evidence for which to reject the null hypothesis indicating no deviation from the true model (Hu & Bentler, 1998; Steiger, 2007). Thus, in order for a model to be considered to have appropriate fit a researcher would hope to find a non-statistically significant chi-squared value ($p > .05$), though this is uncommon in most CFA models. The CFI compares the misfit between the target model and the baseline model (Bentler, 1990). The closer this number is to 1.0 the better the fit but it should not be lower than .95 (Raykov & Marcoulides, 2006). Other researchers find a CFI of at least .90 to be acceptable (Bentler & Bonett, 1980). The SRMSR looks for values that are lower than .08 (Hu & Bentler, 1999). The RMSEA looks at the degree of misfit in the proposed model (Browne & Cudeck, 1993). The accepted cut off point is ≤ .05 but Browne and Cudeck (1993) also indicate that results between .05 and .08 suggest fair model fit. Others believe .06 to be an acceptable cut off point for the RMSEA (Hu & Bentler, 1999). The results of this research relied mostly on the aforementioned indices and looked at the results for well-fitting models to be within the fit measures specified.

Initially, the entire sample of teachers in the district who had an observation score regardless of grade or subject taught (*N*=6441) was used for identifying the psychometric properties of the observational rubric for each of the five underlying constructs in this multilevel setting (teachers within schools evaluated by administrators). The names of the five constructs are: Ability to Assess Instructional Needs (5 Items), Plans and Delivers Instruction (9 Items), Maintains a Student-Centered Learning Environment (11 Items), Performs Professional Responsibilities (2 Items), and Engages in Continuous Improvement for Self and School (7 Items). The factor model representation of the instrument is depicted in Figure 2. Results were inspected in terms of fit, as well as a table containing the unstandardized parameter estimates.

Once the fit of the aforementioned model was inspected, a second model was run. This model contained the sample of teachers in the district who had an observation done

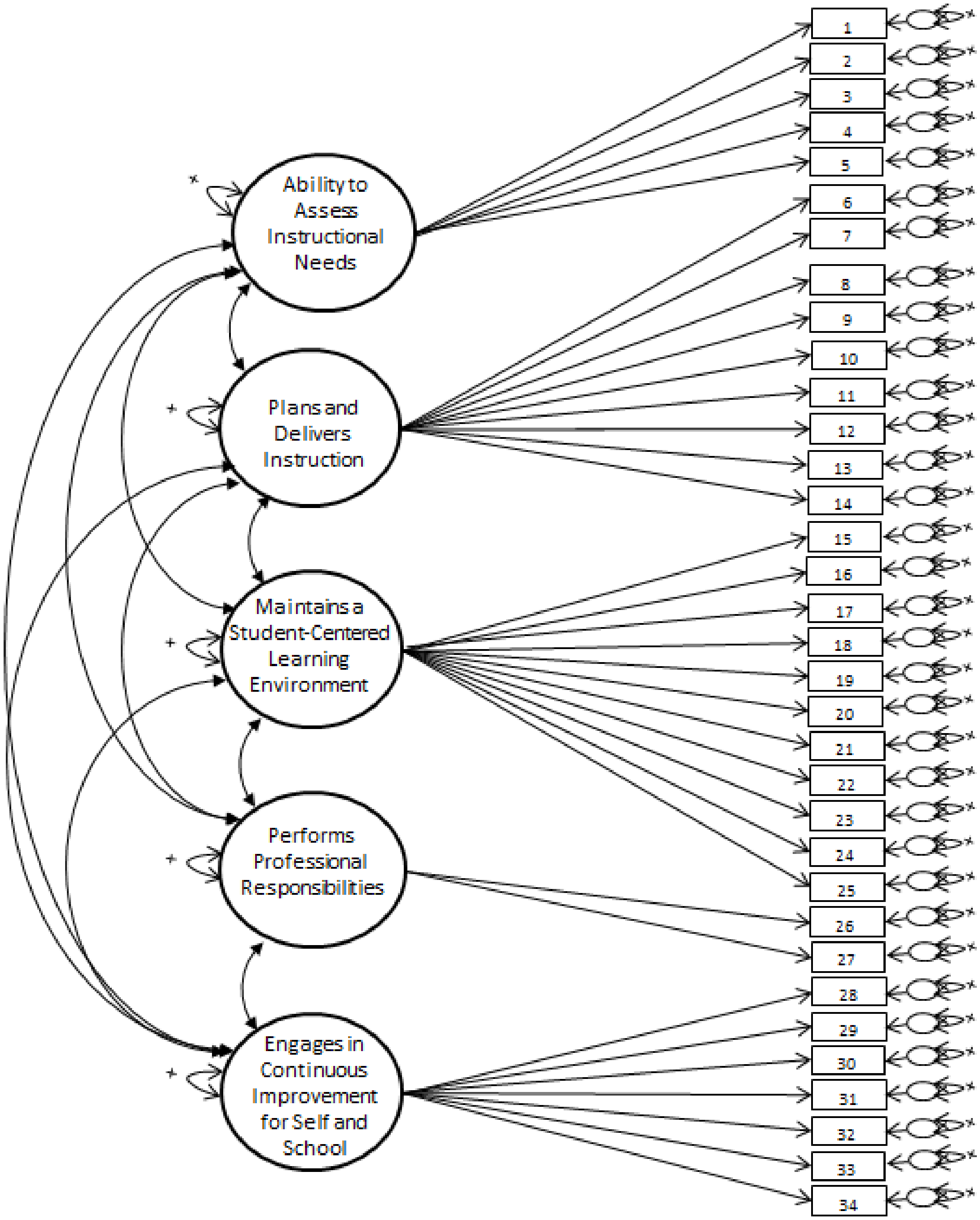

Ability to Assess Instructional Needs
Plans and Delivers Instruction
Maintains a Student-Centered Learning Environment
Performs Professional Responsibilities
Engages in Continuous Improvement for Self and School
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34

Figure 2. Factor model of the observational evaluation instrument.

by an administrator as well as a VAM score in reading or mathematics from the state (*N*=2385). This sample was used for identifying the psychometric properties of the observational rubric for each of the five underlying constructs in this multilevel setting (teachers within schools evaluated by administrators). The sample of 2,385 teachers was used to answer research questions two and three.

The scores from the observational rubric are used as a total composite score in the district. To statistically investigate the observational rubric consistent with how it is actually used, a second-order CFA was analyzed. The second-order latent factor, called Total Score, was made up of the five factors underlying the rubric (Figure 3). The results of this model were provided for the same two samples (*N* =6441 and 2385) previously mentioned.

Internal consistency reliability of the observational rubric. To investigate the internal consistency reliability of the scores, Cronbach alphas were computed. Considering that the focus of this study is at the teacher level, and not the school level, this method is appropriate for reliability estimates of the scores. This study was done in two parts. The entire sample size (*N*=6441) was used for this portion of the study and included all teachers who were evaluated using the observational rubric during the 2011-2012 school year. The second part used the sample of teachers who had an observational score from an administrator as well as a VAM score from the state (*N*=2385). These analyses provided evidence of the internal consistency of the scores.

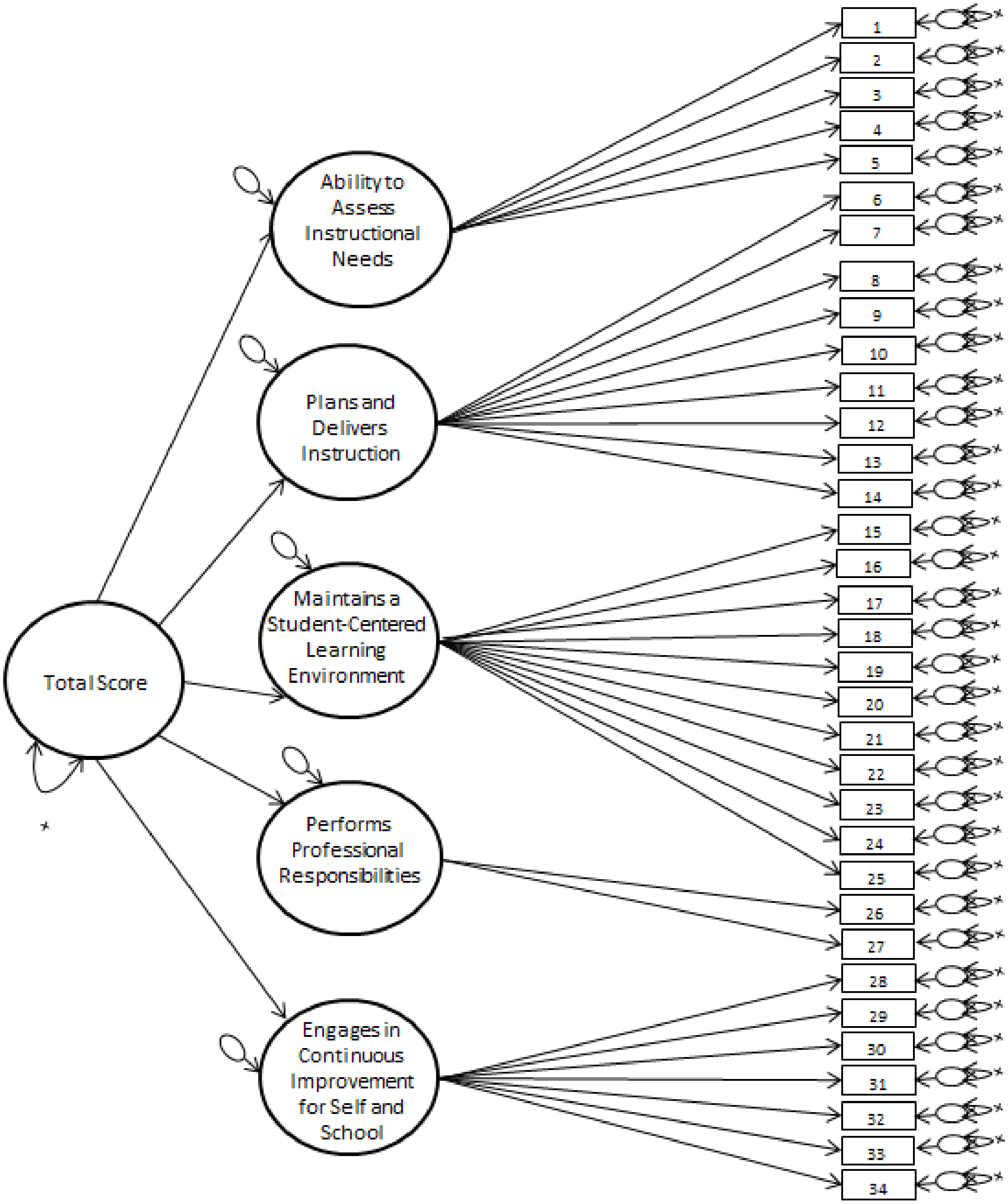

Total Score
*
Ability to Assess Instructional Needs
Plans and Delivers Instruction
Maintains a Student-Centered Learning Environment
Performs Professional Responsibilities
Engages in Continuous Improvement for Self and School
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34

Figure 3. Second-order factor model of the observational evaluation instrument.

Research question two. Question two asked if administrators' ratings on the observational rubric correlated with each teacher's effectiveness score as measured by the scores from the value-added model. Correlations between principal observations and VAM scores provide convergent validity evidence. In theory, what administrators observe and rate on the observational rubric, or their idea of what a good teacher means should correlate to some degree with VAM scores. Established practices hold that a correlation of about .60 or greater shows strong evidence for convergent validity and this is what was used as the benchmark for this study (Hill, Kapitula, & Umland, 2011).

Initially 126 site numbers (schools) were provided by the district for this study with the number of teachers observed per site ranging from one to 122 (*N*=6441). For the analysis of question two, cases which did not have both VAM scores and a score on the observational rubric were removed, resulting in the number of school sites decreasing to 119 with the number of teachers observed per site ranging from one to 49 (*n*=2572). Sites, as counted by the district, contained schools that were joined together by identification (ID) number by specialty type of schools (e.g., virtual school and teleschool were joined as one code). This means that several schools were collapsed into one school code as defined by the state of Florida. Because of this collapsing, the 120 site numbers, as provided by the district, were equivalent to 129 sites as defined by school ID number from the State of Florida. Analyses of the data relevant to research question two used the school identification numbers as defined by the State of Florida.

To ensure that there were enough cases within a school, the decision was made to remove from the sample any schools that had less than nine observations (16 sites and 58 cases). It is important to note that sites that contained less than 9 observations were generally not regular school sites (14 out of 16) and were instead, charter schools, virtual schools, jail schools or other such non-traditional schools, leaving 113 school sites for the analysis. Since this study focused only on traditional elementary, middle, and high schools, the remaining 2513 cases (113 school sites) were then inspected to remove charter schools, exceptional education centers, alternative schools, and career technical and adult education centers still remaining in the sample. This resulted in the deletion of nine school sites and 128 cases for a total remainder of 104 school sites (*N*=2385). The schools varied by number of teacher cases from nine to 49.

All available data were used for this part of the study (teachers with both a VAM score and a score on the observational rubric, *N*=2385). Appropriate power for this study was achieved with this sample size. It is recognized that there were nested data in this study since some administrators using the observational rubric rated several teachers (teachers within schools). The level two sample size was 104 (there were 104 schools in the final sample).

All data were examined for outliers, for missing data, and any major departures from the normal distribution. Relationships between variables were examined for linearity. Cases with missing VAM scores or observational scores were not included for this analysis. Outliers were examined using visual inspection of box and whisker plots as well as Mahalanobis distance analysis.

Mplus software was used to take into account the nested data structure when examining the correlations between the VAM scores and the five factors underlying the observational rubric. The model tested in this analysis can be seen in Figure 4. Initially, the model was tested for fit. Fit indices were used to assess the model whose sets of variances, covariances, and paths fit the data the best. The fit indices that were used in this model included the chi-squared, the standardized root mean square residual (SRMSR), the root mean square error of approximation (RMSEA) index, and the comparative fit index (CFI). The current accepted standards for cut point values for the fit indices were used to assess model fit (Bentler, 1990; Browne & Cudeck, 1993; Hu & Bentler, 1999; Raykov & Marcoulides, 2006).

To also inspect the relationship between these scores consistent with the way the observational rubric is used by the district, the relationship between the VAM scores and the total score on the observational rubric was analyzed. This was done by including a second-order factor to the original CFA with the five factors underlying the observational rubric (Figure 5). The same methods used above were repeated for this analysis.

Once the fit of the model was inspected, the relationship between the VAM scores and the observational rubric scores was examined. In order to ensure that the relationship found was accurate, several methods for observing the relationship were attempted. These methods included the use of VAM scores with and then without the standard error applied, and then inspection of the relationship of VAM scores and the scores from the observational rubric within each school.

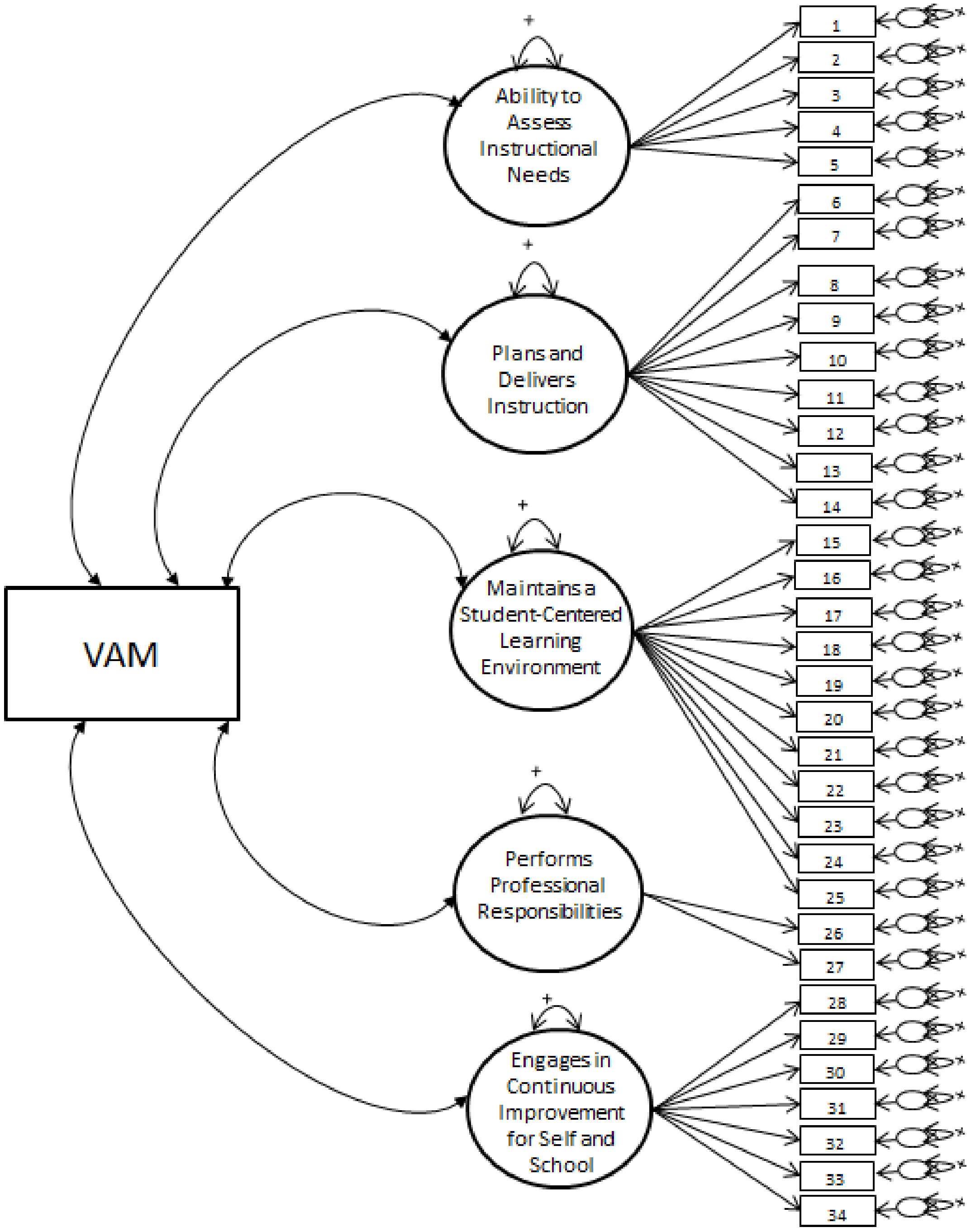

VAM
Ability to Assess Instructional Needs
Plans and Delivers Instruction
Maintains a Student-Centered Learning Environment
Performs Professional Responsibilities
Engages in Continuous Improvement for Self and School
+
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34

Figure 4. Relationship between VAM scores and the subscale scores from the observational rubric (all factors are correlated with each other).

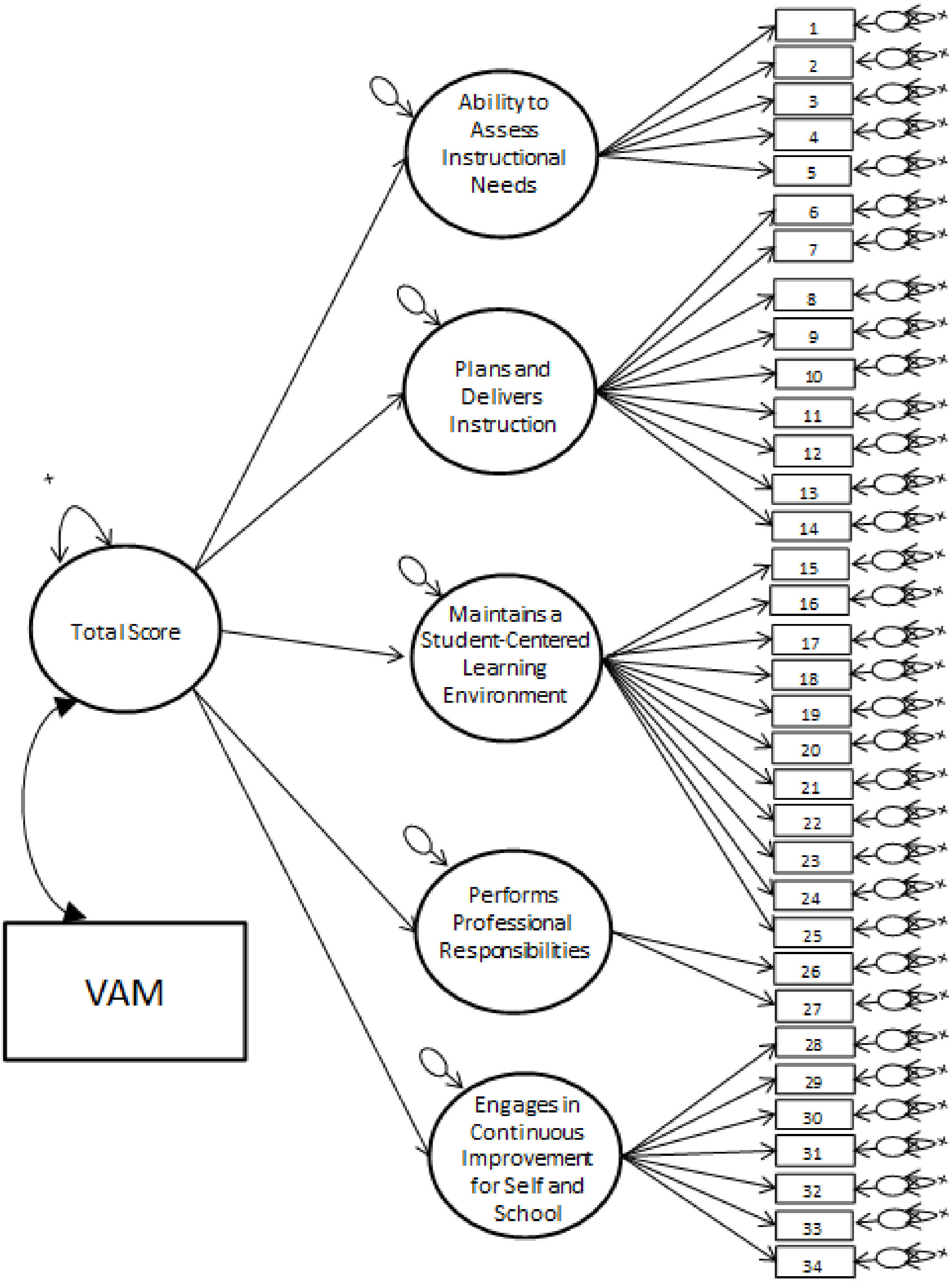

Ability to Assess Instructional Needs
Plans and Delivers Instruction
Maintains a Student-Centered Learning Environment
Performs Professional Responsibilities
Engages in Continuous Improvement for Self and School
Total Score
VAM
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34

Figure 5. Relationship between VAM scores and the second-order scores from the observational rubric.

Research question three. The third research question focused on how the VAM scores and the observational rubric scores related to other theoretically relevant teacher variables and did not relate to theoretically unrelated variables. These relationships were examined for the VAM scores and the observational scores. According to Cronbach and Meehl (1955), a nomological network helps define the meaning of a construct by making clear what the relationships are between constructs, observable variables, or a combination thereof. The nomological network addresses the theories behind these relationships and validity evidence is provided through the interpretation process.

As addressed in the study, though much research has focused on variables that identify quality teaching practices and in turn, quality teachers, there is not unanimous agreement on the topic. Since there is no agreement, further evidence on the relationships between variables that are thought to measure quality teaching should provide validity evidence for VAM scores and the scores from the observational rubric. Further, variables that should have no relationship with either the VAM scores or the scores from the observational rubric, inspected through a nomological network, also provide validity evidence.

This study examined the VAM scores and the scores from the observational rubric and studied how each related to variables that are meant to describe a quality teacher and to theoretically unrelated variables. Since there is no gold standard for a measure that identifies quality teaching, analysis of each measure (VAM scores and the observational rubric) and its relation to other variables could reveal the strengths and weaknesses of each of these measures.

Previous research guided what the expected relationships between the variables should be, and as such, determined if there is validity evidence through a nomological network. Since these variables, which represent correlates of quality teaching, have different support in regard to how they are related to quality teaching, the results of the analysis of the relationship between these variables and VAM scores and the observational rubric were considered equally weighted.

Several variables were used where a hypothesis could be made as to what relationship they would have with VAM scores and the scores from the observational rubric. The variables included National Board Certification (NBC) designation, years of experience, gender, race, and ethnicity. For these analyses some of the relationships were more exploratory and some were more confirmatory. Years of experience and NBC were predicted to have positive relationships to VAM scores and the scores from the observational rubric.

There were variables that theoretically should have little relationship with either the VAM or observational scores. For example, it was not expected that gender, race, or ethnicity would have a strong relation to either VAM scores or the observational rubric scores.

The number of years a teacher has been in the profession can vary greatly. Exploratory analyses focused on the relationship between VAM scores and the observational rubric and years of teaching experience, accounting for the potential ceiling effect of years of teaching experience (Murnane & Phillips, 1981b; Rockoff, 2004; Strong, 2011). These analyses explored both linear and nonlinear relations (e.g., quadratic effects) between years of teaching experience and VAM and observational scores.

Mplus was used to evaluate the relations between the predictor and outcome variables displayed in Figure 6 and Figure 7. Figure 8 depicts the relationship between the predictor variables and the second-order total score of the observational rubric scores.

As was done in previous analyses, preliminary descriptive analyses were conducted and statistical assumptions were evaluated. A structural regression model was examined to identify the patterns and relationships between the variables as well as evaluate the relationships between the constructs (Raykov & Marcoulides, 2006). Initially, the model in Figure 6 was tested for fit (the model in Figure 7 is fully saturated thus producing perfect fit). Fit indices were used to assess the model whose sets of variances, covariances, and paths fit the data the best. The fit indices that were used for the model include the chi-squared, the standardized root mean square residual (SRMSR), the root mean square error of approximation (RMSEA) index, and the comparative fit index (CFI). The current accepted standards for cut point values for the fit indices were used to assess model fit (Bentler, 1990; Browne & Cudeck, 1993; Hu & Bentler, 1999; Raykov & Marcoulides, 2006). The relationships between each of the teacher quality measures (VAM and the observational rubric scores) and the predictor variables were evaluated as a source of validity evidence.

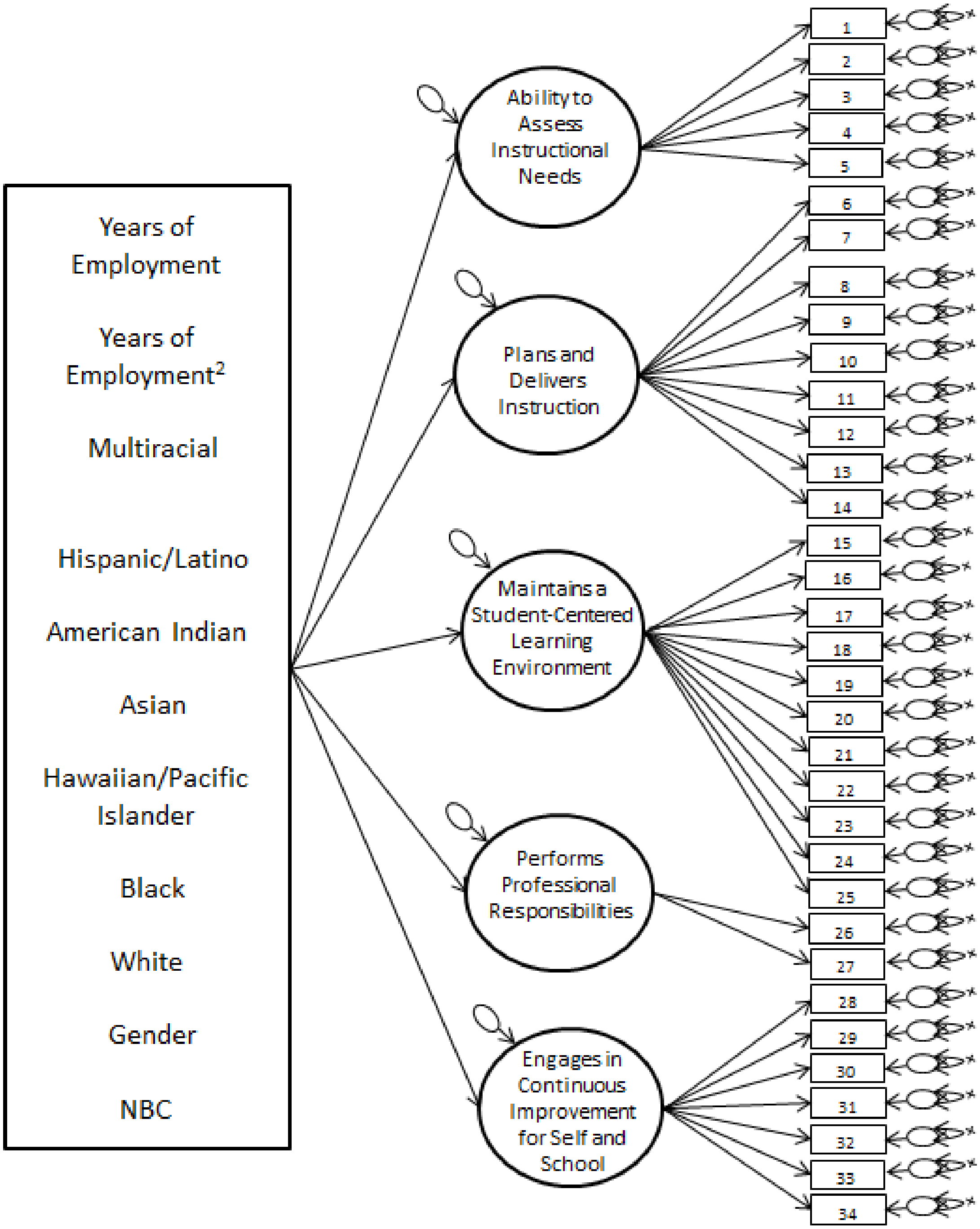

Years of Employment
Years of Employment^2
Multiracial
Hispanic/Latino
American Indian
Asian
Hawaiian/Pacific Islander
Black
White
Gender
NBC
Ability to Assess Instructional Needs
Plans and Delivers Instruction
Maintains a Student-Centered Learning Environment
Performs Professional Responsibilities
Engages in Continuous Improvement for Self and School
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34

Figure 6. Relationship between predictor variables and the observational rubric scores (all factors are correlated with each other).

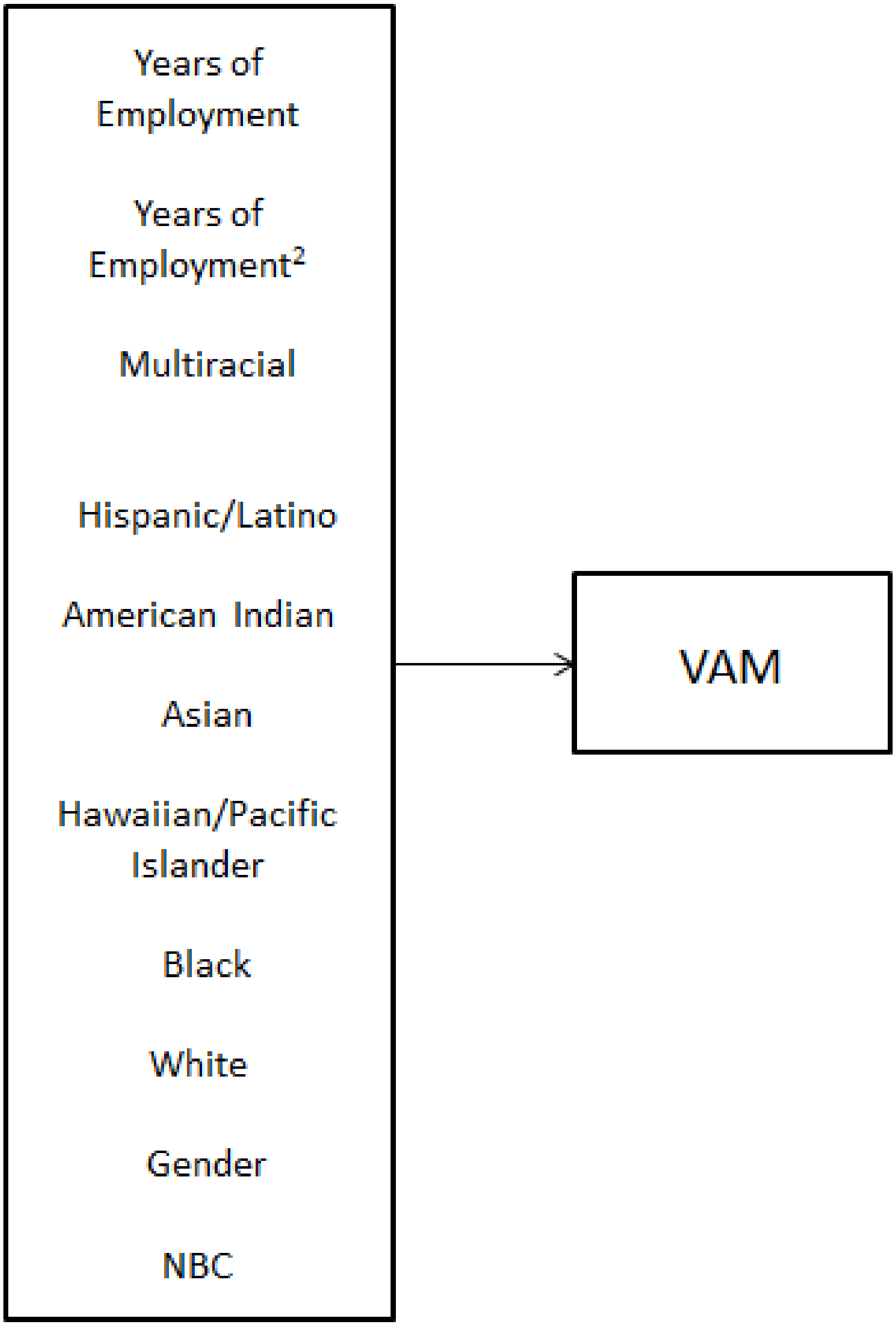


Figure 7. Relationship between predictor variables and VAM scores.

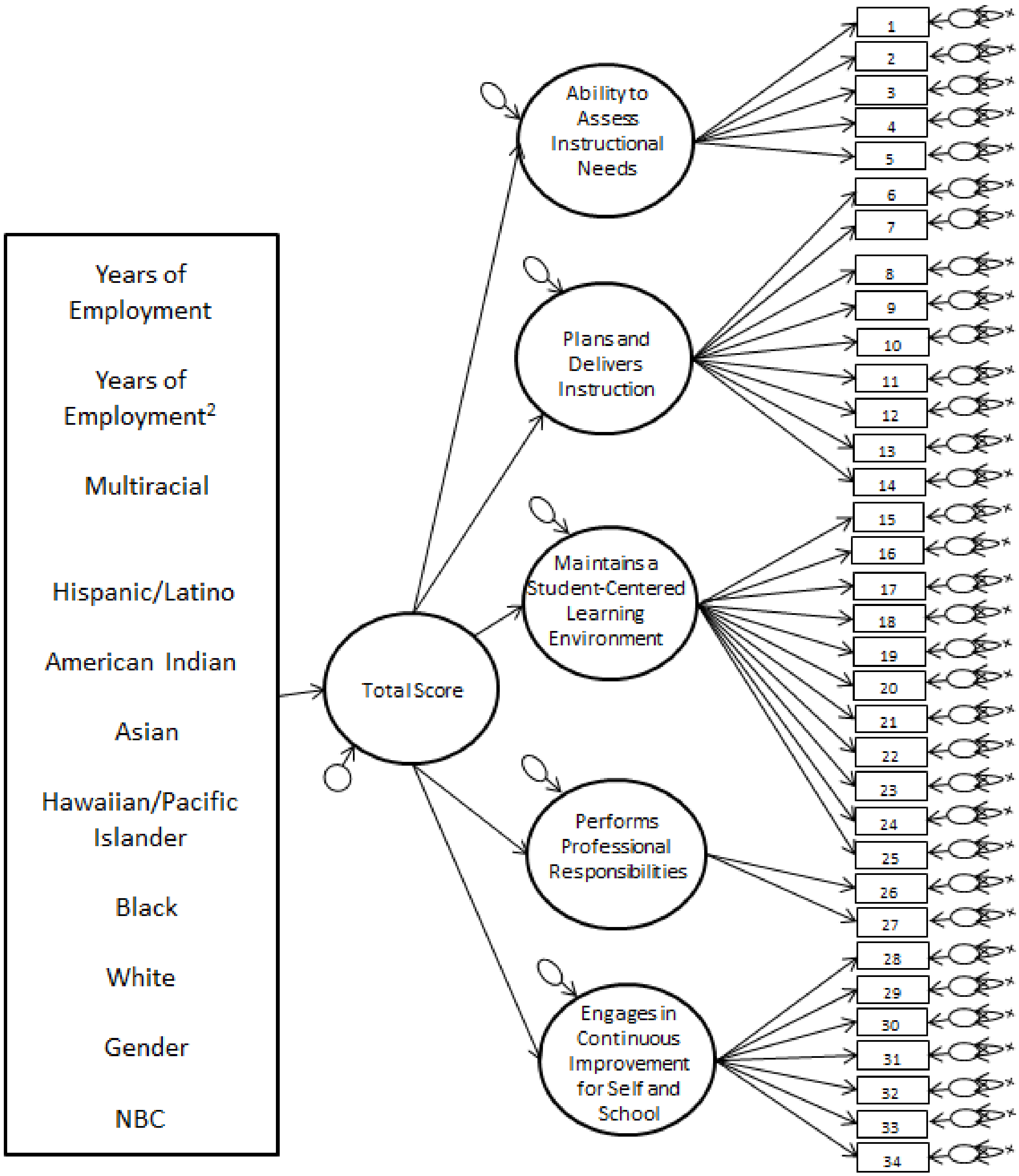

Years of Employment
Years of Employment²
Multiracial
Hispanic/Latino
American Indian
Asian
Hawaiian/Pacific Islander
Black
White
Gender
NBC
Total Score
Ability to Assess Instructional Needs
Plans and Delivers Instruction
Maintains a Student-Centered Learning Environment
Performs Professional Responsibilities
Engages in Continuous Improvement for Self and School
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34

Figure 8. Relationship between predictor variables and the second-order total score of the observational rubric scores.

Summary

The purpose of this study was to evaluate the validity of the VAM scores. Each piece of this study, taken together, can provide a clearer understanding of the relationship of VAM scores to other theoretically established variables of quality teachers. Though there is no gold standard to compare VAM scores to, providing several lines of evidence can add to the knowledge base of these scores.

Chapter Four: Results

The purpose of this study was to examine the validity of value-added scores for use in teacher evaluations. This chapter presents the results of this study organized by each research question. All of these questions are answered using data from a sample of teachers from a large southeastern school district.

The questions addressed by this study include:

1a) To what extent are the administrators' observational ratings of teachers collected during the 2011-2012 school year consistent with the five-factor measurement model underlying the observational rubric?

1b) For the observational rubric, what is the estimated internal consistency reliability of the scores for the five factors collected through observations during the 2011-2012 school year?

2) Do administrators' observational ratings of teachers based on the rubric correlate with teachers' value-added scores from the Florida VAM for the 2011-2012 school year?

3) Do the teachers' VAM scores for the 2011-2012 school year and the observational rubric relate to other theoretically relevant teacher variables (e.g., National Board Certification) and not to theoretically unrelated variables (e.g., gender, race/ethnicity)?

Data Source

The State of Florida provided the district with data for 2,613 teachers who received a value-added score using the FCAT for the 2011-2012 school year. The district provided data for 6,441 teachers who received a score based on the observational rubric. Out of that sample, 2,385 teachers had a VAM score from the state and a score based on the observational rubric administered by the school principal or assistant principal. Because VAM scores are computed using up to three years of prior data, some teachers who received a VAM score from the state were not observed by an administrator during the 2011-2012 school year (they may have retired, transferred, or changed positions within the district to non-instructional staff, etc.).

To answer the questions addressed in this research, different samples of varying sizes were used. For analysis of the observational rubric (i.e., the five-factor model and reliability), the entire sample of 6,441 teachers was used as well as the subset of teachers who had both scores (*N*=2385). For the remaining questions, the sample of teachers with both VAM and observational scores was used.

Research Question One

The first research question was answered in two parts. Part one focused on whether administrators' observational ratings of teachers were consistent with the five-factor model underlying the rubric; part two evaluated the estimated internal consistency reliability of the scores of the five-factor observational rubric. The nested structure of the data (teachers within schools) was taken into account in this analysis by using the type equal complex function in Mplus. "This estimation includes a Taylor series-like function to provide a normal theory covariance matrix for analysis... created by obtaining a weighted covariance matrix that combines the variances and covariances of the [primary sampling unit (Schools)]" (Hancock & Mueller, 2006, p. 352). Ignoring the violation to the independence of the sampling could lead to biased reliability estimates and improperly estimated standard errors (Geldhof, Preacher, & Zyphur, 2013; Hancock & Mueller, 2006; Snijders & Bosker, 1999). Single level analyses are not the most appropriate when sampling constitutes nested data structures.

Fit of the five-factor model. To address the extent that the observational rubric scores were consistent with the five-factor model (as can be seen in Chapter 3, Figure 2), confirmatory factor analysis (CFA), taking into account the clustering (complex/nested sampling) of the data, was conducted using the Mplus maximum likelihood estimation with robust standard errors (MLR). The model was run twice, the first time using the entire sample of teachers who had a score on the observational rubric (*N*=6441) and the second time with the sample of teachers who had a score from an administrator on the observational rubric as well as a VAM score from the state (*N*=2385).

The MLR estimation is robust to non-normal data, missing data, and non-independence of observations (Muthén & Muthén, 1998-2007) and thus appropriate for this analysis as it accounts for violations of the assumptions including all of the items must be univariately normal and all of the items together must be multivariate normal. See Table 8 for skeweness and kurtosis values for the sample of all teachers in the district with a score on the observational rubric, *N*=6441. Also see Table 9 for teachers in the district who had a score on the observational rubric and a VAM score from the state, *N*=2385.

Table 8

Descriptive Statistics for Items From the Observational Rubric for All Teachers With a Score in the District

Item on the Rubric	*N*	*M*	*SD*	Skeweness	Kurtosis	ICC
1.1 Ability to assess instructional needs						
I11A	6440	2.01	0.62	-0.09	-0.11	.27
I11B	6440	2.13	0.55	-0.02	0.48	.23
I11C	6439	2.16	0.57	-0.07	0.22	.25
I11D	6440	2.05	0.52	-0.02	0.96	.25
I11E	6438	2.13	0.52	0.03	1.16	.22
1.2 Plans and delivers instruction						
I12A	6437	2.27	0.58	-0.21	0.02	.21
I12B	6436	2.12	0.59	-0.10	-0.01	.27
I12C	6436	2.13	0.61	-0.16	-0.08	.20
I12D	6435	2.25	0.52	0.22	-0.14	.21
I12E	6435	2.20	0.54	0.09	0.19	.21
I12F	6435	2.02	0.54	-0.04	0.71	.21
I12G	6435	2.36	0.57	-0.27	-0.36	.20
I12H	6435	2.15	0.57	-0.05	0.18	.15
I12I	6436	2.07	0.59	-0.07	0.06	.22
2.1 Maintains a student-centered learning environment						
I21A	6436	2.23	0.58	-0.10	-0.23	.26
I21B	6435	2.31	0.57	-0.19	-0.18	.22
I21C	6436	2.26	0.51	0.23	-0.02	.23
I21D	6435	2.28	0.57	-0.27	0.06	.22
I21E	6434	2.14	0.53	0.00	0.97	.22
I21F	6434	2.15	0.52	0.09	0.72	.25
I21G	6435	2.28	0.52	0.14	-0.21	.20
I21H	6435	2.31	0.52	0.14	-0.47	.24
I21I	6435	2.21	0.52	0.08	0.68	.23
I21J	6434	2.17	0.52	0.17	0.43	.27
I21K	6435	2.07	0.52	-0.01	1.03	.27
3.1 Performs professional responsibilities						
I31A	6436	2.32	0.53	-0.06	0.17	.29
I31B	6435	2.24	0.57	-0.27	0.79	.25
3.2 Engages in continuous improvement for self and school						
I32A	6435	2.35	0.58	-0.24	-0.62	.12
I32B	6436	2.33	0.55	-0.09	-0.31	.13
I32C	6435	2.26	0.55	-0.00	-0.01	.17
I32D	6434	2.31	0.52	0.13	-0.47	.20
I32E	6435	2.14	0.51	0.15	0.74	.22
I32F	6435	2.10	0.47	0.23	1.48	.22
I32G	6435	2.08	0.47	0.13	1.92	.27

Note. ICC=Intraclass correlation coefficient. Response scale ranged from 0 (Unsatisfactory) to 3 (Highly Effective).

Table 9

Descriptive Statistics for Items From the Observational Rubric for All Teachers With a Score in the District who Also Received a VAM Score From the State

Item on the Rubric	*N*	*M*	*SD*	Skeweness	Kurtosis	ICC
1.1 Ability to assess instructional needs						
I11A	2385	2.03	0.64	-0.09	-0.31	.22
I11B	2385	2.15	0.56	-0.06	0.34	.21
I11C	2385	2.19	0.58	-0.10	0.03	.23
I11D	2385	2.06	0.53	-0.04	0.93	.23
I11E	2384	2.15	0.53	0.01	0.92	.21
1.2 Plans and delivers instruction						
I12A	2383	2.26	0.59	-0.26	0.22	.19
I12B	2382	2.10	0.60	-0.14	0.06	.24
I12C	2382	2.13	0.61	-0.20	0.11	.16
I12D	2382	2.25	0.52	0.13	0.09	.19
I12E	2382	2.21	0.54	0.02	0.26	.20
I12F	2382	2.02	0.55	-0.09	0.67	.18
I12G	2382	2.33	0.59	-0.34	-0.09	.18
I12H	2382	2.17	0.55	-0.03	0.39	.11
I12I	2382	2.10	0.59	-0.12	0.12	.18
2.1 Maintains a student-centered learning environment						
I21A	2382	2.19	0.58	-0.08	-0.12	.22
I21B	2382	2.29	0.57	-0.18	0.07	.21
I21C	2383	2.24	0.50	0.26	0.24	.21
I21D	2382	2.27	0.59	-0.32	0.28	.21
I21E	2381	2.13	0.53	-0.02	1.10	.20
I21F	2381	2.13	0.53	0.03	0.82	.22
I21G	2382	2.31	0.54	-0.01	-0.06	.17
I21H	2382	2.31	0.53	0.07	-0.27	.19
I21I	2382	2.20	0.53	-0.06	0.98	.19
I21J	2381	2.18	0.53	0.08	0.43	.25
I21K	2382	2.07	0.54	-0.09	0.90	.26
3.1 Performs professional responsibilities						
I31A	2383	2.31	0.53	-0.11	0.64	.25
I31B	2382	2.22	0.57	-0.28	0.86	.21
3.2 Engages in continuous improvement for self and school						
I32A	2381	2.34	0.58	-0.22	-0.53	.10
I32B	2382	2.33	0.56	-0.19	-0.00	.10
I32C	2381	2.24	0.55	-0.03	0.17	.16
I32D	2381	2.31	0.53	0.01	-0.07	.16
I32E	2381	2.15	0.51	0.16	0.85	.20
I32F	2381	2.11	0.49	0.14	1.41	.19
I32G	2381	2.09	0.47	0.11	2.09	.20

Note. ICC=Intraclass correlation coefficient. Response scale ranged from 0 (Unsatisfactory) to 3 (Highly Effective).

Each of the confirmatory factor analysis (CFA) models consisted of five factors which were scaled by fixing the first item loading to 1.0 using the Mplus version 5.21 software while the remaining factor variances/covariances, factor loadings and residual estimates were freely estimated (Muthén & Muthén, 1998-2007). The defaults of the program were not changed leaving the error covariances set to zero (with the assumption that there should be no correlations amongst the error variances). Missing data were estimated in the model through MLR estimation (same as full information maximum likelihood or FIML where the same parameters are estimated but with the difference being that the Quasi-Newton method is used for the standard errors and the chi-squared when data are missing at random) in Mplus version 5.21, which assumes the data are missing completely at random (MCAR), or missing at random (MAR) (Muthén & Muthén, 1998-2007).

Descriptive statistics for the model with all teachers who received a score on the observational rubric (*N*=6441) are summarized in Table 8. The means of the items ranged from 2.01 (item 1.1.a. involving and guiding students in tracking their own progress) to 2.36 (item 1.2.g. what a teacher does to engage students in learning). The observed variables in this study were approximately normally distributed (see Table 8). Multivariate normality was inspected through box and whisker plots (see Figure 9) and with SPSS 21.0 using Mahalanobis distance. Significant multivariate outliers per latent factor ranged from 14 cases to 29 cases per factor, and 14 cases for the total score, but no cases were removed due to the robustness of the Mplus estimation software.

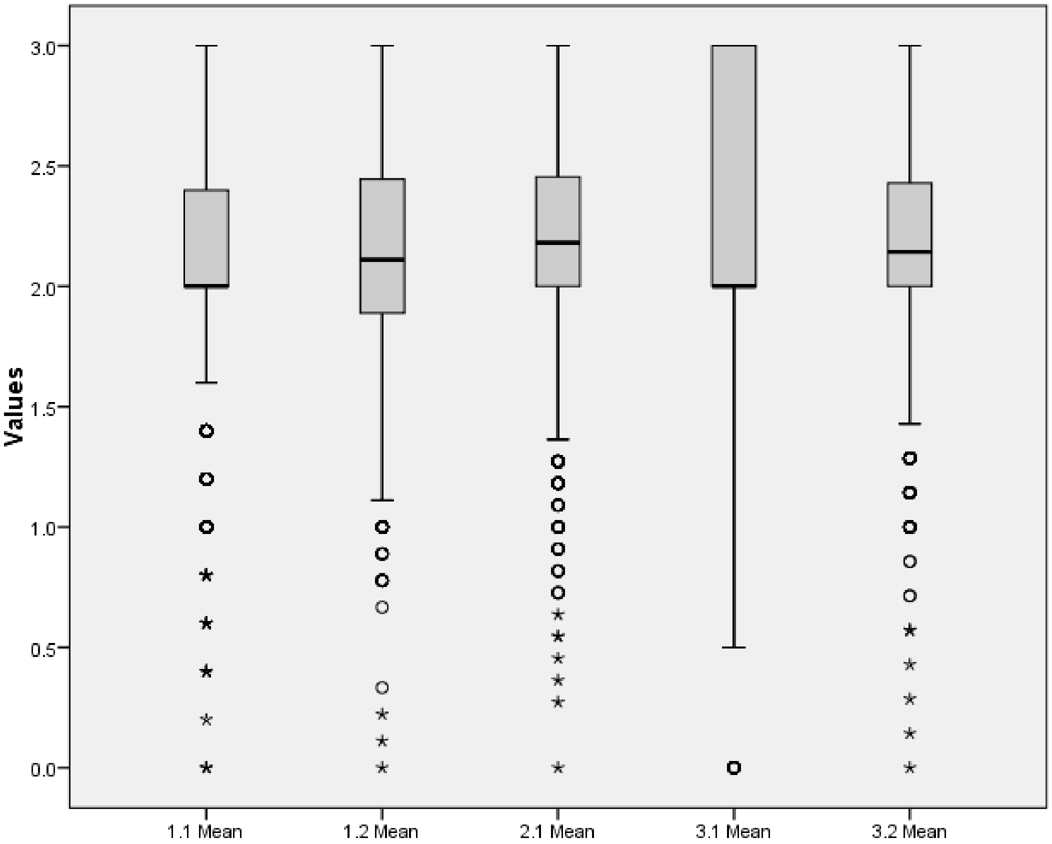


Figure 9. Box and whisker plots for the five subscale scores of the observational rubric (N=6441). The names of the factors are: 1.1 Ability to Assess Instructional Needs; 1.2 Plans and Delivers Instruction; 2.1 Maintains a Student-Centered Learning Environment; 3.1 Performs Professional Responsibilities; 3.2 Engages in Continuous Improvement for Self and School. Means of the subscales are: Factor 1.1=2.10; 1.2=2.17; 2.1=2.22; 3.1=2.28; 3.2=2.22.

The ICC "represents the ratio of a scale score's between-cluster variance relative to its total variability across both levels" (Geldhof, Preacher, & Zyphur, 2013, p. 12). The ICCs observed in Table 8 indicate that about 10% to 30% of the variance of each of the variables can be attributed to the school the teachers belonged to and are considered to be moderate to moderately high in size (Kreft & de Leeuw, 1998). This further supports the use of an analysis approach that takes into consideration the complex nested structure of the data.

Descriptive statistics for the model with all teachers who received a score on the observational rubric and a VAM score from the state (*N*=2385) are summarized in Table 9. The means of the items ranged from 2.02 (item 1.2.f. involving and guiding students in tracking their own progress) to 2.34 (item 3.2.a. what a teacher does to engage students in learning). The observed variables in this study were approximately normally distributed (see Table 9). Multivariate normality was inspected through box and whisker plots (see Figure 10) and with SPSS 21.0 using Mahalanobis distance. Significant multivariate outliers per latent factor ranged from 6 cases to 14 cases per factor, but no cases were removed due to the robustness of the Mplus estimation software.

The ICC's for this sample (*N*=2385) were also computed. The ICCs observed in Table 9 indicate that about 10% to 26% of the variance of each of the variables can be attributed to the school the teachers belonged to and are considered to be moderate to moderately high in size (Kreft & de Leeuw, 1998). This further supports the use of an analysis approach that takes into consideration the complex nested structure of the data.

To assess the fit of the models, several goodness-of-fit indicators were used. For the first model which contained all the teachers from the district who obtained a score from their administrator on the observational rubric, results were as follows. The chi-squared value demonstrated lack of fit of the five-factor model, $\chi^2(517, N = 6441) = 7{,}643.60, p <.001$. The ideal would be a non-significant chi-squared, indicating that the

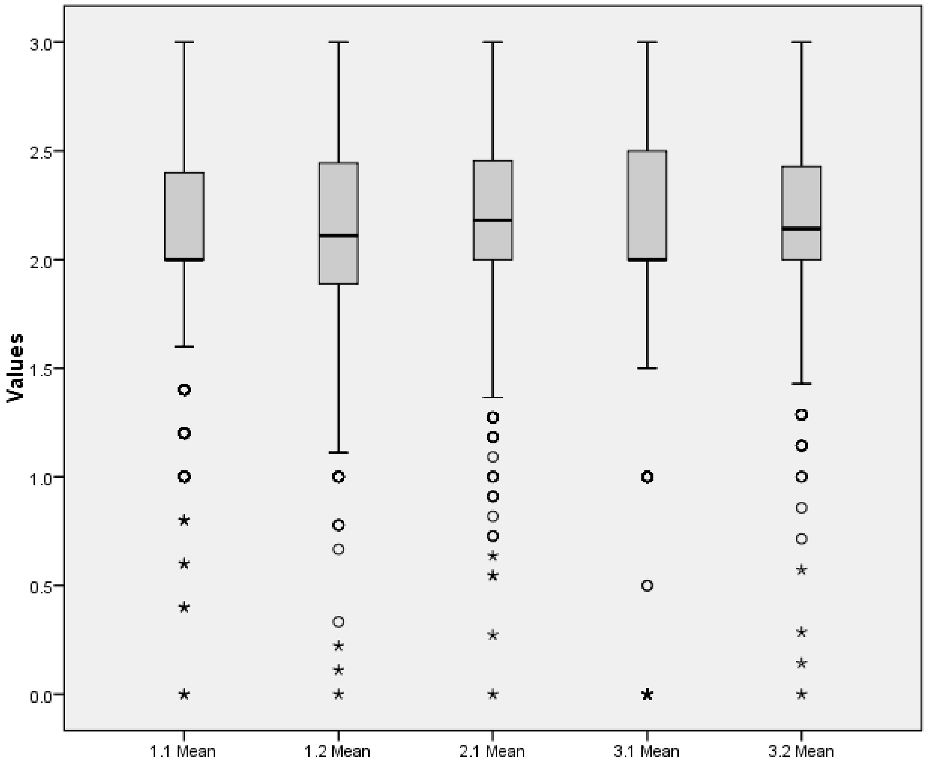


Figure 10. Box and whisker plots for the five subscale scores of the observational rubric (N=2385). The names of the factors are: 1.1 Ability to Assess Instructional Needs; 1.2 Plans and Delivers Instruction; 2.1 Maintains a Student-Centered Learning Environment; 3.1 Performs Professional Responsibilities; 3.2 Engages in Continuous Improvement for Self and School. Means of the subscales are: Factor 1.1=2.11; 1.2=2.18; 2.1=2.21; 3.1=2.26; 3.2=2.22.

model is "reproducing the population matrix of observed variable relationship indices" (Raykov & Marcoulides, 2006, p. 41). The chi-square has been known to be sensitive to sample size, thus other fit estimates were also used to ascertain how well the model fit (Bollen, 1990; Marsh et al., 1988). Other measures of fit suggested that the model presented had appropriate fit. The CFI was .914, higher than the cut off value of .90 (Bentler & Bonett, 1980). The RMSEA of .046 and the SRMR of .039 were both below the acceptable cut off point of .05 (Browne & Cudeck, 1993; Hu & Bentler, 1999). These fit measures are displayed in Table 10.

For the second model, which contained the sample of teachers who had both a score on the observational rubric by their administrators as well as a VAM score, the fit can be seen in Table 10. The chi-squared value demonstrated lack of fit of the five-factor model, $\chi^2(517, N = 2385) = 4{,}020.44$, $p <.001$. Other measures of fit suggest that the model presented had appropriate fit. The CFI was .904, higher than the cut off value of .90 (Bentler & Bonett, 1980). The RMSEA of .053 and the SRMR of .040 were both below the acceptable cut off point of .06 and .08, respectively (Browne & Cudeck, 1993; Hu & Bentler, 1999).

All loadings, variances, covariances and correlations between the latent factors were statistically significantly different from zero ($p < .01$). The unstandardized factor loadings can be seen in Table 11 for both models, as can the residual variances and the R^2, representing the proportion of the variance that can be explained by the indicator's factor.

Table 10

Confirmatory Factor Analysis: Fit Indices for the Five-Factor Model Underlying Administrators' Observational Ratings of Teachers for Sample With Only Observational Scores and Sample of Teachers With Observational Score and VAM Scores

Sample	X^2	*df*	CFI	RMSEA	SRMR
Sample of teachers in the district with observational rubric scores (*N*=6441)	7643.59	517	.914	.045	.039
Sample of teachers in the district with observational rubric scores and VAM scores (*N*=2385)	4,020.44	517	.904	.053	.040

Table 11

Confirmatory Factor Analysis: Unstandardized Factor Loadings, Residual Variances and R^2 for the Five-Factor Model Underlying Administrators' Observational Ratings of Teachers

	All teachers in the District with a score on the observational rubric (*N*=6441)			All teachers in the District with a score on the observational rubric and a VAM score (*N*=2385)		
Item on the Rubric	Factor Loading	Residual Variance	R^2	Factor Loading	Residual Variance	R^2
1.1 Ability to assess instructional needs						
I11A	1.00[a] (-)	0.20 (0.01)	0.47	1.00[a] (-)	0.21(0.01)	0.49
I11B	0.97 (0.03)	0.13 (0.01)	0.56	0.94(0.04)	0.14(0.01)	0.56
I11C	1.03 (0.04)	0.13 (0.01)	0.59	0.98(0.04)	0.14(0.01)	0.57
I11D	0.83 (0.04)	0.15 (0.01)	0.45	0.81(0.04)	0.15(0.01)	0.47
I11E	0.84 (0.04)	0.14 (0.01)	0.48	0.81(0.04)	0.15(0.01)	0.47
1.2 Plans and delivers instruction						
I12A	1.00[a] (-)	0.16 (0.01)	0.52	1.00[a] (-)	0.16(0.01)	0.54
I12B	0.98 (0.03)	0.18 (0.01)	0.48	0.94(0.04)	0.20(0.01)	0.45
I12C	1.00 (0.03)	0.20 (0.01)	0.47	0.97(0.04)	0.20(0.01)	0.46
I12D	0.89 (0.03)	0.13 (0.01)	0.51	0.91(0.03)	0.12(0.01)	0.56
I12E	0.91 (0.03)	0.14 (0.01)	0.51	0.93(0.04)	0.14(0.01)	0.54
I12F	0.84 (0.04)	0.16 (0.01)	0.43	0.88(0.04)	0.16(0.01)	0.47
I12G	1.00(0.02)	0.15 (0.01)	0.54	1.02(0.03)	0.16(0.01)	0.55
I12H	0.73 (0.03)	0.23 (0.01)	0.29	0.70(0.04)	0.21(0.01)	0.30
I12I	0.79 (0.04)	0.24 (0.01)	0.31	0.78(0.04)	0.24(0.01)	0.32
2.1 Maintains a student-centered learning environment						
I21A	1.00[a] (-)	0.20 (0.01)	0.39	1.00[a] (-)	0.24(0.01)	0.36
I21B	1.15 (0.05)	0.16 (0.01)	0.52	1.18(0.06)	0.15(0.01)	0.53
I21C	0.91 (0.05)	0.15 (0.01)	0.41	0.92(0.06)	0.15(0.01)	0.41
I21D	1.21 (0.05)	0.16 (0.01)	0.55	1.31(0.07)	0.14(0.01)	0.60
I21E	1.04 (0.06)	0.14 (0.01)	0.51	1.08(0.06)	0.14(0.01)	0.51
I21F	1.04 (0.06)	0.13 (0.01)	0.51	1.07(0.06)	0.14(0.01)	0.50
I21G	0.94 (0.04)	0.16 (0.01)	0.41	1.01(0.05)	0.16(0.01)	0.43
I21H	1.00 (0.05)	0.14 (0.01)	0.48	1.05(0.06)	0.15(0.01)	0.48
I21I	0.96 (0.05)	0.15 (0.01)	0.44	1.05(0.07)	0.15(0.01)	0.47
I21J	1.01 (0.04)	0.14 (0.01)	0.49	1.09(0.06)	0.14(0.01)	0.51
I21K	1.00 (0.05)	0.14 (0.01)	0.48	1.10(0.06)	0.15(0.01)	0.50
3.1 Performs professional responsibilities						
I31A	1.00[a] (-)	0.10 (0.01)	0.65	1.00[a] (-)	0.10(0.01)	0.65
I31B	1.12 (0.04)	0.09 (0.01)	0.71	1.12(0.04)	0.10(0.01)	0.71
3.2 Engages in continuous improvement for self and school						
I32A	1.00[a] (-)	0.24 (0.01)	0.29	1.00[a] (-)	0.24(0.01)	0.28
I32B	1.12 (0.04)	0.18 (0.01)	0.40	1.16(0.05)	0.18(0.01)	0.41
I32C	1.16 (0.05)	0.17 (0.01)	0.44	1.22(0.06)	0.16(0.01)	0.46
I32D	1.09 (0.05)	0.16 (0.01)	0.42	1.12(0.06)	0.17(0.01)	0.41
I32E	1.11 (0.06)	0.14 (0.01)	0.45	1.15(0.09)	0.13(0.01)	0.48
I32F	1.04 (0.06)	0.12 (0.01)	0.46	1.09(0.09)	0.13(0.01)	0.46
I32G	0.95 (0.06)	0.13 (0.01)	0.39	0.98(0.09)	0.13(0.01)	0.40

Note. Numbers in parentheses represent the standard error.

[a]Factor loading fixed to 1.0

In order to compare the relative strength of the loadings across the measured variables, the standardized model results were inspected. Standardized factor loadings represent the amount of change in the dependent variable per standard deviation unit of the independent variables (Acock, 2008). The following are the results for the model with all teachers who received a score on the observational score in the district (*N*=6441). Loadings for the first factor (Ability to assess instructional needs) ranged from .67 to .77, for the second factor (Plans and delivers instruction) from .54 to .73, for the third (Maintains a student-centered learning environment) from .62 to .72, the fourth (Performs professional responsibilities) from .81 to .85, and the fifth (Engages in continuous improvement for self and school) from .54 to .68. Factor variances/covariances and correlations for the model can be seen in Table 12. Correlations between the factors ranged from .60 to .92 indicating strong positive correlations between the factors.

The following are the results for the model with all teachers who received a score on the observational score in the district and a VAM score from the state (*N*=2385). Loadings for the first factor (Ability to assess instructional needs) ranged from .68 to .76, for the second factor (Plans and delivers instruction) from .54 to .74, for the third (Maintains a student-centered learning environment) from .60 to .77, the fourth (Performs professional responsibilities) from .81 to .84, and the fifth (Engages in continuous improvement for self and school) from .53 to .69. Factor variances/covariances and correlations for the model can be seen in Table 13. Correlations between the factors ranged from .62 to .93 indicating strong positive correlations between the factors.

Table 12

Factor Variances/Covariance and Correlations for the Five-Factor Model Underlying Administrators' Observational Ratings of Teachers for all Teachers With Observational Rubric Scores

Factor	1.1	1.2	2.1	3.1	3.2
1.1	.18 (0.02)	.92	.86	.60	.87
1.2	0.16 (0.01)	0.18 (0.01)	.92	.62	.87
2.1	0.13 (0.01)	0.14 (0.01)	0.13 (0.01)	.65	.88
3.1	0.11 (0.01)	0.11 (0.01)	0.10 (0.01)	0.19 (0.01)	.66
3.2	0.11 (0.01)	0.11 (0.01)	0.10 (0.01)	0.09 (0.01)	0.10(0.01)

Note. (*N*=6441). Variances are presented as the diagonal elements. Covariances are presented below the diagonal while correlations are presented above the diagonal. Standard errors are in parentheses. The names of the construct are: 1.1 Ability to Assess Instructional Needs; 1.2 Plans and Delivers Instruction; 2.1 Maintains a Student-Centered Learning Environment; 3.1 Performs Professional Responsibilities; 3.2 Engages in Continuous Improvement for Self and School.

Table 13

Factor Variances/Covariance and Correlations for the Five-Factor Model Underlying Administrators' Observational Ratings of Teachers for Sample With Observational and VAM Scores

Factor	1.1	1.2	2.1	3.1	3.2
1.1	.20 (0.02)	.93	.87	.62	.88
1.2	0.18 (0.01)	0.19 (0.01)	.92	.65	.88
2.1	0.14 (0.02)	0.14 (0.01)	0.12 (0.02)	.69	.88
3.1	0.12 (0.01)	0.12 (0.01)	0.10 (0.01)	0.19 (0.02)	.70
3.2	0.12 (0.01)	0.12 (0.01)	0.09 (0.01)	0.09 (0.01)	0.09 (0.01)

Note. (*N*=2385). Variances are presented as the diagonal elements. Covariances are presented below the diagonal while correlations are presented above the diagonal. Standard errors are in parentheses. The names of the construct are: 1.1 Ability to Assess Instructional Needs; 1.2 Plans and Delivers Instruction; 2.1 Maintains a Student-Centered Learning Environment; 3.1 Performs Professional Responsibilities; 3.2 Engages in Continuous Improvement for Self and School.

Modification indices (to see if the models would have better fit if a path, covariance, or correlation were added) were also inspected for each of the models. For both models (the sample with all 6441 teachers and the sample of 2385 teachers with observational rubric scores and VAM scores from the state), two of the resulting modifications made theoretical sense as they were the correlations of the residuals for the items within the same latent construct that asked similar questions. Modification indices revealed that correlating the residuals for item I12H (Using available technology tools and resources to engage students in learning) with the residuals for item I12I (Providing students with opportunities to use technology to support learning) would create a better fitting model with a chi-squared difference of 1,176.35 points for the larger sample and 514.56 points for the smaller sample. Further, the modification indices revealed that correlating the residuals for item I21E (Applying consequences for lack of adherence to rules and procedures) with the residuals for item I21F (Acknowledging adherence to rules and procedures) would also create a better fitting model with a chi-squared difference of 654.32 for the larger sample and 346.26 points for the smaller sample.

Though the suggested changes were plausible theoretically, no post-hoc changes were made to the confirmatory model as fit was determined to be adequate. Regardless, inspection of the model fit was examined, to understand the potential difference in model fit after correlating the errors of the items that were a major source of misfit (items I12H and I12I) for each of the models. The resulting improved fit indices can be seen in Table 14. The correlation of these errors did not make significant changes to the path loadings in the models.

Table 14

Confirmatory Factor Analysis: Fit Indices for the Five-Factor Model Underlying Administrators' Observational Ratings of Teachers for Sample With Only Observational Scores and Sample of Teachers With Observational Score and VAM Scores With Correlated Errors for Items I12H and I12I

Sample	X^2	*df*	CFI	RMSEA	SRMR
Teachers in the district with observational rubric scores (*N*=6441)	6,291.38	516	.930	.042	.037
Teachers in the district with observational rubric scores and VAM scores(*N*=2385)	3,437.01	516	.920	.049	.038

The district uses the sum of the observational rubric indicator scores. For this reason, a second-order CFA was inspected to take into consideration the total score, and not simply each of the subscales of the observational rubric. A second-order latent construct called "Total Score" was created in the model that accounted for the variation in the five first-order factors of the observational rubric. This was completed for both models (the sample with all of the teachers, *N*=6441, and the sample with all of the teachers with observational rubric scores and VAM scores from the state, *N*=2385). The second-order model included the correlated errors of I12H and I12I. Fit indices for both of the second-order CFA models can be seen in Table 15.

Table 15

Second-Order Confirmatory Factor Analysis: Fit Indices for the Total and the Five-Factor Model Underlying Administrators' Observational Ratings of Teachers for Sample With Only Observational Scores and Sample of Teachers With Observational Score and VAM Scores With Correlated Errors for Items I12H and I12I

Sample	X^2	*df*	CFI	RMSEA	SRMR
Teachers in the district with observational rubric scores (*N*=6441)	6,395.15	521	.929	.042	.037
Teachers in the district with observational rubric scores and VAM scores(*N*=2385)	3,506.51	521	.918	.049	.038

The second-order CFA models had adequate fit. The unstandardized factor loadings, residual variances and R^2 for the second-order CFA models can be seen in Table 16. Standardized factor loadings between the total score and each of the five constructs underlying the model for the sample with all teachers in the district with a score on the observational rubric (*N*=6441) ranged from .67 to .97. Standardized factor loadings between the total score and each of the five constructs underlying the model for the sample with all teachers in the district with a score on the observational rubric and a VAM score (*N*=2385) ranged from .70 to .97.

Reliability of the observational rubric. Score reliability can reveal the consistency of a measure. Though the data in this study were clustered (teachers in schools), the focus of the analysis was not on any school-level variables. For this reason, inspection of reliability using Cronbach's alpha (α) was an appropriate technique. Cronbach's alpha coefficients can have values ranging from 0 to 1 (Cronbach, 1951). Alpha values of .7 and higher have been found to be acceptable (Nunnally, 1978). Results of the reliability coefficient of the five factors, as well as for the entire instrument, for each of the samples (all teachers with a score on the observational rubric and teachers with both a score on the rubric and a VAM score from the state) can be seen in Table 17. The resulting alpha coefficients can be categorized as "Good" to "Excellent" for each of the samples for each individual factor as well as for the instrument in its entirety. Table 17 also shows the values of the corrected item-to-total correlations. This value indicates the relationship of the items in the factors with the

Table 16

Second-Order Confirmatory Factor Analysis: Unstandardized Factor Loadings, Residual Variances and R^2 for the Total Score of Administrators' Observational Ratings of Teachers

	All teachers in the District with a score on the observational rubric (*N*=6441)			All teachers in the District with a score on the observational rubric + VAM score (*N*=2385)		
Item on the Rubric	Factor Loading	Residual Variance	R^2	Factor Loading	Residual Variance	R^2
1.1 Ability to assess instructional needs						
I11A	1.00[a] (-)	0.20 (0.01)	0.47	1.00[a] (-)	0.21(0.01)	0.49
I11B	0.97 (0.03)	0.13 (0.01)	0.56	0.94(0.04)	0.14(0.01)	0.56
I11C	1.03 (0.04)	0.13 (0.01)	0.59	0.98(0.04)	0.14(0.01)	0.57
I11D	0.83 (0.04)	0.15 (0.01)	0.45	0.80(0.04)	0.15(0.01)	0.46
I11E	0.85 (0.04)	0.14 (0.01)	0.48	0.82(0.04)	0.15(0.01)	0.47
1.2 Plans and delivers instruction						
I12A	1.00[a] (-)	0.16 (0.01)	0.53	1.00[a] (-)	0.16(0.01)	0.55
I12B	0.97 (0.03)	0.18 (0.01)	0.48	0.93(0.04)	0.20(0.01)	0.45
I12C	1.00 (0.03)	0.20 (0.01)	0.47	0.96(0.04)	0.20(0.01)	0.47
I12D	0.89 (0.03)	0.13 (0.01)	0.52	0.91(0.03)	0.12(0.01)	0.56
I12E	0.91 (0.03)	0.14 (0.01)	0.51	0.93(0.04)	0.13(0.01)	0.55
I12F	0.84 (0.04)	0.16 (0.01)	0.44	0.88(0.04)	0.16(0.01)	0.48
I12G	1.00(0.02)	0.15 (0.01)	0.54	1.01(0.03)	0.15(0.01)	0.56
I12H	0.69 (0.03)	0.24 (0.01)	0.26	0.66(0.04)	0.22(0.01)	0.27
I12I	0.75 (0.04)	0.25 (0.01)	0.28	0.74(0.04)	0.25(0.01)	0.29
2.1 Maintains a student-centered learning environment						
I21A	1.00[a] (-)	0.20 (0.01)	0.39	1.00[a] (-)	0.21(0.01)	0.36
I21B	1.15 (0.05)	0.16 (0.01)	0.52	1.18(0.06)	0.15(0.01)	0.53
I21C	0.91 (0.05)	0.15 (0.01)	0.41	0.92(0.06)	0.15(0.01)	0.41
I21D	1.21 (0.05)	0.16 (0.01)	0.55	1.31(0.07)	0.14(0.01)	0.60
I21E	1.04 (0.06)	0.14 (0.01)	0.51	1.08(0.06)	0.14(0.01)	0.51
I21F	1.03 (0.06)	0.13 (0.01)	0.51	1.06(0.06)	0.14(0.01)	0.50
I21G	0.93 (0.04)	0.16 (0.01)	0.41	1.00(0.05)	0.17(0.01)	0.43
I21H	1.00 (0.05)	0.14 (0.01)	0.47	1.05(0.06)	0.15(0.01)	0.48
I21I	0.96 (0.05)	0.15 (0.01)	0.43	1.05(0.07)	0.15(0.01)	0.47
I21J	1.01 (0.04)	0.13 (0.01)	0.50	1.08(0.06)	0.14(0.01)	0.51
I21K	1.00 (0.05)	0.14 (0.01)	0.48	1.10(0.06)	0.15(0.01)	0.50
3.1 Performs professional responsibilities						
I31A	1.00[a] (-)	0.10 (0.01)	0.65	1.00[a] (-)	0.10(0.01)	0.65
I31B	1.11 (0.04)	0.09 (0.01)	0.71	1.12(0.04)	0.10(0.01)	0.71
3.2 Engages in continuous improvement for self and school						
I32A	1.00[a] (-)	0.24 (0.01)	0.29	1.00[a] (-)	0.24(0.01)	0.28
I32B	1.11 (0.04)	0.18 (0.01)	0.40	1.16(0.05)	0.19(0.01)	0.40
I32C	1.16 (0.05)	0.17 (0.01)	0.44	1.22(0.06)	0.16(0.01)	0.46
I32D	1.09 (0.05)	0.16 (0.01)	0.42	1.12(0.06)	0.17(0.01)	0.41
I32E	1.11 (0.06)	0.14 (0.01)	0.45	1.15(0.09)	0.13(0.01)	0.48
I32F	1.04 (0.06)	0.12 (0.01)	0.46	1.08(0.09)	0.13(0.01)	0.46
I32G	0.95 (0.06)	0.13 (0.01)	0.39	0.98(0.09)	0.13(0.01)	0.40
Total Score						
1.1	1.00[a] (-)	0.02 (0.00)	0.87	1.00[a] (-)	0.03(0.01)	0.87
1.2	1.04 (0.04)	0.01 (0.00)	0.93	1.00(0.04)	0.01(0.00)	0.93
2.1	0.87 (0.04)	0.01 (0.00)	0.90	0.80(0.04)	0.01(0.00)	0.90
3.1	0.74 (0.04)	0.10 (0.01)	0.45	0.73(0.05)	0.09(0.01)	0.49
3.2	0.73 (0.03)	0.02 (0.00)	0.84	0.68(0.04)	0.01(0.00)	0.86

Note. Numbers in parentheses represent the standard error. [a]Factor loading fixed to 1.0

Table 17

Summary of all Cronbach Alphas by Scales and Total for the Observational Rubric Completed by Administrators by Sample of All Teachers in the District as Well as Teachers With a Score on the Observational Rubric and a VAM Score From the State

Sample	Factors	# of Items in the Scale	Cronbach Alpha	Range of values of corrected item-to-total correlation	*N*
Teachers in the District with a score on the Observational Rubric	1.1 Ability to assess instructional needs	5	.84	.61-.68	6538
	1.2 Plans and delivers instruction	9	.88	.53-.67	6435
	2.1 Maintains a student-centered learning environment	11	.91	.58-.70	6434
	3.1 Performs professional responsibilities	2	.81	.69-.69	6435
	3.2 Engages in continuous improvement for self and school	7	.82	.51-.60	6434
	Entire Instrument	34	.96	.49-.70	6433
Teachers in the District with a score on the Observational Rubric and a VAM score from the state	1.1 Ability to assess instructional needs	5	.84	.61-.67	2384
	1.2 Plans and delivers instruction	9	.88	.53-.69	2382
	2.1 Maintains a student-centered learning environment	11	.91	.56-.73	2381
	3.1 Performs professional responsibilities	2	.81	.68-.68	2382
	3.2 Engages in continuous improvement for self and school	7	.83	.49-.61	2381
	Entire Instrument	34	.96	.49-.73	2381

summed score for all other items. An industry rule of thumb is to have at least a .40 value for this correlation. All values in the reliability were within acceptable ranges.

Research Question Two

Question two addressed the relationship between the scores from the VAM and the observational rubric. Only traditional schools who had at least nine observations per school were included in the sample of schools (e.g., Non charter or special need schools were excluded). The total number of schools was 104 with 2385 cases. To answer this question both VAM scores with the standard error applied and VAM scores without the standard error applied were modeled to better understand the relationship. Table 18 depicts the skeweness and kurtosis values of the VAM data with and without the standard error applied as well as histograms in Figures 11 and 12 representing the distribution for each of the variables. The correlation between the VAM scores with and without the SE was .51.

Table 18

Descriptive Statistics for the Two Types of VAM Scores Used

Indicator	VAM + SE	VAM without SE
N	2385	2385
Mean	0.35	-0.06
Median	0.25	-0.05
SD	0.46	0.28
Skewness	3.54	-0.91
Kurtosis	19.96	22.80
Range	5.89	5.90
Minimum Value	-0.81	-3.85
Maximum Value	5.07	2.05

Note. SE= Standard error; VAM + SE = VAM+(SE*1.96). *SD* = Standard deviation

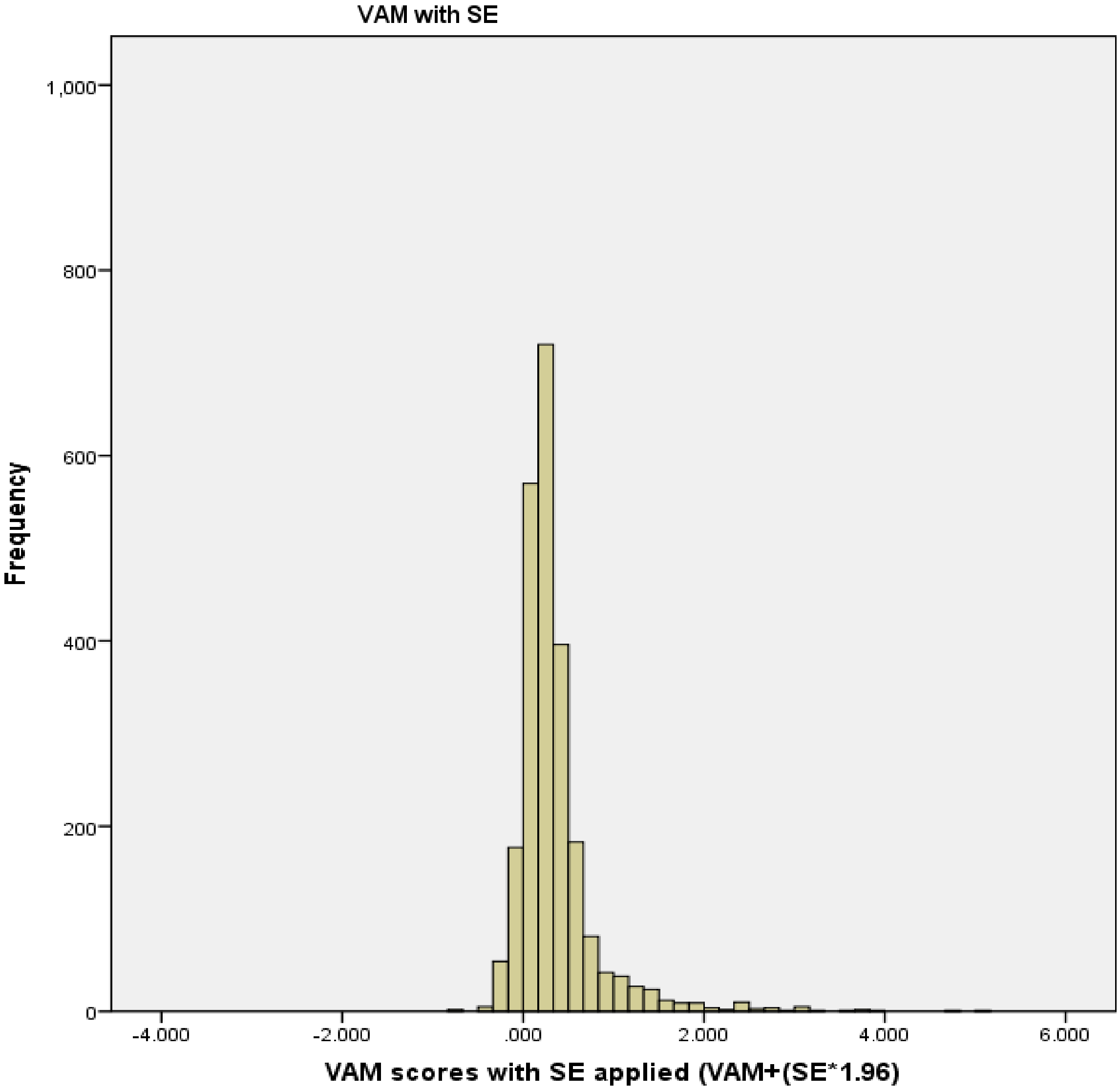


Figure 11. Histogram of VAM scores with standard error applied.

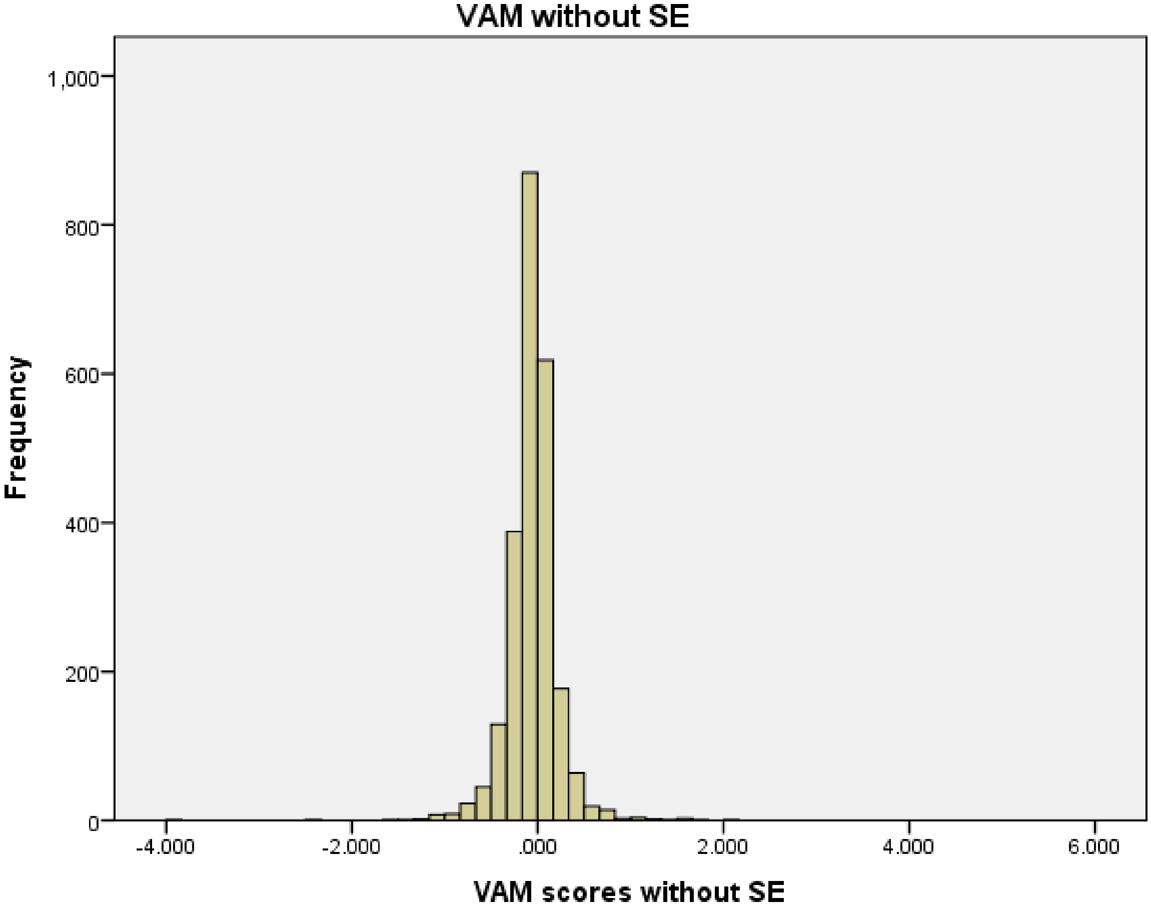


Figure 12. Histogram of VAM scores without standard error applied.

For the first model, the VAM scores analyzed included the standard error of measurement. Since VAM scores delivered by the State of Florida to the distrit contained a score representing the standard error by case, the final VAM score used for the analysis was computed at the top of the band of the 95% confidence interval, VAM=VAM+(SE*1.96). To further support the findings as presented, and because VAM scores could be calculated in several different ways, the first model was replicated using the VAM scores as presented to the district (no standard error applied).

Application of VAM scores using these two methods (95% confidence band and original VAM score with no standard error applied) against the observational rubric was used to evaluate the sensitivity of the relationship between the two variables in terms of how the VAM score was calculated.

The general model for the observational rubric consisted of five factors that were scaled by fixing the first item loading to 1.0 using the Mplus version 5.21 software while the remaining factor variances/covariances, factor loadings, and residual parameters were freely estimated (Muthén & Muthén, 1998-2007). The defaults of the program were not changed leaving the error covariances set to zero (with the assumption that there should be no correlations between the error variances). MLR estimation with robust standard errors was also used because it is robust to non-normal data, missing data, and non-independence of observations (Muthén & Muthén, 1998-2007).

The VAM variable with standard error applied, and then the VAM variable without the standard error applied, were added to the five-factor CFA model estimated for research question one. As stated previously, these CFAs took into account the nested data structure (Raudenbush, 1995; Raudenbush, Rowan, & Kang, 1991). The clustering variable used in this study was the teachers' school. Results for the fit of both models can be seen in Table 19.

Using the same criteria for the estimation of fit for this question as was used for question one, results indicated that both models had relatively adequate fit. Though the chi squared was statistically significant, which indicates misfit for both of the models, the

Table 19

Fit Indices for the Model: Observational Rubric With VAM Scores With and Without Standard Error (SE)

Sample	X^2	*df*	CFI	RMSEA	SRMR
Five-factor model With VAM with SE applied	4084.06	546	.903	.052	.039
Five-factor model With VAM without SE applied	4100.99	546	.903	.052	.039

Note. SE= Standard Error; VAM=VAM+(SE*1.96).

large sample size is likely contributing to this result. For this reason, other measures of fit were also inspected. For both models, the CFI of .903 indicated an acceptable fit (Bentler & Bonett, 1980), and the RMSEA of .052 and the SRMR of .039 were below the accepted cut off values thus indicating acceptable fit (Browne & Cudeck, 1993; Hu & Bentler, 1999).

Modification indices were inspected for each of the models used to answer the second research question. For each of the models, modification indices involved the same pair of items (I12H and I12I) as was determined in the previous models. The chi-squared difference for the model using VAM scores with the standard error applied would result in an improvement in model fit of 517.39 points while for the model using VAM scores without the standard error applied, an improved fit of 521.40 points. Again, no post-hoc modifications were made to either model.

Correlations between all of the factors underlying the observational rubric and both VAM scores were inspected. The results of the correlations indicate that though the correlations between the VAM without the standard error and the factors underlying the observational rubric scores were stronger than the VAM scores with the standard error applied, the correlations would still be classified as small for both versions of VAM. The correlations can be seen in Table 20 for each of the VAM scores with the five factors underlying the observational rubric as well as the correlation of the VAM scores to each other.

To understand if there were differences in these correlations by school level, the same analysis was replicated separating the sample even further into elementary schools, middle schools and high schools. The fit indices found for the three new subsets were similar to those for the entire sample. Results of the correlations by level and the number of teachers represented in each of the samples can be seen in Table 21.

Table 20

Correlations for the Five Factors Underlying the Administrators' Observational Ratings of Teachers and VAM Scores With and Without SE Applied

Observational Scale	VAMS with SE Correlation	VAMS without SE Correlation
1.1	.05	.16
1.2	.06	.18
2.1	.09	.18
3.1	.05	.15
3.2	.06	.14

Note. The names of the observational scales are: 1.1 Ability to Assess Instructional Needs; 1.2 Plans and Delivers Instruction; 2.1 Maintains a Student-Centered Learning Environment; 3.1 Performs Professional Responsibilities; 3.2 Engages in Continuous Improvement for Self and School. *N*=2385. SE=Standard error.

Table 21

Correlations for the Five Factors Underlying the Administrators' Observational Ratings of Teachers and VAM Scores With and Without SE Applied by School Level

	Elementary (*n*=1056)		Middle (*n*=671)		High (*n*=609)	
	Correlation		Correlation		Correlation	
Observational Scale	VAM with SE	VAM no SE	VAM with SE	VAM no SE	VAM with SE	VAM no SE
1.1	.13	.25	-.07	.05	.07	.23
1.2	.13	.26	-.05	.08	.06	.22
2.1	.16	.27	-.01	.08	.13	.20
3.1	.04	.17	.02	.08	.10	.20
3.2	.08	.21	-.06	.04	.11	.20
Total (Second-Order)	.14	.27	-.04	.07	.10	.23

Note. The names of the observational scales are: 1.1 Ability to Assess Instructional Needs; 1.2 Plans and Delivers Instruction; 2.1 Maintains a Student-Centered Learning Environment; 3.1 Performs Professional Responsibilities; 3.2 Engages in Continuous Improvement for Self and School. *N*=2385. SE=Standard error.

To determine if the method in which the standard error was applied had an effect on the correlations, new value-added scores were calculated using the lower end of the confidence interval. If these correlations were much stronger than the scores for the upper end of the confidence interval, this may be an indication that the method of applying the standard error to the scores had an effect on the relationship. This new score was computed by subtracting the standard error of each score from the provided VAM score, VAM=VAM-(SE*1.96).

The mean for the new variable was -0.458 with a standard deviation of 0.512. The correlation between the VAM score at the lower end of the confidence interval with VAM score without the standard error was .626, and the correlation between the VAM score at the lower end of the confidence interval with the VAM with the standard error at the top end of the confidence interval was -.357. The fit indices for the first order CFA model demonstrated appropriate fit, $\chi^2(546, N=2385) = 4108.225$, p<.001 (CFI=.903; RMSEA=.052; SRMR=.039), as did the indices for the second-order model, $\chi^2(554, N=2385) = 3599.830$, p<.001 (CFI=.917; RMSEA=.048; SRMR=.038). Results of the correlations for the VAM scores using the lower end of the confidence interval can be seen in Table 22.

Because the district uses the results of the observational rubric as a total score, and not as individual constructs, it was also important to understand the relationship between the VAM scores with and without the standard error and the Total score on the observational rubric. This was accomplished using a second-order CFA where the five underlying constructs made up the second-order latent construct called “Total Score.” In order to obtain the best fitting models, a correlated error term of I12H and I12I was added to the models. The fit indices for these new models can be seen in Table 23. The correlation between the total score and VAM scores with the standard error applied was .07 while the

Table 22

Correlations for the Five Factors Underlying the Administrators' Observational Ratings of Teachers and VAM Scores With SE Applied as the Lower end of the Confidence Band

Observational Scale	Correlation
1.1 Ability to Assess Instructional Needs	.129
1.2 Plans and Delivers Instruction	.142
2.1 Maintains a Student-Centered Learning Environment	.112
3.1 Performs Professional Responsibilities	.116
3.2 Engages in Continuous Improvement for Self and School	.098
Total (Second-Order)	.131

Note. N=2385. SE=Standard error.

Table 23

Fit Indices for the Second-Order Model: Total Score for the Observational Rubric With VAM Scores With and Without SE

Sample	X^2	*df*	CFI	RMSEA	SRMR
One second-order factor and five first-order factors and VAM with SE applied	3,566.12	554	.917	.048	.038
One second-order factor and five first-order factors and VAM without SE applied	3,590.94	554	.917	.048	.038

Note. SE= Standard Error; VAM=VAM+(SE*1.96).

correlation between the total score and the VAM scores without the standard error was .18. These correlations are consistent with the first-order confirmatory factor analysis results.

In order to visually understand the relationship between the total score and each of the VAM scores (with and without the SE applied), scatterplots were created. Figure 13 shows the relationship between each of the VAM scores with the total score on the observational rubric.

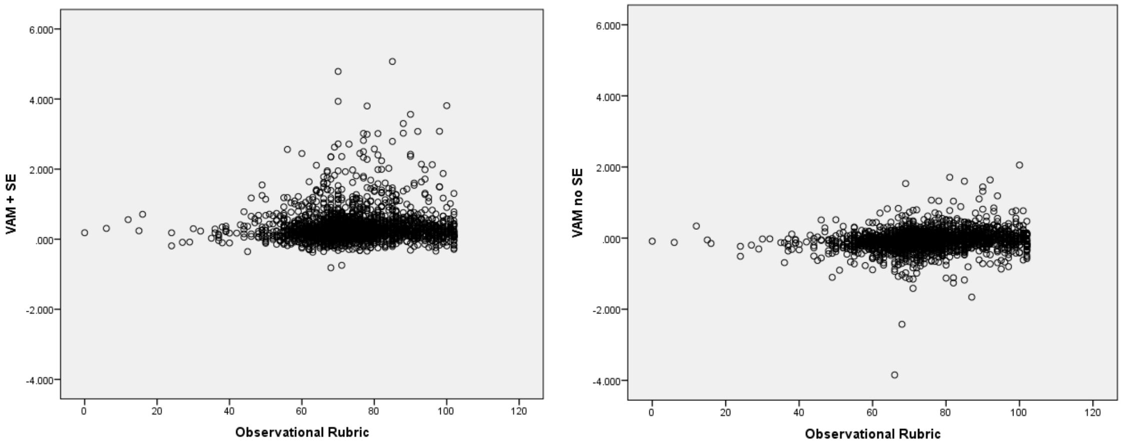


Figure 13. Scatterplot of VAM scores with total score on the administrative review.

One last attempt was made to investigate the relationship between the VAM scores and the scores from the observational rubric and ensure that the results of the findings were an actual representation of the relationship and not due to the estimation methods. This approach involved analyzing by school, the correlations between the VAM scores and the scores from the observational rubric (a composite for each of the five factors as well as the total). The goal was to investigate if the relationship between the VAM and observational scores varied between schools.

Using SPSS 21, the file was split into the 104 schools and a correlation was calculated between the VAM scores with the standard error applied (upper end), VAM scores without the standard error applied, and the composite scores of each of the five observational factors as well as the total composite score. Results demonstrated that there were many differences between schools in how the VAM variables correlated with the factors of the observational rubric and the instrument as a whole. Though the majority of the correlations were relatively weak across most schools, this was not the case for all of the schools. Maximum and minimum values for the correlations can be seen in Table 24 and stem-and-leaf plots depicting all of the correlations between the factors and VAM with and without the standard error applied can be seen in Figure 14 and Figure 15. A summary of the correlations can be seen in Table 25. Further, correlations for the schools by school level (Elementary, Middle, or High) between the five factors and each of the VAM scores (with and without the standard error applied) can be seen in Appendix C.

Table 24

Maximum and Minimum Correlations for VAM and Observational Scores

Observational Scale	VAM Original Correlation	VAM with SE Correlation
Factor 1.1	-.874 to .832	-.833 to .755
Factor 1.2	-.535 to .850	-.696 to .786
Factor 2.1	-.494 to .897	-.693 to .853
Factor 3.1	-.651 to .742	-.603 to .697
Factor 3.2	-.627 to .852	-.521 to .735
Total Instrument	-.757 to .908	-.843 to .836

Note. SE=standard error; VAM with SE = VAM + (1.96*SE). The names of the observational scales are: 1.1 Ability to Assess Instructional Needs; 1.2 Plans and Delivers Instruction; 2.1 Maintains a Student-Centered Learning Environment; 3.1 Performs Professional Responsibilities; 3.2 Engages in Continuous Improvement for Self and School. *N*=2385. Number of schools=104. Composite score calculated by summing the scores of the items in each construct and the instrument as a whole.

Factor 1.1- VAM Score Without SE Applied

Stem	Leaf
- 1.0	
- 0.9	
- 0.8	7
- 0.7	
- 0.6	
- 0.5	
- 0.4	
- 0.3	1 2 3
- 0.2	0 0 1 4 8
- 0.1	2 4 5 8
- 0.0	1 1 2 2 3 4 4 4 4 5 5 8
0.0	0 0 1 1 1 1 2 2 2 2 2 2 4 5 6 7 7 8 8 9
0.1	0 0 0 1 2 3 3 3 4 4 6 7 7 8 9
0.2	0 0 0 0 2 2 2 2 3 4 4 5 5 6 6 6 8 8 9 9 9 9
0.3	2 3 4 6 6 8 8
0.4	0 1 2 2 2 5
0.5	1 1 1 5 6 8
0.6	4
0.7	
0.8	2 3
0.9	
1.0	

Factor 1.2- VAM Score Without SE Applied

Stem	Leaf
- 1.0	
- 0.9	
- 0.8	
- 0.7	
- 0.6	
- 0.5	4
- 0.4	2
- 0.3	
- 0.2	2 2 2 2 3 4 4 6
- 0.1	1 1 3 5 7 9 9
- 0.0	1 1 2 3 3 4 6 6 7 9 9 9 9
0.0	0 1 3 4 4 5 7 8 9
0.1	0 1 2 3 3 5 5 5 6 6 7 7 7 9 9
0.2	0 0 1 1 1 2 3 3 4 4 5 6 7 7 7 7 7 7 8 8
0.3	1 1 2 3 4 4 5 6 7 8 8 9 9
0.4	0 0 0 2 4 6 7 7 8
0.5	9
0.6	1 2 6 8
0.7	1 4
0.8	5
0.9	
1.0	

Factor 2.1- VAM Score Without SE Applied

Stem	Leaf
- 1.0	
- 0.9	
- 0.8	
- 0.7	
- 0.6	
- 0.5	
- 0.4	4 9
- 0.3	0 6
- 0.2	0 1 2 4 6
- 0.1	1 4 4 5 6
- 0.0	1 2 2 2 3 3 3 4 6 7 7 8 8
0.0	0 2 2 2 3 3 4 5 5 5 5 6 8 8
0.1	0 0 0 0 1 3 3 3 4 4 4 5 5 5 5 5 5 6 6 6 6 7 8 8
0.2	0 0 1 3 4 7 7 9 9
0.3	0 0 0 2 2 2 2 3 3 5
0.4	0 0 1 2 4 4 5 7
0.5	0 1 5 7
0.6	0 1 3 9
0.7	1 4
0.8	1
0.9	0
1.0	

Factor 3.1- VAM Score Without SE Applied

Stem	Leaf
- 1.0	
- 0.9	
- 0.8	
- 0.7	
- 0.6	5
- 0.5	
- 0.4	6
- 0.3	3 3 5 6
- 0.2	0 2 2 3 4 4 9
- 0.1	0 0 0 1 3 4 7 8 9 9
- 0.0	1 1 3 4 6 6 6 7 8
0.0	0 1 1 1 1 3 3 3 3 4 5 6 6 6 8 8 8 9
0.1	0 0 0 1 2 3 3 8 9
0.2	0 1 1 1 1 2 2 3 4 5 5 5 5 5 7 7 7 8 8 9
0.3	0 1 3 3 4 5 5 5 5 6 7
0.4	0 1 5 6 8
0.5	0 0 3
0.6	6 7
0.7	3 4
0.8	
0.9	
1.0	

Factor 3.2- VAM Score Without SE Applied

Stem	Leaf
- 1.0	
- 0.9	
- 0.8	
- 0.7	
- 0.6	3
- 0.5	3
- 0.4	0 6
- 0.3	0 9
- 0.2	2 4 5 6
- 0.1	0 0 0 0 1 1 1 4 5 6 7 9
- 0.0	2 4 4 5 5 5 6 6 8 9 9 9
0.0	0 1 1 2 3 3 4 5 5 6 6 6 7 8 8 9
0.1	0 1 1 2 2 4 4 4 5 6 7 7 8 8 9 9
0.2	0 2 2 3 3 3 4 4 4 4 4 5 6 7 8
0.3	1 1 2 5 6 6 6
0.4	0 2 2 4 6 8 8 9
0.5	0 2 8
0.6	6 6
0.7	7 9
0.8	5
0.9	
1.0	

Total Score- VAM Score Without SE Applied

Stem	Leaf
- 1.0	
- 0.9	
- 0.8	
- 0.7	6
- 0.6	
- 0.5	
- 0.4	0
- 0.3	9
- 0.2	0 0 3 4 5 7 9
- 0.1	5 7 8 9 9
- 0.0	2 2 3 4 5 5 6 7 7 8 8 8 9
0.0	3 3 3 3 4 4 5 7 7 8 8 9 9 9
0.1	0 0 2 4 4 4 4 5 5 5 6 7 8 9 9
0.2	0 0 1 1 2 2 4 4 5 5 5 6 6 7 8 9
0.3	0 0 0 3 3 4 6 6 9 9
0.4	0 1 1 1 2 3 3 7 8
0.5	0 1 4 5 6 7
0.6	1
0.7	3 4 4 9
0.8	
0.9	1
1.0	

Figure 14. Stem-and-leaf plot of correlations between observational scores and VAM scores without the standard error applied by school.
SE=Standard Error.

Factor 1.1 - VAM score with Standard Error Applied

Stem	Leaf
- 1.0	
- 0.9	
- 0.8	3
- 0.7	
- 0.6	
- 0.5	
- 0.4	3 6
- 0.3	0 2 8 8 9
- 0.2	0 1 4 6 7 7
- 0.1	0 0 0 1 3 4 4 5 5
- 0.0	1 1 1 1 3 3 5 6 7 7 8 9
0.0	0 0 1 1 1 1 1 2 2 3 3 3 3 3 4 4 5 5 5 5 7 7 8 8 8 9 9 9
0.1	0 0 0 0 0 1 1 1 2 3 3 6 6 7 8 9
0.2	1 1 2 3 3 4 8 8
0.3	0 0 1 3 4 4
0.4	1 1 3 4 6 8
0.5	2 4
0.6	1
0.7	3 6
0.8	
0.9	
1.0	

Factor 1.2 -VAM score with Standard Error Applied

Stem	Leaf
- 1.0	
- 0.9	
- 0.8	
- 0.7	0
- 0.6	0
- 0.5	5
- 0.4	2 5
- 0.3	0 2 3 6 6
- 0.2	0 1 3 4 5 6 6 6 7 7 8
- 0.1	0 0 2 4 5 7 7 9 9 9
- 0.0	1 1 1 2 4 6 6 6 6 7 8 8
0.0	0 0 0 0 1 2 3 3 3 5 5 6 6 6 7 7 9 9
0.1	0 0 0 0 1 1 2 3 3 4 7 8 8 8 9 9
0.2	1 1 2 3 4 9 9
0.3	0 0 2 4 5 5 6 6 8
0.4	2 3 3 3 4
0.5	2 3 3
0.6	2 6
0.7	9
0.8	
0.9	
1.0	

Factor 2.1 -VAM score with Standard Error Applied

Stem	Leaf
- 1.0	
- 0.9	
- 0.8	
- 0.7	
- 0.6	7 9
- 0.5	3
- 0.4	4
- 0.3	0 3
- 0.2	1 1 2 5 7 7
- 0.1	0 1 1 2 2 3 4 5 9 9 9
- 0.0	1 2 3 3 3 3 3 4 6 7 7
0.0	1 2 2 2 3 3 3 4 5 5 5 5 6 7 7 8 8 8 8
0.1	0 0 0 1 2 2 4 4 4 5 5 5 5 6 7 7 7 9
0.2	0 1 1 2 3 3 4 4 4 7 8 8
0.3	0 0 1 4 6 6 7
0.4	0 2 3 5 6 6
0.5	0 1 3 5
0.6	3 4
0.7	7
0.8	5
0.9	
1.0	

Factor 3.1-VAM score with Standard Error Applied

Stem	Leaf
- 1.0	
- 0.9	
- 0.8	
- 0.7	
- 0.6	0
- 0.5	7
- 0.4	0 2 5 5 9
- 0.3	3 5 6
- 0.2	0 0 1 3 4 6
- 0.1	1 1 2 2 2 3 4 4 6 6 6 7 8 8 8
- 0.0	1 2 2 4 4 5 5 7 7 7 7 8 8 8
0.0	0 0 1 1 2 3 3 4 4 6 8 9
0.1	0 0 1 2 2 3 3 3 4 5 6 6 7 8 8 9 9
0.2	0 1 1 1 1 3 4 4 5 6 6 6 7 7 8 9 9
0.3	0 9
0.4	0 0 4 5 9
0.5	2 6
0.6	8
0.7	0
0.8	
0.9	
1.0	

Factor 3.2 - VAM score with Standard Error Applied

Stem	Leaf
- 1.0	
- 0.9	
- 0.8	
- 0.7	
- 0.6	
- 0.5	0 2
- 0.4	2 4
- 0.3	1 1 1 2 4 4 5
- 0.2	0 0 0 2 3 4 7 8
- 0.1	0 2 3 3 4 5 6 6 7 8 8 9 9
- 0.0	1 1 1 3 3 4 4 7 7 9
0.0	0 1 1 2 3 3 3 3 4 4 4 4 5 5 5 6 7 8 8 8 9
0.1	0 0 1 1 3 4 5 5 6 6 8 8 8 9
0.2	1 2 2 2 5 5 5 7 7 8 8 8 9
0.3	0 4 4 5 6 8
0.4	5 7 8
0.5	1
0.6	9
0.7	1 2 4
0.8	
0.9	
1.0	

Total Score -VAM score with Standard Error Applied

Stem	Leaf
- 1.0	
- 0.9	
- 0.8	4
- 0.7	
- 0.6	
- 0.5	1 8
- 0.4	1
- 0.3	0 0 1 2
- 0.2	0 0 0 1 1 3 4 4 5 7 7 8 9
- 0.1	0 1 1 1 5 5 6 6 8
- 0.0	1 1 1 3 3 4 5 6 6 6 7
0.0	2 2 3 3 4 5 5 6 6 7 7 7 9 9
0.1	0 0 1 1 3 4 4 5 5 6 6 6 6 6 6 7 7 8 9 9
0.2	0 2 2 3 3 4 5 5 5 9
0.3	0 1 3 3 4 4 7
0.4	0 5 5 6 6 7 8
0.5	6
0.6	1
0.7	2 3
0.8	4
0.9	
1.0	

Figure 15. Stem-and-leaf plot of correlations between observational scores and VAM scores with the standard error applied by school.
SE=Standard Error.

Table 25

Number and Percentage of Schools With Strong Positive or Negative Correlations Between Observational and VAM Scores

	Strong Positive		Strong Negative		Positive		Negative	
	VAM Without SE	VAM With SE	VAM Without SE	VAM With SE	VAM Without SE	VAM With SE	VAM Without SE	VAM With SE
Factor	# (%)	# (%)	# (%)	# (%)	# (%)	# (%)	# (%)	# (%)
1.1	9 (8.65%)	5 (4.81%)	1 (0.96%)	1 (0.96%)	79 (75.96%)	69 (66.35%)	25 (24.04%)	35 (33.65%)
1.2	8 (7.69%)	6 (5.77%)	1 (0.96%)	3 (2.88%)	74 (71.15%)	61 (58.65%)	30 (28.85%)	43 (41.35%)
2.1	12 (11.54%)	8 (7.69%)	0 (0.00%)	3 (2.88%)	77 (74.04%)	70 (67.31%)	27 (25.96%)	53 (50.96%)
3.1	7 (6.86%)	4 (3.92%)	1 (0.98%)	2 (1.96%)	70 (68.63%)	57 (55.88%)	32 (31.37%)	45 (44.12%)
3.2	8 (7.69%)	5 (4.81%)	2 (1.92%)	2 (1.92%)	70 (67.31%)	62 (59.62%)	34 (32.69%)	42 (40.38%)
Total	12 (11.54%)	5 (4.81%)	1 (0.96%)	3 (2.88%)	76 (73.08%)	62 (59.62%)	28 (26.92%)	41 (39.42%)

Note. *N*=104 schools. Strong Correlations (| *r* | > .50). SE=standard error; VAM with SE = VAM + (1.96*SE). The names of the observational scales (factors) are: 1.1 Ability to Assess Instructional Needs; 1.2 Plans and Delivers Instruction; 2.1 Maintains a Student-Centered Learning Environment; 3.1 Performs Professional Responsibilities; 3.2 Engages in Continuous Improvement for Self and School. *N*=2385. Composite score calculated by summing the scores of the items in each construct and the instrument as a whole.

Correlations between VAM scores with the standard error applied and observational measures by school. Out of the 104 schools in the sample, 41 of the schools had negative correlations between the VAM with the standard error applied and the total score for the observational instrument (correlations ranged from -.01 to -.84). There were 62 schools with positive correlations for the total score of the observational rubric with the VAM score with the standard error applied (correlations ranged from .02 to .84). Out of the same 104 schools in the sample, correlations within five of the schools were strong and positive while three had strong negative correlations ($| r | > .50$). For both positive and negative correlations, a mix of strong and weak correlations can be seen.

Composite score of Factor 1.1 (Ability to Assess Instructional Needs) had 69 schools with a positive correlation and 35 schools with a negative correlation. Out of those schools, five had a strong positive and one had a strong negative correlation ($| r | > .50$).

Composite score of Factor 1.2 (Plans and Delivers Instruction) had 61 schools with a positive correlation and 43 schools with a negative correlation. Out of those schools, six had a strong positive and three had strong negative correlations ($| r | > .50$).

Composite score of Factor 2.1 (Maintains a Student-Centered Learning Environment) had 70 schools with a positive correlation and 34 schools with a negative correlation. Out of those schools, eight had a strong positive and three had strong negative correlations ($| r | > .50$).

Composite score of Factor 3.1 (Performs Professional Responsibilities) had 57 schools with a positive correlation and 45 schools with a negative correlation. Out of those schools, four had a strong positive and two had strong negative correlations (| *r* | > .50). Two of the schools had no variance for this factor and correlations could not be calculated.

Composite score of Factor 3.2 (Engages in Continuous Improvement for Self and School) had 62 schools with a positive correlation and 42 schools with a negative correlation. Out of those schools, five had a strong positive and two had strong negative correlations (| *r* | > .50).

Correlations between VAM scores without the standard error applied and observational measures by school. Of the 104 schools in the sample, 28 schools had negative correlations between the total score of the observational rubric and the VAM score with no standard error applied (correlations ranging from -.02 to -.76). Positive correlations of the total score on the observational rubric with VAM score without the standard error applied ranged from .03 to .91. From the same sample, 12 schools had strong positive correlations and one school had a strong negative correlation (| *r* | > .50). For both positive and negative correlations, a mix of strong and weak correlations can be seen.

Composite score of Factor 1.1 (Ability to Assess Instructional Needs) had 79 schools with a positive correlation and 25 schools with a negative correlation. Out of those schools, nine had strong positive correlations and one had a strong negative correlation (| *r* | > .50).

Composite score of Factor 1.2 (Plans and Delivers Instruction) had 74 schools with a positive correlation and 30 schools with a negative correlation. Out of those schools, eight had a strong positive correlation and one had a strong negative correlation (| *r* | > .50).

Composite score of Factor 2.1 (Maintains a Student-Centered Learning Environment) had 77 schools with a positive correlation and 27 schools with a negative correlation. Out of those schools, 12 had a strong positive correlation and none had a strong negative correlation (| *r* | > .50).

Composite score of Factor 3.1 (Performs Professional Responsibilities) had 70 schools with a positive correlation and 32 schools with a negative correlation. Out of those schools, seven had a strong positive correlation and one had a strong negative correlation (| *r* | > .50). Two of the schools had zero variance and the correlation could not be calculated.

Composite score of Factor 3.2 (Engages in Continuous Improvement for Self and School) had 70 schools with a positive correlation and 34 schools with a negative correlation. Out of those schools, eight had a strong positive correlation and two had a strong negative correlation (| *r* | > .50).

The relationship between the VAM scores and the observational rubric was analyzed in several different ways. The results of all three methods led to the same conclusion. The relationship between VAM scores and the observational rubric was relatively weak. Established guidelines suggest that a correlation of about .60 or greater shows strong evidence for convergent validity (Hill, Kapitula, & Umland, 2011) and these values were not met in any of the three analyses.

Research Question Three

The third research question focused on how the VAM and observational scores related to other theoretically relevant teacher variables and not to other variables that they should theoretically not relate to. In theory, both VAM and the observational rubric scores should relate in the same fashion to variables that measure the same construct and to those that are completely unrelated to teacher effectiveness. There were several predictor variables used in this study. Binary variables included: National Board Certification, Race/Ethnicity (Multiethnic, Hispanic/Latino ethnicity, American Indian, Asian, Hawaiian/Pacific Islander, Black, White, multiracial) and Gender (Female coded as 1). The non-binary variable used in this study was the years of experience of a teacher. The dependent variables were the five factors of the observational rubric and VAM scores. The VAM scores used in this part of the analysis took into account the standard error at the 95% confidence interval, VAM=VAM + (SE*1.96).

The same data set used to answer research question two was again used to answer question three. The model for the observational rubric consisted of five factors which were scaled by fixing the first item loading to 1.0 using the Mplus version 5.21 software while the remaining factor variances/covariances, factor loadings, and residual estimates were freely estimated (Muthén & Muthén, 1998-2007).

The defaults of the program were not changed leaving the error covariances set to zero (with the assumption that there should be no correlations amongst the error variances). Mplus MLR maximum likelihood estimation with robust standard errors was used for the analysis of this question because of robustness to non-normal data, missing data, and non-independence of observations (Muthén & Muthén, 1998-2007).

Because the categories of race were not mutually exclusive (a person could identify himself/herself using multiple categories), the data were recoded so that each person was in only one racial category (any person who marked more than one race was coded as multi-racial). Further, because of the small sample of individuals who were Hawaiian/Pacific Islander and American Indian, these two categories were joined into one. In the models, the race of White was used as the reference category for the other races.

To calculate the years of employment of a teacher, taking into consideration the potential ceiling effect encountered after a certain amount of years in the field, the variable years of experience of a teacher (with values from 0 to 40 years) was transformed by squaring the variable and then including this quadratic component into the equation (years of experience^2).

Due to estimation problems resulting from the magnitude of the years of teaching experience variable, this variable was transformed using the mean of the variable (*M*=11.79 years). The transformed variable was equal to (Years Teaching Experience-11.79). This transformed variable ranged from -11.79 to 28.21, and the squared variable ranged from .04 to 795.80.

Descriptive statistics for the predictor variables used to answer question three can be seen in Table 26. The sample size, means of the variables, standard deviations and normality values for the indicators of the observational rubric as well as the VAM scores did not change from those reported for question two. Independent variables presented in Table 26 depict whether or not they are binary (1=Yes, 0=No) and the number of participants who said they belonged to the particular group.

Results of the fit statistics for the model relating the predictor variables to the observational rubric can be seen in Table 27. The model comparing the predictor variables to the VAM scores does not have fit statistics as all variables were measured or observed variables (no latent variables). Using the same criteria for this question as was

Table 26

Descriptive Statistics for Item Characteristics Used in This Study

Measurable Indicator	*n*	*M*	*SD*	Skeweness	Kurtosis
Years of Experience		11.79	8.75	0.92	0.07
Years of Experience-mean		0.00	8.75	0.92	0.07
(Years of Experience-mean)^2		76.60	110.07	3.00	10.99
National Board Certification (1=Yes, 0=No)	140	0.06	0.24		
Hispanic/Latino (1=Yes, 0=No)	95	0.04	0.20		
Multi Ethnicity (1=Yes, 0=No)	19	0.01	0.09		
Asian (1=Yes, 0=No)	28	0.01	0.11		
Hawaiian/Pacific Islander OR American Indian (1=Yes, 0=No)	15	0.00	0.05		
Black (1=Yes, 0=No)	201	0.08	0.28		
White (1=Yes, 0=No)	2122	0.89	0.31		
Gender Binary (Female=1)	1994	0.84	0.37		

Note. N=2385

Table 27

Fit Indices for Predictor Variables for the Observational Rubric

Model	X^2	*df*	CFI	RMSEA	SRMR
Predictor variables for Observational rubric scores	4,786.54	778	.897	.046	.034

Note. N=2385.

used for questions one and two, it can be established that the fit of the model for the observational rubric scores was adequate and demonstrated appropriate model fit. Though the chi squared was statistically significant which indicates misfit, the large sample size is likely contributing to that. For this reason other fit indices were inspected. The CFI, RMSEA and SRMR were all within their individual acceptable cut off values (Bentler & Bonett, 1980; Browne & Cudeck, 1993; Hu & Bentler, 1999). In general, it was concluded that this model had adequate fit. No post-hoc modifications were made to either model.

Each of the models was calculated independently of each other, one model only looking at the relationships between the predictor variables and VAM scores, while the other model looked at the relationship between the predictor variables and the scores on the observational rubric. The results of the models can be viewed in parallel to see the relationships between the predictor variables and (a): the five factors of the observational rubric, and (b) VAM scores. Theoretically, some of the predictor variables were expected to have positive relationships with VAM and the observational rubric, while others were expected to have no relationship. In all cases, it was hypothesized that since

VAM scores and the observational rubric theoretically represent the same construct of effective teaching, the relationships should be similar between each of them and the predictor variables.

Observational rubric scores with predictor variables. Standardized regression coefficients for the predictors can be seen in Table 28 along with the correlations of the predictor variables (see Table 29). The coefficients presented are the standardized coefficients in order to be able to judge the differences in paths between the variables. The R^2 values, representing the percent of the variance that can be explained by the predictors, were relatively small (Factor 1.1: .045; Factor 1.2: .042; Factor 2.1: .032; Factor 3.1: .032; Factor 3.2: .042).

Years of experience (calculated by subtracting the mean) was statistically significant across all five factors and had a positive effect on the observational rubric scores. The quadratic effect ([year of experience-Mean]**2) was negative across all five factors and statistically significant across 4 out of the 5 factors. Given that the coefficient for the squared years of experience was negative across all five factors, the quadratic equation represents one of diminishing returns as time goes on. This is to say that in general, after accounting for the ceiling effect, the more experience a teacher has, the higher the scores on the rubric, with decreasing effectiveness.

The next variable inspected was National Board Certification. Though not statistically significant across all five factors of the rubric (four of the five were statistically significant), in general possessing National Board Certification had a small

Table 28

Standardized Factor Loadings for the Model With the Observational Rubric

Item on the Rubric	Factor Loading
1.1 Ability to assess instructional needs	
I11A	.702(.018)*
I11B	.746(.020)*
I11C	.757(.019)*
I11D	.680(.026)*
I11E	.683(.027)*
1.2 Plans and delivers instruction	
I12A	.736(.016)*
I12B	.670(.024)*
I12C	.682(.023)*
I12D	.742(.019)*
I12E	.736(.020)*
I12F	.687(.023)*
I12G	.744(.014)*
I12H	.544(.028)*
I12I	.567(.025)*
2.1 Maintains a student-centered learning environment	
I21A	.600(.027)*
I21B	.726(.018)*
I21C	.637(.028)*
I21D	.772(.014)*
I21E	.716(.021)*
I21F	.705(.022)*
I21G	.653(.025)*
I21H	.692(.020)*
I21I	.688(.020)*
I21J	.713(.022)*
I21K	.703(.023)*
3.1 Performs professional responsibilities	
I31A	.803(.022)*
I31B	.844(.017)*
3.2 Engages in continuous improvement for self and school	
I32A	.530(.026)*
I32B	.636(.020)*
I32C	.681(.021)*
I32D	.640(.023)*
I32E	.694(.028)*
I32F	.678(.029)*
I32G	.634(.030)*

Note. * Indicates statistically significant loadings (*p*<.05). *N*=2385. Numbers in parentheses represent the standard error.

Table 29

Standardized Regression Coefficients (Beta) for the Predictor Variables of the Observational Rubric Factors

	Factor 1.1		Factor 1.2		Factor 2.1		Factor 3.1		Factor 3.2	
Variable	*r*	Beta Coefficient	*r*	Beta Coefficient	*r*	Beta Coefficient	*r*	Beta Coefficient	*r*	Beta Coefficient
Years of Experience-Mean	.14	.183(.036)*	.13	.166(.035)*	.13	.167(.034)*	.14	.137(.034)*	.10	.144(.032)*
Years of (Experience-Mean) ^{2}	.03	-.098(.031)*	.02	-.093(.031)*	.03	-.079(.032)*	.08	-.015(.030)	.01	-.089(.032)*
Gender (Female=1)	.10	.082(.029)*	.10	.088(.026)*	.09	.082(.028)*	.07	.057(.024)*	.10	.091(.027)*
National Board Certification	.09	.056(.022)*	.11	.077(.021)*	.08	.042(.021)*	.08	.047(.030)	.08	.055(.023)*
Hispanic/Latino	.00	.010(.019)	-.01	.000(.021)	.01	.018(.022)	-.02	-.013(.023)	.01	.019(.023)
White	.08	Reference Group	.06	Reference Group	.02	Reference Group	.05	Reference Group	.04	Reference Group
Black	-.07	-.070(.036)	-.05	-.042(.032)	-.01	-.008(.031)	-.04	-.037(.024)	-.04	-.029(.029)
Hawaiian/Pacific Islander OR American Indian	-.06	-.065(.019)*	-.05	-.051(.017)*	-.04	-.043(.019)*	-.02	-.015(.025)	-.04	-.040(.013)*
Asian	-.00	.005(.023)	.01	.021(.023)	.01	.018(.022)	.00	.009(.024)	.02	.030(.025)
Multi Race	-.01	-.004(.021)	-.01	-.009(.017)	-.02	-.011(.023)	-.03	-.025).031)	-.02	-.019(.023)

Note. *N*=2385. *= statistically significant (*p*<.05). The names of the observational scales are: 1.1 Ability to Assess Instructional Needs; 1.2 Plans and Delivers Instruction; 2.1 Maintains a Student-Centered Learning Environment; 3.1 Performs Professional Responsibilities; 3.2 Engages in Continuous Improvement for Self and School. *r*= correlation of indicator variables with the observational rubric. Numbers in parentheses represent the standard error.

positive effect on the scores from the observational rubric. These results match the hypothesis presented for variables that are meant to signify correlates of quality teaching.

For both models, it was also hypothesized that several predictor variables (race/ethnicity and gender) would not have any relationship with either the observational rubric or VAM scores. For all of the race/ethnicity predictors, across the majority of the five factors in the observational rubric, these indicators were not statistically significant. Hispanic/Latino ethnicity and Asian, Black, and multi-racial were not statistically significant across all five factors. Hawaiian/Pacifica Islander or American Indian were statistically significant across four of the five factors and had a negative effect as compared to White teachers. In general, when the predictor variables were statistically significant, the relationships were usually relatively weak (positive or negative) on the five underlying factors of the observational rubric. This result matched the hypothesis that race/ethnicity should have a weak relationship with the effectiveness of a teacher.

Being female was found to be a statistically significant predictor across all five factors underlying the observational rubric. For all five indicators, female teachers had slightly higher observational scores than male teachers. This finding was not what was predicted in the hypothesis as gender was not expected to relate to more effective teaching.

Correlation coefficients can be interpreted as effect sizes. Cohen's (1992) guidelines indicate that an effect size can demonstrate the strength of the relationship between two variables, where .10 can be considered small, .25 medium and anything larger than .40 can be considered large where a "medium effect size represents an effect likely to be visible to the naked eye of the careful observer" (p. 156). The correlations between each of the predictor variables and each of the factors on the observational rubric can be classified as small to medium. Specifically, the variables of year of experience, gender, and National Board Certification have medium effect sizes.

Because the district uses the results of the observational rubric as a total score, and not as individual constructs, it was also important to understand the relationship between the predictor variables and the observational rubric as a total score. This was accomplished using a second-order CFA where the five underlying constructs made up the latent construct called "Total" which represents the total score on the observational rubric. In order to get the best fitting model, a correlated error term between I12H and I12I was added to the model. The fit indices for the new model can be seen in Table 30.

The results of the model for the predictor variables of the total score can be seen in Table 31. Results of the relationship between the predictor variables and the second-order total score for the observational rubric demonstrated the same pattern as for each of

Table 30

Fit Indices for the Second-Order Model with Predictors of the Total Score for the Observational Rubric

Model	X^2	*df*	CFI	RMSEA	SRMR
Second-order model of total observational score with predictor variables	4,280.58	818	.911	.042	.033

Note. *N*=2385.

Table 31

Standardized Regression Coefficients for the Predictor Variables of the Total Score from the Observational Rubric

Variable	Total Score Beta Coefficient
Years of Experience-Mean	.177 (.033)*
Years of Experience-Mean ^{2}	-.090 (.031)*
Gender (Female=1)	.091 (.027)*
National Board Certification	.061 (.021)*
Hispanic/Latino	.010 (.020)
Black	-.037 (.032)
Hawaiian/Pacific Islander OR American Indian	-.051 (.015)*
Asian	.019 (.022)
Multi-Racial	-.013 (.021)

Note. *N*=2385. *= statistically significant (*p*<.05). Numbers in parenthesis represent the standard error.

the individual factors underlying the observational rubric. White was the reference category.

VAM scores with predictor variables. The second model looked at the relationships between the predictor variables and VAM scores. The R^2 for the VAM scores was very small (.004) stating that less than 1% of the variance can be explained by this variable. Inspection of the standardized regression coefficients (Table 32) revealed that none of the predictor indicators (years of experience, years of experience quadratic, National Board Certification, multi-racial, gender, Hispanic/Latino ethnicity, and the races of Asian, Hawaiian/Pacific Islander or American Indian, Black or White) were significantly related to the VAM scores. This was not the case for the observational rubric which had relationships with the predictor variables similar to what was hypothesized.

Table 32

Standardized Regression Coefficients for the Model With the VAM Scores and Predictor Variables

Predictor Variable	*r*	Beta Coefficient
Years of Experience-Mean	.01	.027 (.031)
(Years of Experience-Mean)^2	.00	-.015 (.028)
Gender (Female=1)	-.01	-.008 (.019)
National Board Certification	-.01	-.011 (.023)
Hispanic/Latino	.03	.033 (.020)
White	-.03	Reference Group
Black	.02	.019 (.034)
Hawaiian/Pacific Islander OR American Indian	.01	.006 (.019)
Asian	-.01	-.009 (.018)
Multi Race	.05	.048 (.027)

Note. * Indicates statistically significant loadings (*p*<.05). *N*=2385. *r*= correlation of indicator variables with VAM. Numbers in parentheses represent the standard error.

It was also hypothesized that several predictor variables (race/ethnicity and gender) would not have any relationship with either the observational rubric or VAM scores. As previously stated, none of the indicators (not race, ethnicity or gender) had a statistically significant relationship with the VAM scores.

The correlation between the VAM scores and each of the variables can be interpreted as an effect size with the guidelines previously stated. The effect sizes for VAM scores were small. There were no effects that could be visible to the careful observer.

Chapter Five: Discussion

The purpose of this study was to examine the validity of value-added model (VAM) scores for use in teacher evaluations. This chapter presents a summary of the study and results, discussion, implications of this study, and finally recommendations for future research.

Summary of the Study

Research has shown that teachers have a strong influence on student achievement (e.g., Rowan, Correnti, & Miller, 2002). The importance of this influence has brought about increased focus on the ability to properly evaluate teachers' performance with regard to the educational effect teachers have on students. One of the goals of this focus is on properly identifying effective teachers.

Efforts to appropriately identify effective classroom teachers have been made at the Federal, State, and District levels. At the Federal level, incentives have included pay-for-performance plans (with the intent to pay more effective teachers higher salaries than less effective teachers). At the State level, laws have been passed mandating certain aspects be included in a teacher evaluation, including the use of value-added modeling data in teacher evaluations (Senate Bill 0736, n.d.). At the District level, observational rubrics, based on research, have been implemented, which aim at appropriately identifying effective teaching (Danielson, 2011; Marzano, 2007). The aim of all of these initiatives is to identify teachers who best perform their job and have a positive educational effect on the students they instruct.

The State of Florida has laws that require certain components be included in a teacher evaluation, but they also provide relative freedom in what indicators should be placed in teacher observational rubrics. Many districts across the state use different combinations of research-based indicators in their rubrics. State laws also stipulate that scores from the value-added models, created and approved by the Commissioner of Education, be used as a part of teacher evaluations. The school year 2011-2012 was the first year that VAM scores were used in teacher evaluations across the State of Florida.

Because of the high-stakes decisions that are to be made from VAM scores, and teacher evaluations as a whole (teachers can get incentive pay or be let go), evidence of the validity and appropriateness of these VAM scores is imperative. The purpose of this study was to evaluate how value-added scores relate to accepted empirical evidence of effective teaching in order to provide evidence to support or question the use of value-added scores in teacher evaluations. Data for this study consisted of teacher evaluations based on an observational rubric and VAM scores from teachers in a large southeastern Florida district.

Prior to examining the relation between the VAM and observational scores as a way of evaluating the convergent validity of the VAM scores, it was necessary to examine the psychometric properties of the observational rubric. Exploring the fit of the factor structure of the observational rubric model prior to any analysis with VAM scores was an important first step since the school district had been given the freedom to create the rubric and no prior factor analyses had been conducted on the measure. Once the observational rubric was inspected for statistical appropriateness, it was then correlated to the VAM scores as a measure of convergent validity.

A second source of validity evidence for the VAM scores were the correlations between teacher variables that were hypothesized to be related to effective teaching (e.g., National Board Certification status) and those teacher variables that were hypothesized to be unrelated to effective teaching (e.g., teacher gender). These hypothesized relationships formed a nomological network (Cronbach & Meehl, 1955) that was used to evaluate the construct validity of the VAM scores. These relationships also were examined for the observational rubric scores.

The data were analyzed using Mplus 5.21 to account for the nested structure of the data (teachers were nested within schools). Maximum likelihood estimation with robust standard errors (MLR) was used to account for missing data and non-normality of the data. The fit of the models as well as the strength of the factor loadings and correlations between the variables (e.g., dimensions of the observational rubric and total score with VAM scores) were analyzed to answer each of the questions.

Discussion of the Results

Question One. The first research question was analyzed in two parts. The first part inspected the extent to which the teachers' scores from the observational rubric were consistent with the five factors underlying the model, while the second part inspected the internal consistency reliability of the scores from the instrument. Confirmatory factor analyses of the observational rubric scores were conducted using the entire sample of teachers in the district who received a score on the observational rubric (*N*=6441) and the sample that included only the teachers who received a score on the observational rubric and a VAM score from the state (*N*=2385). The models, which were evaluated using multiple measures of fit, indicated that the five-factor model fit the data appropriately for the entire sample size of teachers receiving a score on the observational rubric.

Standardized factor loadings were strong with Factor 1 ("Ability to assess instructional needs") loadings ranging from .67 to .77; Factor 2 ("Plans and delivers instruction") ranging from .54 to .73; Factor 3 ("Maintains a student-centered learning environment") ranging from .62 to .72; Factor 4 ("Performs professional responsibilities") ranging from .81 to .85; and Factor 5 ("Engages in continuous improvement for self and school") ranging from .54 to .68. These results provide preliminary evidence of the factorial validity of the observational rubric instrument.

Correlations between the factors ranged from .60 to .92 indicating strong positive correlations between the factors. The strong correlations (.92) between two pairs of factors ("Ability to assess instructional needs" with "Plans and delivers instruction" and "Plans and delivers instruction" with "Maintains a student-centered learning environment") suggest that these factors shared considerable variance and have limited discriminant validity.

Plans and delivers instruction contains items such as, "What do I do to plan and organize for effective instruction?" and "What do I do to establish and communicate learning goals?" which, if done successfully, would indicate success in the "ability to assess instructional needs" and "maintain a student-centered learning environment."

Conceptually, these strong correlations make sense given that for a teacher to plan and deliver instruction it would be necessary for the teacher to assess the instructional needs of students. Further, in order to maintain a student-centered learning environment, planning is essential.

Similar fit indices for the CFA model were obtained using the sample of teachers receiving a score on the observational rubric and a VAM score from the state. Standardized factor loadings were strong: Factor 1: “Ability to assess instructional needs” ranged from .68 to .76; Factor 2: “Plans and delivers instruction” ranged from .54 to .74; Factor 3: “Maintains a student-centered learning environment” ranged from .60 to .77; Factor 4: “Performs professional responsibilities” ranged from .81 to .84; and Factor 5: “Engages in continuous improvement for self and school” ranged from .53 to .69. Correlations between the factors ranged from .62 to .93 indicating strong positive correlations between the factors.

Comparable results questioning discriminant validity were also found with this sample. The strong correlations of .93 between two of the factors (“Ability to assess instructional needs” with “Plans and delivers instruction”) and .92 between two of the factors (“Plans and delivers instruction” with “Maintains a student-centered learning environment”) would suggest limited discriminant validity. Brown (2006) indicates that correlations between factors higher than .80 to .85 may be an indication of weak discriminant validity.

Even though factor analyses have not been previously conducted on this particular observational rubric, researchers who have conducted exploratory factor analyses (EFA) and confirmatory factor analyses (CFA) on similar observational measures have found similar results (high correlations specifically with the planning factor) in terms of model fit and limited discriminant validity (e.g., Sabo & Lawton, 2013). Observer error in the form of a response set, such as the halo effect (i.e., an observer forms an early impression of the teacher that influences ratings on other dimensions), may play a role in the limited discriminant validity of the five-factor observational measure. Training observers to be aware of this observational error and other types of observational errors (e.g., error of central tendency, observer drift, observer contamination by outside data) may result in improved discriminant validity of this observational measure.

As Guilford (1946) notes, inspection of both reliability and validity is important in evaluating the psychometric properties of a measure. The second part of the first research question looked at the internal consistency reliability of the instrument. Reliability indicators were calculated using both the entire sample, which included all teachers in the district receiving a score on the observational rubric, as well as the sample of teachers receiving both a VAM score from the state and a score on the observational rubric.

Reliabilities for each of the factors, as well as the instrument as a whole for the sample with all of the teachers receiving a score on the observational rubric, were deemed satisfactory (factor alphas of .84, .88, .91, .81, and .82, and for the entire instrument .96). Similar reliabilities for each of the factors, as well as the instrument as a whole were obtained for the sample of teachers with both observational rubric and VAM scores from the state (factor alphas of .84, .88, .91, .81, and .83, and for the entire instrument .96). Evidence of the internal consistency reliability of the observational rubric scores was strong.

Although internal consistency reliability (e.g., Cronbach alpha) is widely used with educational and social science measures, a Cronbach alpha may not be the most appropriate or the most informative measure to use in understanding the reliability of scores from an observational instrument. A Cronbach alpha does not measure the variability in teaching behaviors between days, between lessons, between observers, or between observations. A more appropriate approach, which was not feasible for the present study, would have been to use generalizability theory (GT) to analyze the multiple sources of measurement error that may affect the reliability of the observational scores. GT could reveal different sources of information that could reveal a clearer picture of how well the measurement system as a whole is working. This argument is supported by Hill, Charalambos, and Kraft (2012) who argued that the use of generalizability theory (using multiple raters, during several observations, and rating several teachers) can lead to more reliable scores and can also provide evidence of the appropriate number of facets (raters/teachers/occasions) that should be used to obtain desired levels of reliability (e.g., > .90). For example, to achieve reliabilities greater than .90 it may be necessary to observe on more than one occasion as was the case in the present study.

This argument is further supported by the research conducted by the Measuring Effective Teachers Project (2013), which was funded by the Bill and Melinda Gates Foundation. This three year study focused on a number of issues related to measures of effective teaching including the creation and validation of observational measures of teachers. In this study, researchers investigated the best combinations of several measurement facets, which included number of lessons to observe, number of observers, and time spent observing, in order to achieve the best reliability for the observation (Bill & Melinda Gates Foundation, 2013). A single measure of internal consistency, such as a Cronbach alpha, is not sufficient to identify the many other aspects that are critical to a valid and reliable observation system.

Question Two. The second question focused on the relationship between VAM scores and the scores from the observational rubric. The sample size used to answer this question only included teachers who had a score on the observational rubric as well as VAM scores in reading, mathematics, or both (combined score). This question built upon the five-factor CFA model analyzed in question one by adding the VAM scores and examining the relation between these scores and the five-factors from the observational rubric. Given that VAM scores can be utilized in many different formats, this question was answered using two models analyzed in parallel: the first time with VAM scores with the standard error applied, the second time with VAM scores without the standard error applied.

Fit of the CFA model with the inclusion of the VAM scores was acceptable. Results of the relationship between the observational rubric and VAM scores with the standard error applied at the 95% confidence interval showed low positive correlations to the factors underlying the observational rubric (correlations of .054, .059, .088, .045, and .055, respectively for each factor with VAM scores). The weak correlations between the

VAM scores and the observational rubric scores can lead to questions about the validity of the VAM scores, the validity of the observational rubric scores, or both of these measures since these measures are used to identify effective teachers, and thus should theoretically be moderately to highly correlated (teachers with higher scores on one measure should also have higher scores on the other). The low correlations between the two indicators of teacher effectiveness indicated that the scores did not have a linear relation. This weak correlation raises questions for the VAM and observational scores and their appropriateness in making high-stakes decisions.

To evaluate the sensitivity of this correlation to different scoring methods for the VAM scores, the same model was analyzed using the original VAM scores as provided by the state, without any application of the standard error. Fit of the statistical model was acceptable. The resulting relationship from this model between the five factors from the observational rubric and the VAM scores with no standard error applied to them were low and positive but stronger than the relationships of the VAM scores with the standard errors applied (correlations per factor of .164, .181, .178, .145, and .136). These weak correlations again call for caution in the use of the VAM scores for teacher evaluations. The two measures did not correlate as expected.

To see if the weak correlations were due in part to the method in which the standard error was applied or the educational level of the schools, two more attempts to inspect the strength of the correlation between the VAM and the observational rubric scores were attempted. The model fit indices for these two attempts were similar to the previous indices and indicated appropriate fit. Correlations were not found to be any stronger when the results were analyzed by level (elementary, middle, high) or when the lower end of the confidence band (VAM-SE*1.96) was used. These results indicated that the weak correlations were not related to grade level or to the method in which the standard error was applied to the VAM scores.

Another attempt to understand the relationship between these two scores (VAM and observational rubric scores) was made by comparing the correlations within each school in the sample. Correlations were calculated between the VAM scores with the standard error applied as well as the VAM scores without any the standard error for each of the five factors of the observational rubric as well as the total score. A few schools had very strong and positive correlations between the two variables, while other schools had strong negative correlations. The majority of the schools in the study had very weak correlations, either positive or negative.

Established practices (Hill, Kapitula, & Umland, 2011) suggest that a correlation of about .60 or greater shows strong evidence for convergent validity, and this criterion was used in this study. The resulting correlations for the five factors and total score from the observational rubric with the VAM scores in this study did not meet this criterion. This result demonstrated that there was no strong evidence of convergent validity between VAM scores, with or without the standard error applied, and the observational rubric scores. It is of note that when the correlations between the VAM scores and the total observational rubric score were inspected by school, six schools out of 102 showed evidence of strong convergent validity with VAM without the SE applied, while five showed evidence of strong convergent validity with VAM with the SE applied.

Additional research is needed to understand why the relation between the VAM and observational scores was strong in these few schools.

It is important to note that the majority of the schools had weak correlations between the VAM scores in either format and the factors underlying the observational rubric. This further supports the findings of the previous two correlational analyses that the relationships between VAM scores and scores from the observational rubric are weak. Since both of these measures are utilized in identifying effective teachers, results call for caution in the application of VAM scores for high-stakes decision making until additional research can support their use.

The results of this research, attempted through several analyses, were relatively consistent across all methods. This indicated no to very low correlations between VAM scores (with and without the SE applied) and the observational rubric. Though these results were robust as each analysis provided a similar outcome, the reasons why there was little relation between the two scores were unclear. One potential reason for the low correlation between these two measures is that the measurement model underlying teacher quality may be a formative measurement model rather than a reflective measurement model (Edwards, 2011). In a formative measurement model indicators such as the scores from the observational rubric and the VAM scores are viewed as *causes* of the latent construct of teacher quality. These indicators represent distinct aspects of the construct of teacher quality and because of this distinctness may not necessarily correlate with each other. In contrast, with a reflective measurement model, indicators such as the scores from the observational rubric and the VAM scores are viewed as the *effects* of the latent construct of teacher quality and therefore according to this model these indicators should correlate. These alternative measurement models represent different conceptualizations of teaching quality and the decision to use one over the other is complex that needs to be made based on statistical and theoretical criteria. These different measurement models will need to be part of the discussion as researchers strive to define teaching quality and develop meaningful ways to measure this construct. Another possibility for the low correlation between the VAM and observational scores is that there may be large amounts of random error in the observational measure that attenuated the relation between the observational and VAM scores.

Observations, because they require human judgment, have the capability to introduce large amounts of error into a score. Observers must be trained in order to reduce the effects of measurement error. These effects can include the personal bias of the observer, the desire to rate the majority of the participants on the high end of the scale, the tendency for an observer's initial impression of a person to carry into subsequent observations, and the tendency to rate all individuals around the midpoint of the scale (Gall, Borg, & Gall, 1996). These sources of error could affect the observational scores used in this study, thus attenuating the relationship between VAM scores and the observational rubric.

Whatever the reasons are for the low correlations between the VAM and observational scores, the results from the present study are consistent with those from other research studies that examined the relationship between scores from different forms of VAM and different observational rubric. For example, Milanowski (2004) found correlations between VAM scores based on reading and observational scores between .03 to .45, in mathematics between .20 to .56, and in science between -.01 to .33, with a sample size of 212 teachers; the correlational analyses consisted of 16 to 55 teachers depending on grade and subject. Correlations aggregated by grade in reading were .32 (95% confidence interval = .18 to .45), mathematics was .43 (95% confidence interval= .29 to .55), and science was .27 (95% confidence interval = .09 to .46) (Milanowski, 2004). Though some of the correlations in Milanowski's study, as compared with this study, were slightly stronger (they were mostly still considered weak), the sample sizes used to determine these correlations were much smaller than those used in the present study.

The study by Kimball, White, and Milanowski (2004) also had similar results (teacher *N*=328) showing very weak to weak correlations between VAM and observational scores (3rd grade reading and mathematics *r*=.10; 4th grade reading *r*=.28; 4th mathematics *r*=.07; 5th grade reading *r*=.28; 5th grade mathematics *r*=.37). Some of the correlations by grade were slightly higher than those found in the present research, but they were not sufficiently robust to provide evidence of convergent validity.

Gallagher (2004) found one moderate and several weak correlations (teacher *N*=34) between an observational rubric and VAM scores (reading *r*=.50, mathematics *r*=.21, language arts *r*=.18, composite *r*=.36). The small sample size calls for caution with the interpretation of these correlations. Regardless, these relationships were relatively weak in nature and did not provide strong convergent validity evidence.

All of these studies together have demonstrated that there is a general lack of relationship between VAM scores and the scores from various observational rubrics. Previous research has generally been conducted with smaller sample sizes of teachers, while this study was based on a much larger sample size providing more robust results. This research, supported by previous research, suggests caution when using VAM scores for high-stakes decision making.

Question Three. The third question focused on the relationship between VAM scores and the scores from the observational rubric (using each of the five factors and the total score) as dependent variables and several theoretically relevant variables as predictor variables. These analyses were part of the nomological network and examined the relationship between the independent variables of National Board Certification, years of employment, race/ethnicity (Hispanic/Latino, American Indian, Asian, Hawaiian/Pacific Islander, Black, White) and gender and each of the following two dependent variables: VAM scores and observational rubric scores. The VAM score used to answer this question included the standard error at the upper 95% confidence interval. Years of employment considered the ceiling effect found in research and was included as a quadratic term (years^2) in the regression equation.

It was hypothesized that possession of National Board Certification and years of employment (considering the ceiling effect) would have positive effects on both the VAM scores and the observational rubric scores. It was also hypothesized that race/ethnicity, and gender would not be related to either the VAM or the observational rubric scores. The standardized regression coefficients between the dependent variables

(VAM and observational rubric scores) and the predictor variables showed that none of the covariates had a statistically significant relation to VAM scores. This did not match the hypothesis because it was expected that there would be a positive relationship between the dependent variable of VAM scores and the following two independent variables: National Board Certification status and years of experience. This finding, which calls into question the validity of the VAM scores, again calls for caution in the use of VAM scores for teacher evaluations in a high-stakes context.

The standardized coefficients of the predictors with the observational rubric scores behaved much more as predicted. Years of experience was positive and statistically significantly related to the observational rubric factors (standardized loadings = .183, .166, .167, .137, and .144, respectively). The quadratic portion of years of experience was statistically significant for four of the five factors and proved to be a negative coefficient further supporting the ceiling effect discussed in previous research (standardized coefficients = -.098, -.093, -.079, -.015 [not significant], and -.089). National Board Certification also matched the hypothesis by having a positive relationship to the observational rubric scores in all cases and being statistically significant in four of the five cases (standardized coefficients = .056, .077, .042, .047 [not significant], and .055).

In general, the two predictors of effective teaching had the expected positive relationship with the scores from the observational rubric. These two predictor variables were not statistically significant on the observational factor that was composed of only two indicators (Performs Professional Responsibilities).

It was also predicted that race/ethnicity would not have a relationship with the scores on the observational instrument. As demonstrated in the results, most race/ethnicity categories did not have statistically significant relations to any factor of the observational rubric. Hispanic/Latino, Black, Asian, and Multi-racial race/ethnicities were not statistically significant predictors for any of the factors. This matches the hypothesized relationship between the factors of the observational rubric and the predictor of race/ethnicity.

The last predictor of gender was hypothesized to have no relationship with the scores on the observational rubric. In contrast to what was expected, gender was significantly related to all five factors of the observational rubric (gender coded as female=1 had standardized coefficients of .082, .088, .082, .057, .091). This means that female teachers had higher scores on the observational rubric. These effects were statistically significant, but they were not large in magnitude.

Based on the analyses guided by the nomological network, there is little support for the validity of the VAM scores. In contrast, there was some support for the observational scores based on relations with several variables. These variables include National Board Certification, years of experience including the ceiling effect, and the majority of the race/ethnicity categories.

In view of the fact that the observational rubric is mainly used in the district as a total score and not as individual subscale scores, all analyses involving the observational rubric were rerun adding a second-order factor to obtain a total score. This was replicated for all questions in the study and all comparisons between the observational rubric and other variables (VAM or predictor variables such as gender, race/ethnicity, etc.). The results of the second-order CFA were consistent with the results of the first-order model for each of the analyses in this research. This provided evidence of the robustness of the results from this study.

Conclusion

In conclusion, the observational rubric performed well in this study. The rubric had appropriate model fit indicating that the measured variables loaded properly on their respective factors. Internal consistency reliability of the scores from the observational rubric was also acceptable. Lastly, the scores from the observational rubric generally had the expected relationships with the predictor variables, thus providing support for the validity of the observational scores.

On the other hand, VAM scores did not perform well statistically. VAM scores were not statistically significantly related to the indicators of quality teaching used in this study. The non-statistically significant relationships were weak in nature.

When both scores, VAM and the observational rubric, were compared to each other in an attempt to determine the correlation between the two, VAM scores had low to no relationships with the observational rubric scores. This was the case across several different analytic approaches, which included different applications of VAM scores, separation of VAM scores by educational level (elementary, middle and high school) and inspection of the relationships within each of the schools. Given that the correlation between the two scores were very low, and that scores from the observational rubric functioned appropriately and had the expected relationships with predictor variables while VAM scores did not, the results from this study call for caution in the use of VAM scores for high-stakes decision-making.

Implications of the Study

The results of this validity study indicate that caution should be taken in the use of VAM scores for teacher evaluations, especially for high-stakes decision-making. The 2011-2012 school year was the first year the VAM scores were used as part of teacher evaluations. This study began the process of providing information related to the validity of these scores.

Implications of this research include the reconsideration of teacher observation systems. Teaching is complex and so is correctly identifying quality and/or effective teachers. Observation systems need to be able to provide valid and reliable evidence regarding teachers. Because of this, observation systems need to be carefully inspected and should include the best combination of raters and number of time points to make appropriate evaluation decisions.

Based on the results of this study, districts in the state of Florida should consider using the VAM scores at the minimum percentage allowable by law of an overall teacher's evaluation until more evidence can be provided to support that these scores measure what they purport to measure. If more validity evidence is gathered which supports the use of VAM scores, teacher evaluations might then include a higher percentage of points coming from the scores of these models.

Currently there is a movement in Florida to remove these models from teacher evaluations. For example, the Florida Education Association has filed a lawsuit against the State of Florida Department of Education stating that SB736 is unconstitutional as this bill, in part, takes away the rights of teachers to bargain concerning their evaluations (Robinson, et al. v. Robinson, 2011). Another lawsuit (Peek, Weatherstone and Florida Education Association v. Florida State Board of Education and Florida Department of Education, 2012) has stated that the value-added formula is unlawful as it was never adopted appropriately by rule. The most recent lawsuit filed to date, which includes as one of the plaintiffs the recipient of the "teacher of the year" award from Hernando County, is challenging the VAM scores used in teacher evaluations (Cook et al. v. Bennett et al., 2013), stating that in some district plans teachers' scores are sometimes not derived from the students they actually teach (e.g., a district may apply the school-wide VAM score to an art teacher who does not have an individual teacher VAM score). One potential implication of this research is that the results could be used to support litigation in the controversy over value-added models.

Recommendations for Future Research

Future research should begin with replicating this study with different observational rubrics from different districts across the state of Florida for comparison with their teachers' VAM scores. Additional research would continue the process initiated by this research to create a clearer picture of the validity of VAM scores across different districts that, in turn, have different observational rubrics. If the same results can be found when VAM scores are compared to different observational rubrics, this would provide more evidence to recommend caution in the use of these scores for high-stakes decision-making. If, on the other hand, evidence of a strong correlation between

VAM scores and scores from an observational rubric from another district was found, this would provide evidence that the large southeastern school district where this study was based should reassess its observational rubric.

Researchers can, in future studies, focus on schools that had very strong positive correlations and those that had very strong negative correlations between VAM scores and the observational rubric. This research can be qualitative in nature with interviews and focus groups to understand why some schools have strong positive correlations while others have strong negative correlations. These differences could be related to characteristics of the administrator in the school, the school culture, student demographics within a school, or a variety of other possible reasons.

To investigate why some schools had very strong positive or negative correlations within schools, this study could be replicated using a multilevel statistical model (i.e., two-level), with predictor variables at the school level. These variables could include school SES, school demographic characteristics, or other such school-type variables. Using school-level variables might identify which variables are related to the strength of the relation between the VAM scores and scores from the observational rubric.

Future research could also look at the unexpected findings in this study, such as the relation between gender and scores on the observational rubric, using qualitative, quantitative, or mixed methods. Results of this study indicated that females had higher scores on the observational rubric. The results also demonstrated that Hawaiian/Pacific Islander or American Indian teachers received a lower score on their observational rubric. Additional psychometric analyses of the observational rubric that include examining differential item functioning (DIF) or measurement invariance by teacher gender and race/ethnicity are needed to identify potential biases in the observational measure. If allowed, future studies could investigate the actual VAM model from the State of Florida and the scores from the state as a whole. Assumptions underlying the model need to be examined along with any patterns of misfit in the models. This type of research would allow transparency with the value-added models, and the scores produced from them.

A longitudinal study replicating this same analysis with the use of subsequent year VAM scores would also provide more information. This study could reveal if VAM scores begin to provide positive validity evidence for their use in high-stakes decisions. Further, a longitudinal study might reveal trends on the VAM scores that could not be identified in this cross-sectional study.

Although the observational rubric used in this study demonstrated good model fit based on the confirmatory factor analyses and adequate internal consistency reliability, these statistical tests do not evaluate inter-observer reliability or the consistency over time of the teachers' ratings. Future studies need to provide more rigorous tests of the psychometric qualities of the observational rubric. Generalizability theory is one approach that could be used to evaluate the multiple sources of error (e.g., observer, occasion, item, subject matter, school level) that may impact the measurement system.

Considering the complexity involved in teaching, one summative observation and one formative observation, as used in this research, may not be sufficient to capture the true essence of a particular teacher. Also, one rater may not provide the evidence needed as the observer may not be as accurate as usual on a particular day, or may interpret an indicator slightly differently on a particular day. For this reason, different ways to inspect the observational rubric and the evaluation system as a whole need to be considered in future studies.

Recently, several attempts have been made to produce more robust scores from teacher observational rubrics. One example is from Hill, Charalambos, and Kraft (2012) who through the use of the generalizability theory, and a small sample of teachers and observers were able to identify an appropriate number of facets (raters/teachers/occasions) that should be used to maximize the reliability of the teachers' rating. With this method, the individual variance components can be identified, thus making it possible to determine what changes (adding raters or observations) would improve the reliability of the system as a whole.

Another example is by Ho and Kane (2013) who present several methods in which observations can be carried out while retaining a certain level of reliability. They used generalizability theory to identify the combination of raters and observations needed for the desired reliability level. Results of this study indicate that in general, the more raters and the more observations the better the reliability of the scores derived from the instrument. This study also demonstrated that additional research should be conducted to find situations where reliability can be maximized.

In another study funded by the Bill and Melinda Gates Foundation, the Measuring Effective Teachers Project (2013) found that "adding a second observer increases reliability significantly more than having the same observer score an additional lesson" (p. 5). This study found that reliability for only one observer during one time period (much like the present study) was .51 (Bill & Melinda Gates, 2013). After several years of research on the topic, there is more clarity on the topic (ideal number of observers and occasions) yet the highest reliability achieved by this study was .72 which is still not ideal for high-stakes decision making (Bill & Melinda Gates, 2013). More research is still needed.

Future research could inspect the observational rubric in more detail using generalizability theory. District decisions on the number of yearly teacher observations captured and utilized for high-stakes decision making should be based on the outcomes from a generalizability study and not out of minimum compliance with state laws. Since the cost of having additional observations may hinder results from a generalizability study, the suggestions provided by Ho and Kane (2013) could be used for lowering district costs to include more observations that are shorter in length.

Closing Remarks

Teaching is a highly complex job, which has serious effects on society. Teachers have the task of educating the future of the nation. Ineffective teachers could have a crippling effect on the nation and because of that, accountability for the profession is imperative.

There are many indicators that can be used to define teacher quality such as observations from a principal, measures of the effect a teacher has on student achievement, and student and/or parental input, just to name a few. It is reasonable to desire that all of these sources of data be included in teacher evaluations, yet each of these sources of data is not free of flaws. There could be errors in the timing of tests, the students assigned to the teachers, or the observations of the teacher. Because of the imperfections, it is critical to continue collecting validity evidence of the measures of teacher quality.

This research looked at one aspect of the incredibly complex process involved in appropriately identifying quality/effective teachers. As demonstrated, there are significant measurement and research design challenges in the task of developing and validating an accountability system for teachers. This study has raised a number of important questions that will need ongoing research using qualitative, quantitative, and mixed method approaches and which will need the involvement of policy makers, teachers, students, parents, and various other stakeholders.

References

Aaronson, D., Barrow, L., & Sander, W. (2007). Teachers and student achievement in the Chicago Public High Schools. *Journal of Labor Economics*, *25,* 95–135.

Acock, A. C. (2008). *Introduction to Mplus.* Retrieved from http://www.uclaisap.org/slides/caldar/summer%20institute/2008/Day-2%20Aug%2014-2008/Track%201/Mplus%20for%20Windows.pdf

American Educational Research Association, American Psychological Association, & Council on Measurement in Education. (1999). *Standards for educational and psychological testing*. Washington, DC: American Education Research Association.

American Institute for Research. (n.d.). Florida value-added technical report. Retrieved from http://www.fldoe.org/committees/doc/Value-Added-Model-Technical-Report.docx

Amrein-Beardsley, A. (2008). Methodological concerns about the education value-added assessment system. *Educational Researcher*, *37*(2), 65-75.

Anderson, H. M. (1954). A study of certain criteria of teaching effectiveness. *The Journal of Experimental Education, 23*(1), 41-71.

Armor, D., Conry-Oseguera, P., Cox, M., King, N., McDonnell, L., Pascal, A., Pauly, E., & Zellman, G. (1976). *Analysis of the school preferred reading program in selected Los Angeles minority schools*. Santa Monica, CA: Rand.

Approved District Performance Evaluation Systems. (n.d.). Educator recruitment, development, & retention. Retrieved from: http://www.fldoe.org/profdev/pa.asp

Ash, V. (n.d.). Bureau of K-12 assessment. Retrieved from http://fcat.fldoe.org/

Ballou, D. (2002, Summer). Sizing up test scores. *Education Next*. Retrieved from http://educationnext.org/files/ednext20022_10.pdf

Benson, J. (1998). Developing a strong program of construct validation: A test anxiety example. *Educational Measurement: Issues and Practice, 17*(1), 10-22.

Benson, J., & Hagtvet, K. (1996). The interplay between design, data analysis and theory in the measurement of coping. In M. Zeidner & N. Endler (Eds.), *Handbook of coping: Theory, research, applications* (pp. 83-106). New York, NY: Wiley.

Bentler, P. M. (1990). Comparative fit indexes in structural models. *Psychological Bulletin*, *107*, 238-246.

Bentler, P. M., & Bonett, D. G. (1980). Significance tests and goodness of fit in the analysis of covariance structures. *Psychological Bulletin*, *88*(3), 588-606.

Berger, M. C., & Toma, E. F. (1994). Variation in state education policies and effects on student performance. *Journal of Policy Analysis and Management, 13*(3), 477-491.

Berliner, D. C. (1987). Simple views of effective teaching and a simple theory of classroom instruction. In D. C. Berliner & B. Rosenshine (Eds.), *Talks to teachers* (pp. 93-110). New York, NY: Random House.

Berliner, D. C. (2005). The near impossibility of testing for teacher quality. *Journal of Teacher Education, 56*(3), 205-213.

Bill and Melinda Gates Foundation. (2013). Ensuring fair and reliable measures of effective teaching: Culminating findings from the met project. Seattle, WA: Bill and Melinda Gates Foundation. Retrieved from http://www.metproject.org/downloads/MET_Ensuring_Fair_and_Reliable_Measures_Practitioner_Brief.pdf

Bollen, K. A. (1990). Overall fit in covariance structure models: Two types of sample size effects. *Psychological Bulletin, 107,* 256–259.

Borland, M. V., & Howsen, R. M. (1992). Student academic achievement and the degree of market concentration in education. *Economics of Education Review, 11*(1), 31-39.

Bosshardt, W., & Watts, M. (1990). Instructor effects and their determinants in precollege economic education. *The Journal of Economic Education, 21*(3), 265-276.

Braun, H. I. (2004, December). *Value-added modeling: What does due diligence require?* Princeton, NJ: Educational Testing Service.

Brennan, R. L., National Council on Measurement in Education, & American Council on Education. (2006). *Educational measurement*. Westport, CT: Praeger.

Brookover, W. B. (1945). The relation of social factors to teaching ability. *The Journal of Experimental Education, 13*(4), 191-205.

Brown, T. A. (2006). *Confirmatory factor analysis for applied research*. NY: Guilford Press.

Browne, M. W., & Cudeck, R. (1993). Alternative ways of assessing model fit. In K. A. Bollen & J. S. Long (Eds.), *Testing structural equation models* (pp. 136-162). Newbury Park, CA: Sage.

Butler, M. R., & McNertney, E. M. (1991). Estimating educational production functions: The problem of multicollinearity. *Social Science Journal, 28*(4), 489-499.

Campbell, D. T., & Fiske, D. W. (1959). Convergent and discriminant validation by the multitrait-multimethod matrix. *Psychological Bulletin, 56*(2), 81-105.

Card, D., & Krueger, A. B. (1992). Does school quality matter? Returns to education and the characteristics of public schools in the United States. *Journal of Political Economy, 100*(1), 1-40.

Card, D., & Krueger, A. B. (1996). Labor market effects of school quality: Theory and Evidence (No. w5450). In G. Burtless (Ed.), *Does money matter? The effect of school resources on student achievement and adult success.* Washington, DC: Brookings Institution.

Cavalluzzo, L. C. (2004). Is national board certification an effective signal of teacher quality? Retrieved from http://www.cna.org/documents/CavaluzzoStudy.pdf

Cohen, J. (1992). Methods in psychology. A power primer. *Psychological Bulletin*, *112*(1), 155-159.

Coleman, J. S., Campbell, E. Q., Hobson, C. J., McPartland, J., Mood, A. M., Weinfeld, F. D., & York, R. L. (1966). *Equality of educational opportunity*. Washington, DC: US Government Printing Office.

Competencies for teachers of the twenty-first century. (n.d.). Retrieved from: www.fldoe.org/dpe/publications/preprofessional4-99.pdf

Cook, Brooks, Jefferis, McConnel, Paedae, Plavac, Boehme, Alachua County Education Association, Hernando Classroom Teachers Associaiton & Escambia Education Association v. Bennett, Chartrand, Armas, Bradshaw, Colon, Feingold, Padget, Shanahan (2013). Case Number: 2013cv00072. Retrieved from http://www.nea.org/assets/docs/CookvsBennettComplaint.pdf

Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika*, *16*(3), 297-334.

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

Cronbach, L. J. (1971). Test validation. In R. L. Thorndike (Ed.), *Educational measurement* (2nd ed., pp. 443-507). Washington, DC: American Council on Education.

Crocker, L. M., & Algina, J. (2006). *Introduction to classical and modern test theory*. Mason, OH: Cengage Learning.

Danielson, C. (2006). *Teacher leadership that strengthens professional practice*. Alexandria, VA: Association for Supervision and Curriculum Development.

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

Danielson, C. (2011). *The framework for teaching evaluation instrument*. Retrieved from http://www.danielsongroup.org/article.aspx?page=frameworkforteaching

Darling-Hammond, L. (1997). *Doing what matters most: Investing in quality teaching.* New York, NY: National Commission on Teaching and America's Future.

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

Dee, T. S. (2004). Teachers, race and student achievement in a randomized experiment. *Review of Economics and Statistics, 86*(1), 195-210.

Dee, T. S. (2005). A teacher like me: Does race, ethnicity or gender matter? *The American Economic Review, 95*(2), 158-165.

Dolan, R. C., & Schmidt, R. M. (1987). Assessing the impact of expenditure on achievement: Some methodological and policy considerations. *Economics of Education Review, 6*(3), 285-299.

Edwards, J. R. (2011). The fallacy of formative measurement. *Organizational Research Methods*, *14*, 370-388.

Ehrenberg, R. G., & Brewer, D. J. (1994). Do school and teacher characteristics matter? Evidence from high school and beyond. *Economics of Education Review, 13*(1), 1-17.

Ferguson, R. (1998). Teachers' perceptions and expectations and the black-white test score gap. In C. Jencks & M. Phillips (Eds.), *The black-white test score gap* (pp. 273-317). Washington, DC: Brookings Institution Press.

Florida Department of Education. (n.d.a). The Florida Comprehensive Assessment Test. Retrieved from http://fcat.fldoe.org/fcat/

Florida Department of Education (n.d.b). Frequently asked questions. Retrieved from http://www.fldoe.org/faq/default.asp?Dept=179&ID=985#Q985

Florida's Value-Added Technical Assistance Workshop. (2011). Retrieved from http://www.fldoe.org/committees/pdf/august12tammpres.pdf

Gall, M. D., Borg, W. R., & Gall, J. P. (1996). *Educational research*. New York, NY: Longman.

Gallagher, H. A. (2004). Vaughn elementary's innovative teacher evaluation system: Are teacher evaluation scores related to growth in student achievement? *Peabody Journal of Education, 79*(4), 79-107.

Gardner, D. P., Larsen, Y. W., & Baker, W. (1983). *A nation at risk: The imperative for educational reform.* Washington, DC: US Government Printing Office.

Geldhof, G. J., Preacher, K. J., & Zyphur, M. J. (2013). Reliability estimation in a multilevel confirmatory factor analysis framework. Retrieved from http://www.quantpsy.org/pubs/geldhof_preacher_zyphur_(in.press).pdf

Glaser, R., & Silver, E. (1994). Assessment, testing, and instruction: Retrospect and prospect. *Review of Research in Education*, *20*, 393-419.

Goldhaber, D., & Anthony, E. (2003). *Teacher quality and student achievement* (Urban Diversity Series No. 115). New York, NY: ERIC Clearinghouse on Urban Education. Retrieved from http://www.eric.ed.gov/PDFS/ED477271.pdf

Goldhaber, D., Brewer, D. J., & Anderson, D. (1999). A three-way error components analysis of educational productivity. *Education Economics, 7*, 199–208.

Goldhaber, D., & Brewer, D. (2000). Does teacher certification matter? High school teacher certification status and student achievement. *Educational Evaluation and Policy Analysis, 22*(2), 129-145.

Goldhaber, D., & Anthony, E. (2007). Can teacher quality be effectively assessed? National board certification as a signal of effective teaching. *The Review of Economics and Statistics, 89*(1), 134-150.

Goldhaber, D., & Theobald, R. (2011). Managing the teacher workforce: The consequences of "last in, first out" personnel policies. *Education Next, 11*(4), 78-83.

Gotham, R. E. (1945). Personality and teaching efficiency. *The Journal of Experimental Education, 14*(2), 157-165.

Graham, J. (2008). The general linear model as structural equation modeling. *Journal of Educational and Behavioral Statistics, 33*(4), 485-506.

Grimes, P. W., & Register, C. A. (1990). Teachers' unions and student achievement in high school economics. *Journal of Economic Education, 21*(3), 297-306.

Guilford, J. P. (1946). New standards for test evaluation. *Educational and Psychological Measurement*, *6*(4), 427-438.

Hancock, R. O., & Mueller, R. O. (2006). *Structural equation modeling: A second course*. Greenwich, CT: Information Age Publishing.

Hanushek, E. A. (1971). Teacher characteristics and gains in student achievement: Estimation using micro data. *American Economic Review, 60*(2) 280-288.

Hanushek, E. A. (1986). The economics of schooling: Production and efficiency in public schools. *Journal of Economic Literature, 24*(3), 1141-1177.

Hanushek, E. A. (1992). The trade-off between child quantity and quality. *Journal of Political Economy, 100*(1), 84-117.

Hanushek, E. A. (1997). Assessing the effects of school resources on student performance: An update. *Educational Evaluation and Policy Analysis, 19*(2), 141-164.

Hanushek, E. A. (2007). The single salary schedule and other issues of teacher pay. *Peabody Journal of Education, 82*(4), 574-586.

Hanushek, E. A., & Rivkin, S. G. (2010). Using value-added measures of teacher quality. *CALDER Policy Brief 9*. Washington, DC: National Center for Analysis of Longitudinal Data in Education Research.

Hanushek, E. A., Kain, J. F., & Rivkin, S. G. (1999). *Do higher salaries buy better teachers?* (Working Paper 7082). Cambridge, MA: National Bureau of Economic Research.

Harnisch, D. L. (1987). Characteristics associated with effective public high schools. *Journal of Educational Research, 80*(4), 233-241.

Harris, D. N. (2011). *Value-added measures in education: What every educator needs to know*. Cambridge, MA: Harvard Education Press.

Harris, D. N., & Hill, H. C. (2009). Point/counterpoint: Should "value-added" models be used to evaluate teachers? *Journal of Policy Analysis and Management, 28*(4), 692–712.

Harris, D. N., & Saas, T. R. (2009). What makes for a good teacher and who can tell? Working paper 30. Retrieved from http://www.caldercenter.org/publications.cfm#2009

Harris, D. N., & Rutledge, S. A. (2010). Models and predictors of teacher effectiveness: A comparison of research about teaching and other occupations. *Teachers College Record, 112*(3), 914-960.

Hill, C. W. (1921). The efficiency ratings of teachers. *The Elementary School Journal, 21*(6), 438-443.

Hill, H., Charalambos, C., & Kraft, M. (2012). When rater reliability is not enough: A teacher observation system and a case for the generalizability study. *Educational Researcher, 41*(2), 56-64.

Hill, H., Kapitula, L., & Umland, K. (2011). A validity argument approach to evaluating teacher value-added scores. *American Educational Research Journal, 48*(3), 794-831.

Ho, A. D., & Kane, T. J. (2013). *The reliability of classroom observations by school personnel.* Seattle, WA: Bill and Melinda Gates Foundation.

Holtzapple, E. (2003). Criterion-related validity evidence for a standards-based teacher evaluation system. *Journal of Personnel Evaluation in Education, 17*(3), 207-219.

Hu, L. T., & Bentler, P. M. (1998). Fit indices in covariance structure modeling: Sensitivity to underparameterized model misspecification. *Psychological Methods, 3*(4), 424-453.

Hu, L. T., & Bentler, P. M. (1999). Cutoff criteria for fit indices in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling, 6*, 1–55.

Jacob, B. A., & Lefgren, L. (2005). *Principals as agents: Subjective performance measurement in education.* (Working Paper 11463). Cambridge, MA: National Bureau of Economic Research.

Jacob, B. A., & Lefgren, L. (2008). Can principals identify effective teachers? Evidence on subjective performance evaluation in education. *Journal of Labor Economics* 26(1), 101–136.

Kane, M. (2006). Validation. In R. L. Brennan (Ed.), Educational measurement (4th ed., pp. 18–64). Westport, CT: Praeger.

Kane, T. J., & Staiger, D. O. (2008). *Estimating teacher impacts on student achievement: An experimental evaluation.* (Working Paper 14607). Cambridge, MA: National Bureau of Economic Research.

Kelly, S. (2012). *Assessing teacher quality: Understanding teacher effects on instruction and achievement*. New York, NY: Teachers College Press.

Kimball, S. M., White, B., & Milanowski, A. T. (2004). Examining the relationship between teacher evaluation and student assessment results in Washoe County. *Peabody Journal of Education, 79*(4), 54–78.

Kyriakides, L., & Creemers, B. P. M. (2008). A longitudinal study on the stability over time of school and teacher effects on student learning outcomes. *Oxford Review of Education*, 34, 521-545.

Koedel, C., & Betts, J. R. (2007). *Re-examining the role of teacher quality in the educational production function*. Nashville, TN: National Center on Performance Incentives, Vanderbilt, Peabody College.

Koedel, C., & Betts, J. R. (2011). Does student sorting invalidate value-added models of teacher-effectiveness? An extended analysis of the Rothstein critique. *Education Finance and Policy, 6*(1), 18-42.

Koretz, D. (2003). *Attempting to discern the effects of the NCLB accountability provisions on learning*. Paper presented at the annual meeting of the American Educational Research Association, Chicago, IL.

Kreft, I., & de Leeuw, J. (1998). *Introducing multilevel modeling*. Newburry Park, CA: Sage.

Kupermintz, H. (2003). Teacher effects and teacher effectiveness: A validity investigation of the Tennessee value-added assessment system. *Educational Evaluation and Policy Analysis, 25*(3), 287–298.

Kurth, M. M. (1987). Teachers' unions and excellence in education: An analysis of the decline in SAT scores. *Journal of Labor Research, 8*(4), 351-367.

Linn, R. L. (2006) Validity of inferences from test-based accountability educational accountability systems. *Journal of Personnel Evaluation in Education, 19*(1/2), 5–15.

Lockwood, J. R., Louis, T. A., & McCaffrey, D. F. (2002). Uncertainty in rank estimation: Implications for value-added modeling accountability systems. *Journal of Educational and Behavioral Statistics, 27*(3), 255–270.

Lockwood, J. R., McCaffrey, D. F., Hamilton, L. S., Stecher, B., Le, V., & Martinez, J. F. (2007). The sensitivity of value-added teacher effect estimates to different mathematics achievement measures. *Journal of Educational Measurement, 44*(1), 47–67.

Manatt, R. P., & Daniels, B. (1990). Relationships between principals' ratings of teacher performance and student achievement. *Journal of Personnel Evaluation in Education, 4*(2), 189-201.

Marsh, H. W., Balla, J. R., & McDonald, R. P. (1988). Goodness-of-fit indexes in confirmatory factor analysis: The effect of sample size. *Psychological Bulletin, 103,* 391–410.

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

McCaffrey, D. F., Lockwood, J. R., Koretz, D., Louis, T.A., & Hamilton, L. (2004a). Let's see more empirical studies on value-added modeling of teacher effects: A reply to Raudenbush, Rubin, Stuart, Zanutto, and Reckase. *Journal of Educational and Behavioral Statistics, 29*(1), 139-143.

McCaffrey, D. F., Lockwood, J. R., Koretz, D., Louis, T. A., & Hamilton, L. (2004b). Models for value-added modeling of teacher effects. *Journal of Educational and Behavioral Statistics, 29*(1), 67–101.

McCaffrey, D. F., Sass, T. R., Lockwood, J. R., & Mihaly, K. (2009). The intertemporal variability of teacher effect estimates. *Education, Finance and Policy, 4*(4), 572–606.

McDonald, R. P. (1999). *Test theory: A unified treatment.* Mahwah, NJ: Erlbaum.

Medley, D. M., & Coker, H. (1987). The accuracy of principals' judgments of teacher performance. *Journal of Educational Research, 80*(4), 242-247.

Messick, S. (1981). Evidence and ethics in the evaluation of tests. *Educational Researcher, 10*(9), 9-20.

Messick S. (1993). Validity. In R. L. Linn (Ed.), *Educational measurement* (3rd ed., p. 13-100). Phoenix, AZ: Oryx.

Messick S. (1995). Validity of psychological assessment; Validation of inferences from persons' responses and performances as scientific inquiry into meaning. *American Psychologist 50*(9), 741-749.

Meyer, R. (1997). Value-added indicators of school performance: A primer. *Economics of Education Review*, 16, 183–301.

Meyer, R. H. (2000). Value-added indicators: A powerful tool for evaluating science and mathematics programs and policies. Retrieved from http://www.wcer.wisc.edu/archive/nise/Publications/

Milanowski, A. (2004). The relationship between teacher performance evaluation scores and student achievement: Evidence from Cincinnati. *Peabody Journal of Education, 79*(4), 33–53.

Miller J., McKenna, B., & McKenna, M. (1996). A comparison of alternatively and traditionally prepared teachers. *Journal of Teacher Education, 49*(3), 165-176.

Montmarquette, C., & Mahseredjian, S. (1989). Does school matter for educational achievement? A two-way nested error components analysis. *Journal of Applied Econometrics, 4*(2), 181-193.

Murnane, R., Willett, J., & Levy, F. (1995). The growing importance of cognitive skills in wage determination. *The Review of Economics and Statistics, 77*(2), 251-266.

Murnane, R. J., & Phillips, B. (1981a). What do effective teachers of inner-city children have in common? *Social Science Research, 10*(1), 83-100.

Murnane, R. J., & Phillips, B. R. (1981b). Learning by doing, vintage, and selection: Three pieces of the puzzle relating teaching experience and teaching performance. *Economics of Education Review, 1*(4), 453-465.

Murnane, R. J., Singer, J. D., Willett, J. B., Kemple, J. J., & Olson, R. J. (1991). Who will teach? *Policies that matter.* Cambridge, MA: Harvard University Press.

Muthén, L. K. & Muthén, B. O. (1998-2007). *Mplus user's guide* (5th Edition). Los Angeles, CA: Muthén & Muthén. Retrieved from http://www.statmodel.com/download/usersguide/Mplus%20Users%20Guide%20v5.pdf

National Board. (2013). *2013 Guide to National Board Certification*. Retrieved from http://www.nbpts.org/sites/default/files/documents/Candidate-Center/Guide_to_NB_Certification%203.25.13.pdf

Nixon, L., & Robinson, M. D. (1999). The educational attainment of young women: Role model effects of female high school faculty. *Demography, 36*(2), 185-194.

Nunnally, J. (1978). *Psychometric theory*. New York, NY: McGraw-Hill.

Papay, J. P. (2011). Different tests, different answers. *American Educational Research Journal, 48*(1), 163-193.

Peek, Weatherstone and Florida Education Association v. Florida State Board of Education and Florida Department of Education. (2012). Case number 12-1111RP. Retrieved from http://www.doah.state.fl.us/DocDoc/2012/001111/12001111_0_03302012_04493952_e.pdf

Race to the top fund. (2011). Retrieved from http://www2.ed.gov/programs/racetothetop/index.html

Raudenbush, S. W. (1995). Maximum likelihood estimation for unbalanced multilevel covariance structure models via the EM algorithm. *British Journal of Mathematical and Statistical Psychology, 48*, 359 –370.

Raudenbush, S. W. (2004). Schooling, statistics, and poverty: Can we measure school improvement? *9th Annual William H. Angoff Memorial Lecture*. Educational Testing Service. Retrieved from http://www.ets.org/research/policy_research_reports/pic-ang9

Raudenbush, S. W., & Bryk, A. S. (2002). *Hierarchical linear models: Applications and data analysis methods.* Newbury Park, CA: Sage.

Raudenbush, S. W., Rowan, B., & Kang, S. J. (1991). A multilevel, multivariate model of studying school climate with estimation via the EM algorithm and application to U. S. high school data. *Journal of Educational Statistics, 16*, 296 – 330.

Raykov, T., & Marcoulides, G. A. (2006). *A first course in structural equation modeling*. Mahwah, NJ: Lawrence Erlbaum.

Resnick, D. P. (1982). History of educational testing. In A. K. Wigdor & W. R. Garner (Eds.), *Ability testing: Uses, consequences and controversies, Part II* (pp. 173-194). Washington, DC: National Academy Press.

Rivkin, S. G., Hanushek, E. A., & Kain, J. F. (2005). Teachers, schools, and academic achievement. *Econometrica, 73*, 417–458.

Rockoff, J. E. (2004). The impact of individual teachers on students' achievement: Evidence from panel data. *American Economic Review, 94*, 247–252.

Robinson, Weatherstone, Hall, Lofton, Williams & Mayer. v. Robinson & State of Florida Department of Education (2011). Case number 2011CA2526. Retrieved from http://www.meyerandbrooks.com/documents/Robinson%20vs%20Robinson/Robinson_v_Robinson_Complaint.pdf

Rowan, B., Correnti, R., & Miller, R. J. (2002). What large-scale, survey research tells us about teacher effects on student achievement: Insights from the "Prospects" study of elementary schools. *Teachers College Record, 104*(8), 1525-1567.

Rubin, D. B., Stuart, E. A., & Zanutto, E. L. (2004). A potential outcomes view of value-added assessment in education. *Journal of Educational and Behavioral Statistics, 29*(1), 103–116.

Sabo, K. E., & Lawton, K. (2013). *Confirmatory and Exploratory Factor Analysis of a Teacher Observation Rubric in Value-Added Schools.* Paper presented at the American Educational Research Association Conference, San Francisco, CA.

Sanders, W. L. (1993). Expenditures and student achievement in Illinois: New evidence. *Journal of Public Economics, 52*(3), 403-416.

Sanders, W. L. (1998). Value-added assessment. *The School Administrator, 55*(11), 24-32.

Sanders, W. L., Saxton, A., Schneider, J., Dearden, B., Wright, S. P., & Horn, S. (2002). *Effects of building change on indicators of student achievement growth: Tennessee value-added assessment system*. Knoxville, TN: University of Tennessee Value-Added Research and Assessment Center.

Sato, M., Chung, R. R., & Darling-Hammond, L. (2008). Improving teachers' assessment practices through professional development: The case of National Board Certification. *American Educational Research Journal, 45*(3), 669-700.

Schacter, J., & Thum, Y. M. (2004). Paying for high and low-quality teaching. *Economics of Education Review, 23*, 411-430.

Senate Bill 0736. (n.d.). Retrieved from http://www.flsenate.gov/Session/Bill/2011/736

Schaeffer, B. (2004). Districts pilot value-added assessment. *School Administrator, 61(*11)*,* 20-24.

Smith, T., Gordon, B., Colby, S., & Wang, J. (2005). *An examination of the relationship of the depth of student learning and National Board Certification status.* Boone, NC: Office for Research on Teaching, Appalachian State University. Retrieved from http://www.nbpts.org/UserFiles/File/Applachian_State_Study_Smith.pdf

Snijders, T., & Bosker, R. (1999). *Multilevel analysis: An introduction to basic and advanced multilevel modeling*. Thousand Oaks, CA: Sage.

State Board of Education Presentation. (2012). Retrieved from https://www.fldoe.org/board/meetings/2012_01_24/eval.pdf

Steiger, J. H. (2000). Point estimation, hypothesis testing, and interval estimation using the RMSEA: Some comments and a reply to Hayduk and Glasser. *Structural Equation Modeling, 7,* 149–162.

Steiger, J. H. (2007). Understanding the limitations of global fit assessment in structural equation modeling. *Personality and Individual Differences, 42*, 893–898.

Stern, D. (1989). Educational cost factors and student achievement in grades 3 and 6: Some new evidence. *Economics of Education Review, 8*(2), 149-158.

Strong, M. (2011). *The highly qualified teacher: What is teacher quality and how do we measure it?* New York, NY: Teachers College Press.

Student Success Act, 1008.22(8), F.S., Section 1012.34(3)(a)1. (2011). Retrieved from: http://www.flsenate.gov/laws/statutes/2011/1012.34

Superfine, B. M. (2005). The politics of accountability: The rise and fall of Goals 2000. *American Journal of Education*, *112*(1), 10-43.

Tekwe, C. D., Carter, R. L., Ma, C., Algina, J., Lucas, M. E., Roth, J., et al. (2004). An empirical comparison of statistical models for value- added assessment of school performance. *Journal of Educational and Behavioral Statistics, 29*(1), 11–36.

The framework for teaching. (n.d.). Retrieved from: http://www.danielsongroup.org/article.aspx?page=frameworkforteaching

U.S. Department of Education. (2001). Washington, DC: No child left behind act. Retrieved from http://www2.ed.gov/policy/elsec/leg/esea02/index.html

Value-Added Model White Paper. (n.d.). Student growth. Retrieved from http://www.fldoe.org/committees/sg.asp (pdf)

Vandevoort, L. G., Amrein-Beardsley, A., & Berliner, D. C. (2004). National board certified teachers and their students' achievement. *Education Policy Analysis Archives, 12*, (46).

Webcast. (2011). *Student growth*. Retrieved from http://www.fldoe.org/committees/sg.asp

Webinar Presentation. (2012). *Student growth implementation committee*. Retrieved from http://www.fldoe.org/committees/ppt/VAM-Webinar.ppt

Wilkerson, D. J., Manatt, R. P., Rogers, M. A., & Maughan, R. (2000). Validation of student, principal, and self-ratings in 360 feedback® for teacher evaluation. *Journal of Personnel Evaluation in Education, 14*(2), 179-192.

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, 1*(1), 57-67.

Appendices

Appendix A. Florida Course Codes Used in the Value-Added Model

Table 1. Course Codes Used in the Mathematics Value-Added Model

Year	Course Number	Course Name
2008-09, 2009-10, 2010-11	1200300	Pre-Algebra
2008-09, 2009-10, 2010-11	1200310	Algebra I
2008-09, 2009-10, 2010-11	1200320	Algebra I Honors
2008-09, 2009-10, 2010-11	1200330	Algebra II
2008-09, 2009-10, 2010-11	1200340	Algebra II Honors
2008-09, 2009-10, 2010-11	1200370	Algebra Ia
2008-09, 2009-10, 2010-11	1200380	Algebra Ib
2008-09, 2009-10, 2010-11	1200400	Intensive Mathematics
2008-09, 2009-10, 2010-11	1200410	Math for College Success
2008-09, 2009-10, 2010-11	1200500	Advanded Algebra with Financial Applications
2008-09, 2009-10, 2010-11	1200700	Math College Readiness
2008-09, 2009-10, 2010-11	1201300	Math Analysis
2008-09, 2009-10, 2010-11	1202371	Pre-AICE Additional Math III
2008-09, 2009-10, 2010-11	1204000	M/J Intensive Mathematics (MC)
2008-09, 2009-10, 2010-11	1205010	M/J Mathematics 1
2008-09, 2009-10, 2010-11	1205020	M/J Mathematics 1, Advanced
2008-09, 2009-10, 2010-11	1205040	M/J Mathematics 2
2008-09, 2009-10, 2010-11	1205050	M/J Mathematics 2, Advanced
2008-09, 2009-10, 2010-11	1205070	M/J Mathematics 3
2008-09, 2009-10, 2010-11	1205080	M/J Mathematics 3, Advanced
2008-09, 2009-10, 2010-11	1205090	M/J Mathematics IB
2008-09, 2009-10, 2010-11	1205100	M/J Pre-algebra IB
2008-09, 2009-10, 2010-11	1205370	Consumer Mathematics
2008-09, 2009-10, 2010-11	1205400	Applied Mathematics I
2008-09, 2009-10, 2010-11	1205410	Applied Mathematics II
2008-09, 2009-10, 2010-11	1205500	Explorations in Mathematics I
2008-09, 2009-10, 2010-11	1205510	Explorations in Mathematics II
2008-09, 2009-10, 2010-11	1205540	Business Mathematics
2008-09, 2009-10, 2010-11	1206300	Informal Geometry
2008-09, 2009-10, 2010-11	1206310	Geometry
2008-09, 2009-10, 2010-11	1206320	Geometry Honors
2008-09, 2009-10, 2010-11	1207310	Integrated Mathematics I
2008-09, 2009-10, 2010-11	1207320	Integrated Mathematics II
2008-09, 2009-10, 2010-11	1207330	Integrated Mathematics III
2008-09, 2009-10, 2010-11	1209810	Pre-AICE Mathematics I
2008-09, 2009-10, 2010-11	1209820	Pre-AICE Mathematics II
2008-09	1298010	M/J Great Explorations in Math (GEM) 6th Pre-Algebra
2008-09	1298020	M/J Great Explorations in Math (GEM) 7th Algebra
2008-09	1298030	M/J Great Explorations in Math (GEM) 8th Geometry
2008-09	5012000	Mathematics-Elementary
2008-09	5012010	Functional Basic Skills in Mathematics-Elementary
2008-09, 2009-10, 2010-11	5012020	Math Grade K
2008-09, 2009-10, 2010-11	5012030	Math Grade 1
2008-09, 2009-10, 2010-11	5012040	Math Grade 2
2008-09, 2009-10, 2010-11	5012050	Math Grade 3
2008-09, 2009-10, 2010-11	5012060	Math Grade 4
2008-09, 2009-10, 2010-11	5012070	Math Grade 5
2008-09, 2009-10, 2010-11	7712010	Mathematics K-5
2008-09, 2009-10, 2010-11	7755010	Academics K-5
2008-09, 2009-10, 2010-11	7755030	Academic Skills K-5
2008-09, 2009-10, 2010-11	7755040	Advanced Academic Skills K-5
2008-09, 2009-10, 2010-11	7755050	Developmental Skills K-5
2008-09, 2009-10, 2010-11	7812010	Mathematics: 6-8
2008-09, 2009-10, 2010-11	7855010	Academics 6-8
2008-09, 2009-10, 2010-11	7855030	Academic Skills 6-8
2008-09, 2009-10, 2010-11	7855040	Advanced Academics 6-8
2008-09, 2009-10, 2010-11	7855050	Developmental Skills 6-8
2008-09, 2009-10, 2010-11	7912050	Mathematics 9-12
2008-09, 2009-10, 2010-11	7912340	Life Skills Math: 9-12
2008-09	129800A	M/J Great Explorations in Math (GEM) 6th Pre-Algebra
2008-09	129800B	M/J Great Explorations in Math (GEM) 7th Algebra
2008-09	129800C	M/J Great Explorations in Math (GEM) 8th Geometry

Source: (American Institute for Research, n.d.)

Table 2. Course Codes Used in the Reading Value-Added Model

Year	Course Number	Course Name
2008-09, 2009-10, 2010-11	1000000	M/J Intensive Language Arts (MC)
2008-09, 2009-10, 2010-11	1000010	M/J Intensive Reading (MC)
2009-10, 2010-11	1000020	M/J Intensive Reading and Career Planning
2008-09, 2009-10, 2010-11	1000400	Intensive Language Arts
2008-09, 2009-10, 2010-11	1000410	Intensive Reading
2008-09, 2009-10, 2010-11	1001010	M/J Language Arts 1
2008-09, 2009-10, 2010-11	1001020	M/J Language Arts, 1 Adv.
2008-09, 2009-10, 2010-11	1001030	M/J Language Arts 1, International Baccalaureate
2008-09, 2009-10, 2010-11	1001040	M/J Language Arts 2
2008-09, 2009-10, 2010-11	1001050	M/J Langague Arts 2, Adv
2008-09, 2009-10, 2010-11	1001060	M/J Language Arts 2, International Baccalaureate
2008-09, 2009-10, 2010-11	1001070	M/J Language Arts 3
2008-09, 2009-10, 2010-11	1001080	M/J Language Arts 3, Adv
2008-09, 2009-10, 2010-11	1001090	M/J Language Arts 3,International Baccalaureate
2008-09, 2009-10, 2010-11	1001300	English Skills I
2008-09, 2009-10, 2010-11	1001310	English I
2008-09, 2009-10, 2010-11	1001320	English Honors I
2008-09, 2009-10, 2010-11	1001330	English Skills II
2008-09, 2009-10, 2010-11	1001340	English II
2008-09, 2009-10, 2010-11	1001350	English Honors II
2008-09, 2009-10, 2010-11	1001440	Business English I
2008-09, 2009-10, 2010-11	1001450	Business English II
2008-09, 2009-10, 2010-11	1001560	Pre-AICE English Language
2008-09, 2009-10, 2010-11	1001800	English I Pre-International Baccalaureate
2008-09, 2009-10, 2010-11	1001810	English II Pre-International Baccalaureate
2009-10, 2010-11	1001840	IB Middle Years Program English I
2009-10, 2010-11	1001845	IB Middle Years Program English II
2008-09, 2009-10, 2010-11	1002000	M/J Language Arts 1 through ESOL
2008-09, 2009-10, 2010-11	1002010	M/J Langague Arts 2 through ESOL
2008-09, 2009-10, 2010-11	1002020	M/J Langague Arts 3 through ESOL
2008-09, 2009-10, 2010-11	1002180	M/J Developmental Language Arts Through ESOL (MC)
2008-09, 2009-10, 2010-11	1002300	English I through ESOL
2008-09, 2009-10, 2010-11	1002310	English II through ESOL
2008-09, 2009-10, 2010-11	1002380	Developmental Language Arts Through ESOL
2008-09, 2009-10, 2010-11	1005375	AICE English Literature II
2008-09, 2009-10, 2010-11	1008010	M/J Reading 1
2008-09, 2009-10, 2010-11	1008020	M/J Reading 1, Advanced
2008-09, 2009-10, 2010-11	1008040	M/J Reading 2
2008-09, 2009-10, 2010-11	1008050	M/J Reading 2, Advanced
2008-09, 2009-10, 2010-11	1008070	M/J Reading 3
2008-09, 2009-10, 2010-11	1008080	M/J Reading, Advanced
2008-09, 2009-10, 2010-11	1008300	Reading I
2008-09, 2009-10, 2010-11	1008310	Reading II
2008-09, 2009-10, 2010-11	1008320	Advanced Reading
2008-09, 2009-10, 2010-11	1008330	Reading III
2009-10, 2010-11	1008350	Reading for College Success
2008-09, 2009-10, 2010-11	2400000	Sixth Grade
2008-09, 2009-10, 2010-11	5010010	ESOL English for Speakers of Other Language-Elementary
2008-09, 2009-10, 2010-11	5010020	Functional Basic Skills in Reading-Elementary
2008-09, 2009-10, 2010-11	5010040	Language Arts-Elementary
2008-09, 2009-10, 2010-11	5010050	Reading-Elementary
2008-09, 2009-10, 2010-11	5010060	Integrated Language Arts-Elementary
2008-09, 2009-10, 2010-11	7710010	Language Arts K-5
2008-09, 2009-10, 2010-11	7755010	Academics K-5
2008-09, 2009-10, 2010-11	7755030	Academic Skills K-5
2008-09, 2009-10, 2010-11	7755040	Advanced Academic Skills K-5
2008-09, 2009-10, 2010-11	7755050	Developmental Skills K-5
2008-09, 2009-10, 2010-11	7810010	Language Arts 6-8
2008-09, 2009-10, 2010-11	7810020	Reading: 6-8
2008-09, 2009-10, 2010-11	7910100	Reading 9-12
2008-09, 2009-10, 2010-11	7910110	English 9-12
2008-09, 2009-10, 2010-11	7910400	Life Skills Reading: 9-12

Source: (American Institute for Research, n.d.)

Appendix B. Description of the Covariates in the Value-Added Model

Covariates	Description
The number of subject-relevant courses in which the student is enrolled	Some students are enrolled in multiple courses that, according to the Florida course code directory, are linked to an FCAT test. This variable counts, for each student, the number of courses they are enrolled in that is linked to the FCAT test via the course code directory (see Appendix A).
Two prior years of achievement scores	These are always the scores for the subject from the two prior years. For example, grade 8 math uses grades 6 and 7 FCAT math scores as predictors.
Students with Disabilities (SWD) status	This is a dichotomous variable denoting whether a student receives special education services for a specific disability.
English language learner (ELL) status	This is a dichotomous variable denoting whether students are currently enrolled in an English language learner program or not for less than two years.
Gifted status	This is a dichotomous variable denoting if the student is enrolled in a gifted program or not.
Attendance	This is a continuous variable counting the number of days the student was present during the school year.
Mobility (number of transitions)	This is a continuous variable counting the number of transitions across schools within the same school year.
Difference from modal age in grade (as an indicator of retention)	This is a continuous variable computed as $x_i - x$ where x_i is the age in months for student i and x is the modal age for students enrolled in the same grade across the state.
Class size	A continuous measure counting the number of students linked to teacher j.
Homogeneity of entering test scores in the class	A continuous variable computed as the interquartile range of student entering scores in the class.

Source: American Institute for Research, n.d., p.3, 4.

Appendix C. Correlations By School With Each of the Factors Underlying the Observational Rubric and Each of the VAM Scores

Level	Type of VAM Score	Factor 1.1	Factor 1.2	Factor 2.1	Factor 3.1	Factor 3.2	Total Score
E1	VAM with SE	.08	.14	.42	-.16	.11	.22
	VAM without SE	.42	.47	.57	.08	.49	.54
E2	VAM with SE	-.26	-.33	-.02	-.14	-.44	-.29
	VAM without SE	.05	-.19	-.11	-.19	-.46	-.23
E3	VAM with SE	.23	-.12	-.44	.09	-.52	-.58
	VAM without SE	-.33	.61	-.44	.67	-.63	-.19
E4	VAM with SE	.48	.19	.27	.21	.28	.31
	VAM without SE	.51	.25	.32	.33	.31	.36
E5	VAM with SE	.46	.52	.53	.49	.35	.56
	VAM without SE	.20	.34	.32	.25	.20	.33
E6	VAM with SE	.07	.19	-.07	.02	-.01	.05
	VAM without SE	.20	.31	.15	.01	.10	.22
E7	VAM with SE	.17	.43	.46	.52	.19	.40
	VAM without SE	.10	.22	.20	.36	-.11	.15
E8	VAM with SE	.00	.00	.20	.25	.01	.09
	VAM without SE	.09	.17	.44	.25	.35	.34
E9	VAM with SE	-.08	.53	.17	.26	-.03	.25
	VAM without SE	-.02	.44	.16	.28	.26	.33
E10	VAM with SE	-.32	-.26	-.12	-.04	-.50	-.28
	VAM without SE	-.18	-.15	-.08	.01	-.39	-.19
E11	VAM with SE	.54	.62	.64	.[c]	.69	.73
	VAM without SE	.55	.66	.71	.[c]	.52	.73
E12	VAM with SE	.61	.21	.21	.56	.05	.25
	VAM without SE	.24	.27	.29	.45	.22	.30
E13	VAM with SE	.21	.12	.17	-.07	.10	.14
	VAM without SE	.20	-.06	-.21	.00	-.25	-.09
E14	VAM with SE	.05	-.10	-.03	-.20	-.04	-.06
	VAM without SE	.16	.03	.10	-.36	.07	.09
E15	VAM with SE	-.05	-.06	-.03	.14	.25	.03
	VAM without SE	.01	-.03	-.03	.10	.24	.05
E16	VAM with SE	-.39	-.45	-.21	-.57	-.31	-.41
	VAM without SE	-.05	-.09	-.03	-.07	-.11	-.08
E17	VAM with SE	-.46	-.26	-.10	-.12	-.20	-.24
	VAM without SE	-.21	.15	.13	.18	.09	.12

Level	Type of VAM Score	Factor 1.1	Factor 1.2	Factor 2.1	Factor 3.1	Factor 3.2	Total Score
E18	VAM with SE	-.01	.09	.07	.20	.03	.07
	VAM without SE	.58	.74	.69	.35	.77	.74
E19	VAM with SE	-.03	-.30	.05	.40	.51	.10
	VAM without SE	-.01	-.17	.20	.40	.58	.21
E20	VAM with SE	.31	.35	.77	.16	-.18	.48
	VAM without SE	.51	.27	.81	.21	-.05	.55
E21	VAM with SE	.41	.07	.15	.01	.34	.30
	VAM without SE	.51	.33	.14	.06	.44	.43
E22	VAM with SE	.34	.21	.24	-.14	.14	.25
	VAM without SE	-.01	-.11	-.22	-.10	-.22	-.17
E23	VAM with SE	.21	.17	.16	.06	-.15	.13
	VAM without SE	.26	.27	.24	.08	.06	.24
E24	VAM with SE	.28	.30	.30	-.49	.10	.24
	VAM without SE	.29	.42	.44	-.46	.28	.39
E25	VAM with SE	-.06	-.17	.05	-.33	-.19	-.11
	VAM without SE	.06	.04	.04	-.11	.05	.04
E26	VAM with SE	.10	.01	.05	.28	-.03	.07
	VAM without SE	.29	.32	.17	.13	.12	.24
E27	VAM with SE	.44	.44	.40	.29	.47	.46
	VAM without SE	.41	.48	.47	.46	.42	.50
E28	VAM with SE	-.15	-.21	-.25	-.18	-.09	-.27
	VAM without SE	.33	.19	.13	.21	.23	-.39
E29	VAM with SE	.03	-.42	.02	-.35	-.12	-.18
	VAM without SE	.45	-.02	.33	.03	-.04	.20
E30	VAM with SE	.01	-.10	-.03	-.26	-.07	-.06
	VAM without SE	.00	-.09	-.08	-.23	-.05	-.08
E31	VAM with SE	-.07	-.06	-.19	-.05	-.23	-.15
	VAM without SE	.01	-.03	-.14	-.13	-.09	-.08
E32	VAM with SE	.03	.35	.34	.18	.16	.33
	VAM without SE	.10	.40	.42	.34	.24	.42
E33	VAM with SE	.43	.36	.50	-.11	.45	.45
	VAM without SE	.29	.20	.35	-.22	.24	.26
E34	VAM with SE	.08	.00	-.06	.26	-.31	-.06
	VAM without SE	-.28	-.42	-.16	.22	-.53	-.40
E35	VAM with SE	.02	-.04	.12	-.07	-.24	-.11
	VAM without SE	.02	-.09	-.15	.24	-.09	-.05
E36	VAM with SE	.11	-.23	.03	.[c]	.03	-.01
	VAM without SE	.00	-.22	.03	.[c]	-.09	-.07

Level	Type of VAM Score	Factor 1.1	Factor 1.2	Factor 2.1	Factor 3.1	Factor 3.2	Total Score
E37	VAM with SE	-.27	-.24	-.27	-.36	-.22	-.31
	VAM without SE	-.20	-.24	-.30	-.29	-.11	-.29
E38	VAM with SE	-.27	-.27	.10	-.12	.04	-.11
	VAM without SE	-.24	-.09	.10	-.03	-.06	-.05
E39	VAM with SE	.33	.18	.36	-.42	.11	.20
	VAM without SE	.38	.39	.50	-.18	.36	.47
E40	VAM with SE	.00	-.01	.30	.13	.22	.17
	VAM without SE	.26	.26	.51	.35	.31	.41
E41	VAM with SE	-.83	-.70	-.67	-.60	.27	-.84
	VAM without SE	-.87	-.54	-.49	-.65	.19	-.76
E42	VAM with SE	.11	-.60	-.27	.01	.15	-.21
	VAM without SE	.22	-.19	-.03	.28	.23	.07
E43	VAM with SE	.07	.06	.15	.39	.28	.16
	VAM without SE	.64	.68	.74	.53	.66	.74
E44	VAM with SE	.34	.29	.28	.19	-.01	.29
	VAM without SE	.29	.21	.27	.19	-.06	.25
E45	VAM with SE	-.07	.00	-.33	-.18	-.34	-.20
	VAM without SE	-.02	.24	-.01	-.19	-.16	.03
E46	VAM with SE	.03	-.25	-.21	-.45	-.35	-.27
	VAM without SE	-.14	-.26	-.06	-.33	-.40	-.27
E47	VAM with SE	.28	.43	.51	-.17	.04	.37
	VAM without SE	.20	.62	.32	.13	.08	.39
E48	VAM with SE	.13	.09	.06	.17	-.14	.06
	VAM without SE	.14	.28	.10	.10	-.14	.14
E49	VAM with SE	.30	.42	.24	.29	.71	.47
	VAM without SE	.38	.47	.27	.48	.66	.51
E50	VAM with SE	.30	.13	.17	.16	.22	.22
	VAM without SE	.08	.39	.40	.25	.48	.41
E51	VAM with SE	.04	-.06	-.07	.03	-.07	-.05
	VAM without SE	.13	.00	-.02	.03	.05	.03
E52	VAM with SE	.10	.10	.22	.12	.04	.16
	VAM without SE	.02	.01	.15	-.01	.14	.10
E53	VAM with SE	.09	.34	.43	.13	.21	.33
	VAM without SE	.22	.46	.30	-.04	.06	.30
E54	VAM with SE	.41	.38	.46	.23	.38	.46
	VAM without SE	.56	.59	.55	.33	.46	.61
E55	VAM with SE	-.30	-.08	-.15	-.01	.29	-.04
	VAM without SE	.01	.21	.00	.41	.19	.27

Level	Type of VAM Score	Factor 1.1	Factor 1.2	Factor 2.1	Factor 3.1	Factor 3.2	Total Score
E56	VAM with SE	-.38	-.55	-.53	-.45	-.27	-.51
	VAM without SE	-.15	-.24	-.20	-.14	-.10	-.20
E57	VAM with SE	.13	.32	.08	.40	.06	.19
	VAM without SE	-.03	.31	-.07	.37	-.08	.04
E58	VAM with SE	.18	-.08	.31	.08	.36	.23
	VAM without SE	.42	.23	.30	.10	.36	.41
E59	VAM with SE	.76	.66	.55	.68	.72	.72
	VAM without SE	.83	.71	.60	.74	.79	.79
E60	VAM with SE	.52	.53	.45	.24	.48	.61
	VAM without SE	.24	.34	.63	.66	.24	.56
E61	VAM with SE	-.20	-.19	-.13	-.40	-.13	-.20
	VAM without SE	-.04	-.07	.05	-.20	.01	-.02
E62	VAM with SE	.22	.36	.08	-.11	.03	.16
	VAM without SE	.13	.21	.15	.01	.03	.14
E63	VAM with SE	-.14	.29	.24	.10	.25	.16
	VAM without SE	-.04	.27	.32	.27	.18	.25
E64	VAM with SE	.73	.79	.85	.70	.74	.84
	VAM without SE	.82	.85	.90	.73	.85	.91
E65	VAM with SE	.01	.43	.36	.21	.13	.34
	VAM without SE	.02	.38	.45	.29	.08	.36
E66	VAM with SE	-.10	-.02	.08	.12	.18	.04
	VAM without SE	.02	-.01	.06	.21	.23	.09
E67	VAM with SE	.19	.06	.02	.44	-.13	.06
	VAM without SE	.36	.15	.11	.23	-.04	.17
E68	VAM with SE	.05	-.17	-.69	-.07	.04	-.30
	VAM without SE	-.04	.07	-.36	.20	.11	-.06
H1	VAM with SE	.09	.11	.12	.26	.18	.15
	VAM without SE	.12	-.06	.03	.11	.04	.03
H2	VAM with SE	.08	-.28	-.04	-.08	.07	-.10
	VAM without SE	.18	.23	.05	.01	.27	.19
H3	VAM with SE	-.09	.03	.21	.24	.25	.16
	VAM without SE	.02	.16	.05	.06	.03	.09
H4	VAM with SE	.05	-.01	-.03	.11	-.20	-.03
	VAM without SE	.02	-.01	-.02	.05	-.15	-.03
H5	VAM with SE	.24	.18	.63	.45	.30	.45
	VAM without SE	.40	.38	.61	.50	.50	.57
H6	VAM with SE	.01	-.14	.02	.00	.08	-.01
	VAM without SE	.28	.13	-.02	-.35	.15	.07

Level	Type of VAM Score	Factor 1.1	Factor 1.2	Factor 2.1	Factor 3.1	Factor 3.2	Total Score
H7	VAM with SE	.04	.10	.07	.30	.16	.11
	VAM without SE	.22	.15	.08	.50	.06	.16
H8	VAM with SE	.02	.13	.10	-.02	.27	.14
	VAM without SE	.08	.16	.13	-.24	.22	.14
H9	VAM with SE	-.13	-.19	-.12	-.24	-.31	-.21
	VAM without SE	.17	.35	.16	.04	.32	.26
H10	VAM with SE	-.10	-.20	.01	-.16	-.42	-.23
	VAM without SE	.01	-.13	.02	.22	-.02	-.02
H11	VAM with SE	-.03	-.07	.08	-.18	.03	-.01
	VAM without SE	.32	.24	.21	.12	.18	.25
H12	VAM with SE	.23	.22	.37	.21	.34	.34
	VAM without SE	.42	.36	.41	.35	.48	.48
H13	VAM with SE	-.01	-.15	-.11	-.23	-.19	-.15
	VAM without SE	.22	.17	.18	.25	.16	.21
H14	VAM with SE	.01	.10	.23	.15	.08	.17
	VAM without SE	.19	.27	.16	.09	.17	.29
H15	VAM with SE	-.10	-.01	.03	.00	.28	.05
	VAM without SE	.07	.11	.15	.08	.25	.15
H16	VAM with SE	.05	.23	.14	.27	-.18	.10
	VAM without SE	.23	.27	.18	.30	-.05	.19
EM1	VAM with SE	-.01	.24	.28	.03	.18	.23
	VAM without SE	.10	.13	.23	-.10	.24	.20
EM2	VAM with SE	.11	.18	.19	-.02	.09	.15
	VAM without SE	.17	.12	.08	-.06	.02	.08
M1	VAM with SE	.10	.30	.15	.10	.05	.18
	VAM without SE	.11	.40	.29	.25	.17	.30
M2	VAM with SE	-.21	-.36	-.30	-.07	-.34	-.32
	VAM without SE	-.12	-.22	-.26	-.06	-.30	-.24
M3	VAM with SE	.12	.11	.23	.13	.22	.19
	VAM without SE	.36	.40	.33	.27	.42	.40
M4	VAM with SE	.03	.07	.14	.18	.02	.11
	VAM without SE	.25	.37	.40	.21	.36	.43
M5	VAM with SE	-.43	-.36	-.14	-.12	-.32	-.30
	VAM without SE	-.32	-.23	-.07	-.08	-.24	-.20
M6	VAM with SE	.10	.06	.11	-.05	.08	.09
	VAM without SE	-.04	-.04	.02	-.10	-.10	-.04
M7	VAM with SE	.01	.03	.10	.27	.01	.07
	VAM without SE	.07	.09	.14	.27	.12	.14

Level	Type of VAM Score	Factor 1.1	Factor 1.2	Factor 2.1	Factor 3.1	Factor 3.2	Total Score
M8	VAM with SE	-.11	-.32	.03	-.21	-.04	-.16
	VAM without SE	-.05	-.22	-.04	-.33	-.26	-.18
M9	VAM with SE	.16	.10	.15	.19	.15	.16
	VAM without SE	.34	.20	.16	.31	.40	.28
M10	VAM with SE	.09	.05	-.03	-.04	.00	-.16
	VAM without SE	.13	.10	.02	-.06	.00	-.07
M11	VAM with SE	.10	-.06	-.11	-.16	-.10	-.07
	VAM without SE	.28	.19	.14	-.01	.14	.18
M12	VAM with SE	-.38	-.26	-.19	-.08	-.16	-.24
	VAM without SE	-.31	-.22	-.24	-.17	-.17	-.25
M13	VAM with SE	.16	.03	.14	.04	-.28	.02
	VAM without SE	.25	.28	.30	.03	-.10	.22
M14	VAM with SE	-.24	-.27	-.22	-.08	-.16	-.25
	VAM without SE	.14	.04	.05	.03	.11	.08
M15	VAM with SE	.03	.00	-.01	.21	-.01	.02
	VAM without SE	.04	.17	.10	.35	.14	.15
M16	VAM with SE	-.14	.05	.04	.04	-.17	-.03
	VAM without SE	-.08	.08	.15	.06	-.19	.03
M17	VAM with SE	-.01	.02	.05	-.13	.05	.03
	VAM without SE	.26	.05	.15	-.24	.01	.10
M18	VAM with SE	-.15	-.19	-.19	-.20	-.20	-.20
	VAM without SE	-.20	-.11	-.14	-.22	-.10	-.15