1-1-2014

Validity Of The Educator Evaluation Instrument In The State Of West Virginia

Carla Howe
Wayne State University,

VALIDITY OF THE *EDUCATOR EVALUATION INSTRUMENT* IN THE STATE OF WEST VIRGINIA

by

CARLA HOWE

DISSERTATION

Submitted to the Graduate School

of Wayne State University,

Detroit, Michigan

in partial fulfillment of the requirements

for the degree of

DOCTOR OF PHILOSOPHY

2014

MAJOR: EDUCATIONAL EVALUATION AND RESEARCH

Approved by:

_____

Advisor Date

_____

_____

_____

DEDICATION

For my first and greatest teacher of all, my mother, who continues to be my strongest supporter and most honest critic.

For all of the phenomenal teachers with whom I had the pleasure to work and teach including Jenny Birmelin, Josh Kremer, Tami Lamerato, Scott Long, Rick McCoy, Kim Muncie, Corey Rogers, and Jianna Taylor.

For all of the teachers who were part of my educational career, and for those who work tirelessly to ensure their students learn.

For my colleagues at State Education Agencies who work to support systems, educators, and policymakers to make the informed educational decisions that impact students and student learning.

For the one who changed everything: Juan D'Brot.

A**CKNOWLEDGMENTS**

First and foremost, I would like to thank Dr. Shlomo Sawilowsky, my advisor, who changed my professional trajectory in countless ways, but perhaps most notably just after signing my Master's Thesis. Never hesitant to get right to the point, he put down the pen and said something like, "So when is it that you will begin your doctoral studies?" Before Dr. Sawilowsky asked this question, I did not know it was even possible for me to earn my doctoral degree while working full time. I feel fortunate and grateful that he raised this question—and to have been his student and advisee throughout my studies at Wayne State. He has helped prepare me well for work in the field of Educational Evaluation and Research.

I would also like to thank my committee members, Dr. Ingrid Guerra-López, Dr. Venessa A. Keesler, and Dr. Mark J. Larson, for the time and effort they put in to pushing my work further, and for being part of what will certainly be one of my greatest accomplishments.

I must also mention and thank Kim Reynolds, my classmate at Western Michigan University. It was through Kim's mention of an open position as a research assistant that I applied for a job at The Evaluation Center in Kalamazoo, Michigan. As a research assistant under the guidance of Dr. Gary Miron, I began to see the possibilities of influencing educational change beyond the classroom through research and evaluation. It was there that my interest in this field of study began, and has continued.

Lastly, I'd like to thank my classmate and friend, Jack Sawilowsky, for being part of my success as a doctoral student—and for all the laughs and good times we had throughout our coursework.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

Validity of the *Educator Evaluation Instrument* in the State of West Virginia

Chapter 1

Introduction

In 2009, President Barack Obama introduced the Race to the Top (RTTT) grant funding opportunity for State Education Agencies (SEAs) whereby they could commit to certain assurances for reform-related change. The most notable reform was the introduction of educator evaluations based on student growth measures. Education department officials from many states applied for these grants including officials from West Virginia. However, the state was not awarded the Race to the Top funding like the majority of other states for which an application had been submitted (Hamilton, 2010). Nevertheless, the application process set forth by the U.S. Department of Education required that applicants begin to implement these reforms regardless of whether the SEA was a grant recipient.

Hence, officials at the West Virginia Department of Education (WVDE) began developing a revised educator evaluation system. Beginning in the school year (SY) 2011-2012, the WVDE began piloting an educator evaluation system for teachers of grades Kindergarten through 12. In the first year of the pilot, 25 schools across the State took part in the implementation of a system that included multiple measures, including student growth measures (Meharie & Hixson, 2013). In its second year, the demonstration year, over 100 schools participated.

In 2012, the West Virginia Governor Earl Ray Tomblin approved a bill that required the implementation of a statewide educator evaluation system to begin in the 2013-2014 school year. This statewide teacher evaluation system is required to include student-learning

growth as one of the measures included in the overall summative evaluation (West Virginia Educator Evaluation System for Teachers, 2013).

In addition to the 2012 Legislative session, the need arose to continue to refine the educator evaluation system when officials at the WVDE saw an opportunity to waive some of the restrictions of the No child Left Behind (NCLB) Act of 2001, including the requirement for 100% of students to be proficient on the statewide assessment by 2014. This application, submitted to the U.S. Department of Education in September 2012, would allow the WVDE to waive many of the NCLB restrictions if approved, and set realistic, attainable, yet still challenging goals within the Accountability system (Index Page for the Elementary and Secondary Education Act (ESEA) Flexibility Page, 2013). Applying for Flexibility was appealing because at the time of application, the 100th percentile school in West Virginia was only 75% proficient, making the 100% proficiency in only two years an impossible goal. The application for Flexibility required areas of reform in exchange for the flexibility in three areas: (1) the implementation of college and career-ready standards; (2) an accountability system that considers achievement, growth, and achievement gaps; and (3) supporting the growth of all educators through a comprehensive evaluation system.

The evaluation system was revised after its first year to support educators as part of a comprehensive system of support. The specific purposes of the system included

- setting high standards of performance for both veteran and new teachers;
- ensuring high-quality instruction focused on increasing student achievement;
- encouraging continuous growth and improvement over time.

In an effort to ensure consistency across educator preparation, professional development and professional practice, the system was aligned with the West Virginia Professional Teaching Standards, which are the foundation of the profession (West Virginia

Department of Education, 2013). Included in the system were the seven teaching standards by which teachers are measured within the educator evaluation system. The standards are as follows:

- Standard 1 – Curriculum and Planning
- Standard 2 – The Learner and the Learning Environment
- Standard 3 – Teaching
- Standard 4 – Professional Responsibilities for Self-Renewal
- Standard 5 – Professional Responsibilities for School and Community
- Standard 6 – Student Learning
- Standard 7 – Professional Conduct (West Virginia Professional Teaching Standards, 2013).

The instrument, developed to measure the effectiveness of the teachers as part of the evaluation system, was based on several different measures and components. All teachers are measured with the same instrument that utilizes seven West Virginia Professional Teaching Standards as the main components through observations by principals and evaluators and through the submission of evidences that can be submitted such as lesson plans and portfolios. Although the instrument, based on the West Virginia Professional Teaching Standards, is the same for all teachers the degree of observations and types of evidence required differ depending on a teacher's experience.

Teachers are put into one of three categories based on their number of years teaching: Initial, Intermediate, or Advanced. Teachers who are considered to be in the Initial Progression are those teachers in their first through third years of teaching. Teachers identified in the Intermediate Progression are those with four to five years of teaching experience. Teachers in the Advanced Progression have six years or more of teaching experience.

Depending on the Progression of the teacher, some of the requirements differ. The primary difference is the number of observations required, which lessen as the teacher moves

through the progression. Teachers in the Initial Progression are required to have four observations, which decrease to a required two observations in the Intermediate, then down to none required for the Advanced Progression.

Eighty percent of the overall summative effectiveness score comes from the Professional Teaching Standards 1-5. Each of the first five standards have an equal weight of 17.14%. For these five standards, educators are evaluated by an administrator who then makes a determination of the educator's effectiveness from four categories: Distinguished, Accomplished, Emerging or Unsatisfactory. Each of these categories is defined in a rubric that the evaluator is to use to score the teacher for each standard. Evidences for each of the five teaching standards include Observations, Assessments, Student Feedback, Student Work Samples, and Portfolios in addition to a variety of other evidences, which can be found in Appendix A. Evidences, in addition to the observational information, can be brought forth if the educator disagrees with a determination made by an evaluator.

Standard 6 makes up 20% of the total score for the educator, but is broken down into two parts: Student Learning Goals and Standardized Growth. Student Learning Goals make up 15% of the overall 20%, which the teacher establishes. As described in the West Virginia Educator Evaluation System for Teachers guidance documentation (2012), setting the student learning goals "is standardized for all educators with quality checks to ensure that student learning is part of an overall educator evaluation that is rigorous, consistent and equitable. All evidence for the Student Learning performance standard must meet three criteria that are based on federal requirements:

1. Two data points
2. Rigorous
3. Comparable across classrooms."

These criteria help ensure that the student learning goals that teachers are setting consider at least two data points. This is very important because the teachers' plans must be consistent with these two data points to be able to show measurable progress and a change in student learning. This means the teacher will have to select, in advance, the assessment that will be used to measure the student learning that took place to achieve the goal set.

The second criteria required, rigorous, refers to the assessments that are included within the student learning goals. All assessments must be aligned with the West Virginia Next Generation Content Standards and Objectives.

The intent of assessments comparable across classrooms is for teachers to select, as part of the student learning goal, an assessment that could be used widely across classrooms within a grade. In this example, it could mean the use of district developed common assessments for each grade level in English Language Arts or Socials Studies.

The remaining 5% is determined by Standardized Growth, which is further broken down into 2.5% for a reading score and 2.5% for mathematics. These scores are results of school-level growth data that are determined within the West Virginia Accountability Index (WVAI), which are based on the statewide-standardized assessment results.

Standard 7 is a required component of the system in that a teacher must demonstrate professional conduct as described in the rubric; however, no evidences are required to be brought forth specifically for this standard unless a teacher is not adhering to the performance standard.

The weighting of the standards within the instrument varies as shown in Table 1.

Table 1

Components and Weights of the Educator Evaluation Instrument

Component	Weight	% of Score
Standard 1: Curriculum and Planning	17.14%	80%
Standard 2: The Learner and the Learning Environment	17.14%	
Standard 3: Teaching	17.14%	
Standard 4: Professional Responsibilities for Self-Renewal	11.44%	
Standard 5: Professional Responsibilities for School and Community	17.14%	
Standard 6: Student Learning		
Student Learning Goals		15%
Standardized Growth	2.5% mathematics 2.5% reading	5%
Standard 7: Professional Conduct		Required, but does not count in the overall score
Total		100%

To determine the overall summative performance level of the teacher, the rubric results from the six teaching standards described above are each multiplied by the appropriate weight. The results are totaled up and the overall score will fall into one of four performance levels: Distinguished, Accomplished, Emerging, or Unsatisfactory.

The instrument, as part of the educator evaluation system will be put in place statewide for the purposes of

- setting high standards of performance for all teachers;
- ensuring high-quality instruction focused on increasing student achievement;

- encouraging continuous growth and improvement over time.

In the pilot year implementation of the system, the *Educator Evaluation Instrument* was determined to have reliability of .745. Given that the reliability of the instrument is high, determining the construct validity of the instrument is a natural next step to assure teachers, administrators, parents, policymakers, and developers at the SEA that the decisions made about teachers based on the results of this instrument are accurate. These decisions vary from determining appropriate placement of teachers to the targeted professional development to address areas of weakness. As such, for this instrument to be utilized at the state level, and for the purposes for which it was intended for all teachers, determining the construct validity of the instrument is very important so that monies spent on professional development, either by the teacher or the LEA, are spent appropriately.

In addition to understanding whether the instrument can aid in and for decisions about teachers' placement and targeted professional development, this instrument will be utilized statewide for a multitude of reports for the legislature and the State Board of Education. Both of these audiences drive policy decisions that impact schools and their teachers; therefore determining the validity of the instrument, that it measures what it purports to measure—educator effectiveness—is ultimately necessary so that these audiences can make informed decisions based on the instrument.

In West Virginia, where concern has arisen based on the evaluation results, a teacher may be placed on a Focused Support Plan, or in the case where a teacher received an unsatisfactory rating in one or more of the standards, he or she would be required to develop and adhere to a Corrective Action Plan. Yet, in other states like Tennessee, Delaware, and Michigan, many high-stakes decisions are determined from the results of these evaluations

including merit or performance pay, retention/placement, suspension, and dismissal (Databases on State Teacher and Principal Evaluation Policies, http://resource.tqsource.org/stateevaldb/Compare3States.aspx). Although these high-stakes decisions are not currently part of the educator evaluation system in West Virginia, if the laws and policies were to ever change, having certainty that the instrument used to determine evaluation ratings is valid would be necessary, particularly in a court of law or in the cases of grievances.

Perhaps more importantly, the comprehensive system of support is it meant to be part of what will help educators demonstrate growth through targeted professional development and critical feedback from school leaders. Without a valid instrument, all of the efforts that educators across the state are exerting to implement the system will do little good at ultimately improving student achievement.

Lastly, understanding the validity of this instrument will aid West Virginia during the expansion of the evaluation system statewide during the 2013-2014 school year and after. Having a greater understanding the validity of the instrument itself will help inform future decisions, including possible revisions, about the evaluation instrument, the evaluation system, and the high-stakes implications of the results rendered from the system.

Purpose of the Study

To determine the construct validity of the West Virginia *Educator Evaluation Instrument*, a structural equation model utilizing hierarchical confirmatory factory analysis based on the data from the second pilot year will be developed. The data to be utilized will include the evaluation results of over 3,000 de-identified teachers across the state of West

Virginia. Additional evidence regarding the construct validity of the instrument for determining educator effectiveness will be collected.

Limitations

The sample was not selected randomly, and therefore it is not known to what extent these 3000 teachers are representative of the state's educators in terms of teaching experience or school performance. They represent 15% of the states 20,000 educators, but in the absence of random selection, a large sample size cannot be relied on as being representative.

Assumptions

It is that equal training was provided to the participating districts and evaluators through WVDE training sessions. It is also assumed that the follow-up support and training provided by the Regional Educational Service Agencies (RESAs) was equal across the state.

Definitions

Common Assessments are assessments developed and/or given at a district or school-wide level in a given grade and subject that is often scored collaboratively by teachers, and which the results are often analyzed collaboratively to understand student learning, pacing efforts, and planning.

The Elementary and Secondary Education Act (ESEA) is a U.S. federal legislation enacted in 1965. The ESPEA was enacted as a part of the "War on Poverty" and it is the most far-reaching federal law affecting education. The Act was originally authorized through 1970, however the government has reauthorized the Act every five years since its enactment. The current reauthorization of ESEA is the No Child Left Behind Act of 2001.

ESEA Flexibility is the opportunity from the U.S. Department of Education for SEAs to request flexibility regarding specific requirements of the No Child Left Behind Act of

2001 (NCLB) in exchange for rigorous and comprehensive State-developed plans designed to improve educational outcomes for all students, close achievement gaps, increase equity, and improve the quality of instruction.

A Local Education Agency (LEA) is the legal term for a school district.

Race to the Top is a competitive grant program to encourage and reward States that are implementing significant reforms in the four education areas described in the ARRA: enhancing standards and assessments, improving the collection and use of data, increasing teacher effectiveness and achieving equity in teacher distribution, and turning around struggling schools.

Regional Education Service Agency (RESA) are an administrative agency that supports LEAs within its jurisdiction. In West Virginia, there are eight RESAs that serve all 55 counties.

State Education Agency (SEA) is the legal term for a State Department of Education like the West Virginia Department of Education.

The West Virginia Accountability Index (WVAI) is an assignment of points to schools for progressively higher performance on a balanced set of metrics that will be in place no later than the 2013-14 school year.

The West Virginia Next Generation Content Standards and Objectives are the standards that define the knowledge and skills students should have within their K-12 education careers so that they will graduate high school able to succeed in entry-level, credit-bearing academic college courses and in workforce training programs.

Chapter 2

Review of Literature

In 2001, with the passing of the No Child Left Behind (NCLB) Act, teachers were required to be demonstratively highly qualified in order to teach. This required them to pass content-specific tests indicating their competence in their subject area, along with meeting requirements for collegiate coursework. Even teachers who had received their license or certification prior to the NCLB Act of 2001 were required to demonstrate subject-matter competency in other ways such as providing a portfolio of types of evidences including experience and professional training to become highly qualified (New No Child Left Behind Flexibility: Highly Qualified Teachers, 2004). However, implementing this requirement was not the same as being an effective instructor, nor was it sufficient for ensuring that an effective teacher was in front of students (Palardy & Rumberger, 2008). Over the course of ten years of the NCLB Act being in place, across the country very few state assessment results showed evidence of closing achievement gaps between defined subgroups of students (i.e., White, Black, Hispanic/Latino, Asian/Pacific Islander, Student with Disabilities, Low Income, English Language Learner), which was a main goal of NCLB (NCLB Not Closing Test Score Gaps, 2006).

This led researchers, educators, and policymakers to seek a solution being a teacher being credentialed as highly qualified, because there was little evidence that this credential was enough to make a difference in student achievement gains. Darling-Hammond (2000) found that a teacher's effect on students is additive and cumulative, so if a student has several highly effective teachers in a row, students will show greater gains and the ability to demonstrate achievement; the opposite is also true. Students who experience two to three

unsatisfactory teachers in a row show serious, long-term deficits (Peterson & Peterson, 2006).

This also became a lever for change in how teachers have come to be evaluated across many states. Although being highly qualified is necessary, it is insufficient to determine whether a teacher would be effective in the classroom. Hence, the questions arise: How can a teacher be determined to be effective? How can teacher quality be assured if not through a mechanism like the highly qualified process?

With these questions came the impetus for new educator evaluation systems across states that relied on more than one measure—and certainly more than a status measure. New systems were put into use in schools and districts across the country, each with their own evaluation component.

Determining Construct Validity

Because new educator evaluations systems were put into place so suddenly, concerns about the validity of the instruments for determining effectiveness were raised and continue to be raised (Databases on State Teacher and Principal Evaluation Policies, 2013). For this reason, the construct validity of the *Educator Evaluation Instrument* must be determined through several different mechanisms. "Validity refers to the degree that a test measures what it purports to measure" (Sawilowsky, 2007, p. 178). Construct validity is the extent to which an instrument measures "a fiction used to explain reality" (Sawilowsky, 2007, p. 178), such as aptitude or intelligence). Hypothetical constructs, like educator effectiveness, are not directly observable, and can therefore only be measured indirectly through observed scores or other indicators. Moreover, construct validity depends on the theoretical understanding that underpins the constructs by which the instrument was built. In this case, it is the

investigation into the various constructs of the *Educator Evaluation Instrument* and whether the theoretical understanding of educator effectiveness is being measured appropriately. Ensuring a high level of construct validity is an important for determining appropriate use and decision-making based on the results that stem from the instrument. Validity is important for teachers being evaluated, for those evaluating teachers—whether they are superintendents, principals or peer coaches—and also for the public and for policymakers who are interested to know how teachers are performing. The balance, then, it to ensure that an instrument is valid for what it intends to measure—educator effectiveness.

Validity can be established through statistical analyses. Confirmatory factor analysis (CFA) is a statistical method to determine validity through the examination of factor loadings of each parameter within a structural equation model (SEM). Factor loadings estimate the direct effects on indicators, which are then interpreted as regression coefficients (Kline, 2011).

Hierarchical Confirmatory Factor Analysis (H-CFA) is utilized to examine the relationship among the constructs through higher-order factors with presumed direct causal effects on lower-order factors. The first-order factors are the teaching standards included as part of the instrument, which measure the second-order factor of educator effectiveness. Through H-CFA the second-order factor, educator effectiveness, is measured indirectly through the first-order factors, the professional teaching standards. In order to utilize H-CFA, there must be at least three first-order factors for the model to be appropriately identified. Each of the first-order factors must have at least two indicators. The *Educator Evaluation Instrument* has six first-order factors with at least two indicators, or sub elements, per indicator rending the model identified (Kline, 2011). For the instrument to be valid, each

construct, or first-order factors should load relatively equally on the overall second-order factor, educator effectiveness, as they are weighted equally in the instrument. Additionally, the correlation between the second-order factors should be low, which indicates that each factor is a factor that contributes to the overall effectiveness independently.

In addition to the quantitative measures that determine validity of the instrument, validity can be supported through the examination of evidences theoretically using a framework that defines the construct. Although this evidence gathering process is weak without the accompaniment of the statistical analyses described above, Danielson (2008) argues that for evaluations of teaching to be valid, the data collection instruments must be developed based on a clear definition of good teaching practices rendering this examination and evidence gathering is an important step in establishing whether the *Educator Evaluation Instrument* is valid. Although examining the definitions of good teaching practices upon which the instrument is built would render a weak analysis on its own, the examination of whether the constructs are logical do provide support in determining the validity of the instrument.

Validity

Discriminate capability includes the need for mutual exclusivity and low ambiguity between categories within an instrument, which means each construct should be clearly defined and a separate measure as part of the instrument. This will be determined through the multi-group, hierarchical confirmatory factor analysis (H-CFA). Through the SEM model, the first-order factors should load equally and with positive numbers, but have low correlations. If the factors, or standards, are not mutually exclusive, there would like be a threat to validity because it means that each standard shares qualities with the other

standards; if it cannot be measured as a single, mutually exclusive component of an instrument, the measure itself may not be valid.

Important Practices within an Educator Evaluation System

Although determining the construct validity of the *Educator Evaluation Instrument* is the focus of this study, there are several important components, practices and uses that are part of the educator evaluation system, which require explanation as they are interwoven with the instrument itself, and the basis for how the instrument was developed. These include:

- Teacher-student rostering (i.e. creating a data link between teachers and his/her students) mechanism
- Multiple measures
- Teacher experience
- Observation and artifacts
- Teacher self-reflection
- Student growth measures
- Student growth models

A teacher-student rostering mechanism.

The use of a teacher-student rostering mechanism is necessary for determining validity of an instrument (Odden, 2004, p.134) that purports to determine educator effectiveness because the measures that are to be included in the instrument that quantifies effectiveness must be based on students for which the teacher instructed. Providing a way for teachers to indicate whether the students received instruction will mean that any student achievement or growth scores that will be attributed to the teacher will be based on the correct group of students. Many state departments of education have developed a teacher

rostering mechanism as a required practice within their educator evaluation system (Hawaii DOE | Educator Effectiveness, 2013; Florida Department of Education, 2013; Teacher and Leader Effectiveness (TLE), 2013; Roster verification., 2013). As explained on the opening page of Public Schools of North Carolina Educator Effectiveness (2013) website, roster verification is the

> process that will ensure that teachers are accurately linked to the students they teach. Roster Verification is simply a way for teachers to verify their class rosters and allow schools and teachers to indicate when there are multiple professionals sharing responsibility for a student's instruction.

This practice provides the opportunity, if applicable, for teachers within a district to apply local business rules to rosters. In Michigan's educator evaluation system, for example, a pupil may be removed from the student growth measure component as follows:

> g) The performance evaluation system may allow for exemption of student growth data for a particular pupil for a school year upon the recommendation of the school administrator conducting the annual year-end evaluation or his or her designee and approval of the school district superintendent or his or her designee, intermediate superintendent or his or her designee, or chief administrator of the public school academy, as applicable (HB 4627, 2001).

This process allows for the removal of a student who may have been expelled without services, or who may have been placed in a full-time care, or other institution even though the student may have been required to be enrolled in the school, and would possible still remain linked to the teacher's roster. The inclusion of hand verifying the students for whom the teacher is accountable is an important practice because teachers are provided the opportunity to verify which students' growth scores are attributed to the growth measure component of the evaluation. This process helps to ensure the validity of use of the instrument.

Multiple measures.

The shift in determining educator quality and effectiveness indicate that teacher evaluation should be built around a multitude of measures (Danielson & McGreal, 2000; Darling-Hammond, 2000). Educator evaluation systems that determine an educator's effectiveness should recognize "student achievement, acknowledges good practice, supports teacher goals, shapes performance, motivates to improve on weaknesses, and removes the rare bad teacher from the profession" (Peterson & Peterson, 2006, p. 1).

Darling-Hammond (2000) stated the following are essential for determining a teacher's effectiveness: knowledge of teaching and learning, teacher experience, and certification status (p. 5), where knowledge of teaching and learning is defined as a teacher's understanding of pedagogy and its application in the classroom. In the National Commission on Teaching and America's Future report (1996), *What Matters Most: Teaching for America's Future*, stresses the importance and necessity for a teacher to possess content knowledge and pedagogical knowledge. The understanding of teaching and learning, means not only know what and how to teach a student, but also knowing the *kinds* of mistakes students are going to make—and then what needs to be done to help them understand (Ball & Bass, 2000). The professional teaching standards as constructs not only form the evaluation system built on multiple measures, but it reinforces findings that knowledge of teaching and learning matters for teacher effectiveness.

Having a multitude of measures as part of an evaluation system recognizes the complexity in determining a teacher's effectiveness in that a single measure would be insufficient. States across the county are implementing systems that take into account a teacher's professional practices and student growth measures in varying degrees. State models of educator evaluation across the county vary, but there are many similarities in the

components that they contain. In Colorado, Ohio, Georgia, and Maryland, 50% of an educator's evaluation is based on student growth measures and 50% is based on professional practices (Colorado Department of Education State Model Evaluation System for Teachers, 2013; Ohio Department of Education Teacher Evaluations, 2013; Student Growth Percentiles – Georgia's Student Growth Model, 2013; Maryland state model, 2013). In Delaware, there are five components that compose the educator evaluation system including

- Planning and Preparation
- Classroom Environment
- Instruction
- Professional Responsibilities
- Student Improvement (DPAS II - Delaware Performance Appraisal System, 2013)

Although it appears that the categories extend beyond professional practices and student growth measures, the first four categories could be described as professional practices, and student improvement is akin to student growth measures.

In Oklahoma, there are two approved systems of evaluation available to districts: the state model, or the Tulsa model. The Tulsa model is based on five components that are weighted differently within the system.

- Classroom management - 30%
- Instructional effectiveness - 50%
- Professional growth and continuous improvement - 10%
- Interpersonal skills - 5%
- Leadership - 5% (Tulsa Public Schools TLE Observation and Evaluation Rubric Teachers, 2012)

Within the Tulsa model, Instructional Effectiveness is weighted at 50%, which includes some measures of student growth.

Teacher experience.

A third important practice of an educator evaluation system is to design it such that it is differentiated to account for teachers with varying levels of experience (Danielson, 2008). Teachers with more than three years of experience are more effective than those with three years or fewer (Nye, Konstantopoulus, & Hedges, 2004), which suggests that experience through the act of teaching matters in terms of effectiveness; therefore having a system that recognizes experience is optimal in determining effectiveness. Teachers develop and grow at different rates, taking from five to eight years to master the art and science of teaching (Darling Hammond, 2000) The system of evaluation may differ for Novice or non-tenured teachers from experienced or tenured teachers. Novice teachers may require a prescribed number of observations; whereas, there may be a system for Experienced or tenured teachers that requires a set number of formal observations on a rotating schedule (e.g. every other year), and then self-directed, or self-determined professional growth periods. For these tenured teachers, a periodic comprehensive evaluation would be conducted in the hopes of affirming the experienced teacher's practice. A differentiated system would help to ensure that the experienced teachers are still observed and provided with feedback, but would also free up the principal (or other key evaluators) to focus on the novice teachers for whom more focused attention may be beneficial.

Observation and artifacts.

Although a system should be built around a multitude of measures, Danielson (2008) argues that there are two critical components to an educator's evaluation: observation and

artifacts. The Florida Department of Education (2013) recognized the need for multiple observations of a teacher to gain a comprehensive understanding of his/her teaching practices. Observations are the mechanism to gather important information about a teacher's practice. Observation data can be used in both a formative and summative sense by the teacher and the observer or evaluator to discuss areas for improvement or areas of excellent. The types of observations described within the Florida Department of Education's system include formal, informal and pop-in classroom observations.

Artifacts are another mechanism for gathering important educator effectiveness data because they offer the best and possibly only evidence of certain aspects of teaching (Danielson, 2008). These may include things like planning documentation— both single lesson and long-term planning. These are critical skills for teachers to possess, but show very different skills on the part of the teacher. Only a long-term unit plan can show how the teacher will address teaching standards and how he/she intends to engage students in the learning of large, complex ideas.

Teacher maintenance of records is also important. Observations will not get to this very important aspect of a teacher's responsibilities without simply requiring the collection of this type of evidence. Teachers' maintenance of records is very important for understanding students' learning and growth over time.

Another artifact that is essential in understanding a teacher's effectiveness is how they communicate with parents and families. This would not be evident in an observation, and therefore should be collected as an artifact. Obviously keeping families informed of student progress is key to helping students learn, stay on track, and grow as students over the course of the year. A teacher should be able to show the ways in which he/she stays in contact with

families—whether through emails, newsletters, progress reports, an up-to-date, or online grade book.

Evidence of instruction is another way that artifacts can be used. Teachers can present evidence of student learning from not only the assignment given, but the students' work as well. This act can help an evaluator determine the way in which the teacher plans, assesses, and evaluates the students. Without this type of artifact, it would be difficult at best to know whether a teacher differentiates instruction for a multitude of learners, whether he/she adjusts and adapts the lesson based on student assignments, or whether the teacher just plows ahead with his/her own agenda regardless of the student outcomes.

Teacher self-reflection.

Self-reflection is another aspect of an evaluation system that make it meaningful to teachers. Teachers whose students have high achievement rates continually mention reflection on their work as an important part of improving their teaching (Mitchell, 1998). Additionally, the self-reflection as a component in an evaluation system requires the teacher to take an active role in the evaluation process, which increases the value of the evaluation process for teachers (Danielson & McGreal, 2000). This practice enhances the validity aspect of the instrument as well because the teacher as a way of taking ownership of his/her teaching practices and evaluation. It allows the teacher to be thoughtful and deliberate about his/her personalized plans for continued professional growth (West Virginia Department of Education, wvde.state.wv.us/evalwv). This is still important because teachers rate analyzing and seeking to improve their own teaching as an important factor in their teaching effectiveness (Covino & Iwanicki, 1996).

Student growth measures.

Although experts believe that there are about 5% of unsatisfactory teachers in the teaching population, noneducators assume this percentage to be higher (Person & Peterson, 2006). And, although an elaborate evaluation system is rarely needed for principals to identify unsatisfactory teachers, for credibility sake, objective data, such as student growth measures, are a necessary component of evaluation systems across the country (Peterson & Peterson, 2006). For the past several years, the primary goal of growth analyses has been to determine the amount of student progress that could be attributable to a school or teacher based on complex statistical techniques (Betebenner, 2009). Growth is considered to be an increase in something over time—and in the same way that a child can grow in height, so can he/she in knowledge and achievement (Catellano & Ho, 2013).

Given that growth measures on students show how much gain they have made from one point in time to another, they have become an increasingly important part of educator evaluations across the county. Goe (2007) notes the shift in how evaluations were conducted 30 years ago--that achievement results of pupil were rarely considered within the evaluation; now, it is common practice. Policymakers have come to believe that the failure of evaluating teachers systematically and meaningful in the past can be remedied by calculating growth and achievement measures from standardized test scores (Rothstein et. all, 2010).

Naturally, policymakers have become more involved in school reform efforts, which have required student test scores and student growth in educator evaluations. They argue that a teacher evaluation system should include measures of student achievement for the system itself in order to have any amount of credibility with these audiences (Peterson & Peterson, 2006). They want student achievement data included as an indicator of teacher and school

quality, and because these audiences have expressed concerns, and raised questions, these indicators are part of most state evaluation systems. In fact, 48 of 50 states are mandated by law or policy to include some type of growth measure into the teacher evaluation system (Databases on State Teacher and Principal Evaluation Policies, 2013). Part of the necessity for including student assessment measures is to have an objective measure, one that is not based on an observation by an administrator. Traditional educator evaluations based on the satisfactory/unsatisfactory model were criticized for showing nearly all teachers as satisfactory despite low tests scores and poor student performance. And, although test scores do not capture all facets of student learning, student assessment scores are an available measure and recognized as an important indicator of achievement by educators, policymakers, and the public (Nye, 2004). For this reason, growth measures are a component to some degree in nearly every state's educator evaluation system, may by law (Databases on State Teacher and Principal Evaluation Policies, 2013).

Student growth models.

The growth component of an evaluation system is by far the most controversial with strong opinions by educators and scholars being voiced, however the foundation of these models is essentially the same—to understand student achievement based on student assessment scores (Betebenner, 2009). There are several different types of growth models that SEAs and districts are utilizing as components of their evaluation systems. Castellano and Ho (2013) describe the three main types of growth models being utilized: gain-based, conditional status, and multivariate models (21-22).

Gain-based models are those that take into consideration the gains students make from one point in time to another. For these types of growth models to function properly, the

assessment from which the results are derived must be based on a vertical scale, or a common scale where scores across grades can be compared (Castellano and Ho, 2013)

Conditional status models support a student's conditional status that is framed by a question or a particular context (Castellano & Ho, 2013). Conditional status models utilize past information to contextualize the student's current status by answering the question, "what can be said of a student's current achievement level given their prior achievement?" (Betebenner, 2009, p. 43). In Michigan, for example, a student's growth or improvement is determined by her prior year's score against the current year's score. Because there is no underlying vertical scale in the Michigan Education Assessment Program (MEAP), a transition table was developed to indicate whether a student significantly improved, improved, maintained, declined, or significantly declined based on the prior year's assessment score. In this manner, the transition table acts a conditional status model to indicate whether a student has shown growth, maintained (neither grew nor declined), or declined. The statistical model that undergirds the transition table is based on determining cut scores for the ranges of students. As shown in Figure 1, for a student who was not proficient, and whose score was in the middle, or "Mid" range in the prior year, and who score was partially proficient in the low category, he would have shown improvement.

Figure 1. Michigan Education Assessment Program Transition Table

Year X Grade Y MEAP Performance Level		Year X+1 Grade Y+1 MEAP Performance Level								
		Not Proficient			Partially Proficient		Proficient			Adv
		Low	Mid	High	Low	High	Low	Mid	High	Mid
Not Proficient	Low	M	I	I	SI	SI	SI	SI	SI	SI
	Mid	D	M	I	I	SI	SI	SI	SI	SI
	High	D	D	M	I	I	SI	SI	SI	SI
Partially Proficient	Low	SD	D	D	M	I	I	SI	SI	SI
	High	SD	SD	D	D	M	I	I	SI	SI
Proficient	Low	SD	SD	SD	D	D	M	I	I	SI
	Mid	SD	SD	SD	SD	D	D	M	I	I
	High	SD	SD	SD	SD	SD	D	D	M	I
Advanced	Mid	SD	SD	SD	SD	SD	SD	D	D	M

There are some advantages to conditional status models such as this because it is easy for educators and principals to utilize because no expertise in calculating complex statistical models is necessary—that work was done to create the transition table (MDE - Michigan Educational Assessment Program, 2013).

Another type of conditional status model is the calculation of student growth percentiles. A student's growth percentile describes how normal or abnormal a student's performance is relative his/her academic peers, that is, students whose past academic performance is similar to that of the student (Betebenner, 2009). Student growth percentiles (SGPs) are descriptive in nature because, as Betebenner (2009) argued that stakeholders actually want to know the normative context that helps them understand what the information means in terms of other students at this grade level and in this subject area rather than a precise, statistical measure. With SGPs, a student is considered to show growth if they are performing better than most of his or her academic peers; the opposite would also be true. SGPs are to be utilized at a state-level in order to have a large enough peer group for which a

comparison can be made. Betebenner, the founder of student growth percentiles (SGPs) created open source R-language to be able to run SGPs at the state education agency level (Betebenner, VanIwaarden, Domingue, & Shang, 2013). States including Colorado, New Jersey, Massachusetts, Virginia, Georgia, Washington, and West Virginia calculate student growth percentiles from their state assessment results. (Student Growth Percentiles, 2013).

The third grouping of growth models is the multivariate models, which are calculated to determine the estimates of the value-add of a school or classroom teacher. These types of models are very complex and can require proprietary software like that developed by SAS for states such as Tennessee and North Carolina (Tennessee Department of Education, 2013; SAS Institute Inc., 2013). The value-added growth model in Tennessee is described in the following manner:

> "The Tennessee Value-Added Assessment System (TVAAS) measures the impact schools and teachers have on their students' academic progress. TVAAS is a powerful tool because it measures how much students grow in a year, and shines more light on student progress than solely considering their score on an end of year test. Furthermore, TVAAS only measures what a school can control. Educators are only held accountable for the things that they can control, such as their students' academic progress during the school year. Teachers are not held accountable for the things they cannot change, such as their students' previous achievement."

Betebener (2009) argued, however, that there is a disconnect with value-added measures because they do not truly give educators what they're most interested in—the student growth of individual students, but rather an estimate of the value a teacher added to the students that he/she instructed.

However, there are issues and concerns with incorporating measures of growth and achievement into an evaluation, even if those scores are attributed from individual students through roster verification. One of the main challenges of utilizing a growth, or value-added

score within an evaluation is that students are not randomly assigned (Rothstein, 2009). This challenge leads directly into the next challenge: oftentimes with any of these models, there is often a leap to causal inference (Betebener, 2009), where none can truly be made due to the lack of randomization. However, causal claims are made so as to point accountability to where it supposedly belongs.

Another challenge of both SGPs and value-added models is that complex statistical software is required to run these models for stakeholders. When either of these choices are selected at the state education agency level, oftentimes it is the state education agency that assumes the responsibility for running these types of models; however, there is a decrease in transparency of sorts when results cannot be replicated by stakeholders in the field.

According to Lockwood et. al. (2007), states and districts have increased their reliance on student test scores as part of accountability systems in part due to the requirements of No Child Left Behind (Lockwood, et all, 2007) and therefore have longitudinal data from stronger testing systems data systems, which make determining growth measures possible.

One way that the measures are considered to be fair—and more than showing a bad year is through the use of multiple years of data. In the T-VAAS system or EVAAS system, in place in Tennessee, and North Carolina respectively, at least three years' of student data are captured within the model. Similarly, in states that use student growth percentiles, growth measures are determined based on the current, and at least prior year's score for a student.

As a way to adjust for potential anomalies in data, some approaches take averaging, or multiple years of data into account. In Michigan, when the state's system goes into full swing in school year 2015-2016 with student growth measures weighted as 50% of the

overall evaluation, three consecutive years of growth and achievement data *must* be used if available. If three years of data are not available, any available growth and achievement data are to be used (HB 4267, 2001).

Purposes of Evaluating Educators

It is important to contextualize the purpose and background in the evaluation of teachers. In the 1980s, there was a challenge in the use of teacher evaluation results for both formative and summative purposes (Darling Hammond, 1983). This issue remains the same today, however evaluation has improved over the last 25 years due to the availability of objective data, which is and can be included as one or more component within the teacher evaluation systems (Peterson & Peterson, 2006). When evaluation results of teachers are used for multiple decisions from tenure, to merit-pay, to placement, and retention (Gallagher, 20012), all of which are high-stakes for the teacher (Danielson, 2008), ensuring the validity of the system is a must. In addition to the many high-stakes decisions that might be made from the evaluation results as mentioned above, there is a growth component or purpose in the results of the evaluation as well—to determine targeted professional development for teachers in areas where they may demonstrate weakness or needed growth. As Danielson (2008) noted, teaching is difficult and never perfect, making the need for an evaluation system that promotes professional growth necessary to make change in teachers' practices.

Many systems, such as the one currently implemented in West Virginia, were developed to also promote teacher growth. Although the overall score is a summative score of effectiveness, the detail within the summative evaluation can help inform the teacher of the improvements he/she may need to make. To some, this may seem to conflict with a system that is supposed to assure teacher quality, but focusing on continued improvement

and growth of teachers will help with increasing teacher quality over time (West Virginia Educator Evaluation System for Teachers guidance documentation, (2012).

The Measures for Effective Teaching (MET) project, a three-year project from 2010-2013 sought to understand more about educator effectiveness. With the participation of over 3000 teachers across the country, and many participating on the Advisory Committee, these teachers reported that traditional evaluations, meaning a satisfactory/unsatisfactory model, could not provide usable information to guide improvements in teaching. In fact, these teachers reported that traditional evaluations were perfunctory and disconnected from their work of teaching and learning (2013). The traditional, satisfactory/unsatisfactory model of teacher evaluation simply did not do enough to differentiate teacher performance. This was true in West Virginia prior to the implementation of a statewide educator evaluation system based on multiple measures; a satisfactory/unsatisfactory model was in place (J. D'Brot, personal communications, September 22, 2013).

In Washington, D.C. schools, the IMPACT system and the District of Columbia Public Schools' Effectiveness Assessment System for School-Based Personnel, was developed to help administration identify and reward those teachers who were advancing their students. It was pointed out that nearly all teachers were receiving "satisfactory" ratings despite the low levels of performance by students and schools overall. The IMPACT system forced the conversation around student learning, growth, and performance. Teachers that were able to demonstrate gains for their students were then promoted, offered higher pay to teach in more difficult schools to produce results, or made to be peer coaches for other teachers who didn't perform as strongly (IMPACT, 2013).

Although the *Educator Evaluation Instrument* in West Virginia does not go as far as to measure effectiveness in order to determine pay increases or teacher placement, it is meant to provide feedback to teacher with greater granularity than simply satisfactory or unsatisfactory. Additionally, the evaluations are meant to provide teachers with targeted feedback across the six professional teaching standards, which make the evaluations more meaning full as they reflect other valuable aspects of teaching than just students' test scores, which are often included in new teacher evaluation systems (Nye, 2004).

In summary, ensuring that the instrument used to quantify teachers' effectiveness is valid based on their many practices is critical for all stakeholders with a vested interest in the education of students. Hierarchical confirmatory factor analysis will statistically determine the construct validity of the *Educator Evaluation Instrument* paired with the examination of the constructs that compose the instrument. From policymakers to principals, understanding the validity of the instrument by which teachers are being evaluated will assist in making informed decisions that impact educators and students. A valid instrument will allow those stakeholders to trust and rely upon the results rendered from the *Educator Evaluation Instrument* in order to support educators in improving instruction, attending specialized professional development, and ultimately increasing student achievement.

Chapter 3

Methodology

Sample

The sample in this study consists of the educator evaluation results for 3,848 teachers in the state of West Virginia. These teachers were from over 100 schools across the state at the elementary, middle, and high school levels that participated in the demonstration year of the educator evaluation system rollout, which included the utilization of the *Educator Evaluation Instrument*.

Procedures

The data for this study will be from the school year 2012-2013 educator evaluation results collected by the West Virginia Department of Education from the *Educator Evaluation Instrument*. To obtain these data, a Research Proposal Application will be submitted to the Research Review Committee at the West Virginia Department of Education. The application requires the applicant to explain the purpose of the research study and the data being requested. The data requested as part of the Research Proposal Application for a de-identified teacher-level file that contains the following data elements[1]:

- County Code
- County Name
- School Code
- School Name
- Progression Level (Advanced, Intermediate, Initial)
- Overall Rating (Distinguished, Accomplished, Emerging, Unsatisfactory)
- Overall Rating Calculated Value (0-100)
- Standard 1 Rating (Distinguished, Accomplished, Emerging, Unsatisfactory)
- Indicator 1.1 Rating (Distinguished, Accomplished, Emerging, Unsatisfactory)
- Indicator 1.2 Rating (Distinguished, Accomplished, Emerging, Unsatisfactory)
- Indicator 1.3 Rating (Distinguished, Accomplished, Emerging, Unsatisfactory)
- Standard 2 Rating (Distinguished, Accomplished, Emerging, Unsatisfactory)

[1] Definitions for the Indicators in this list of data elements can be found in the Appendix.

- Indicator 2.1 Rating (Distinguished, Accomplished, Emerging, Unsatisfactory)
- Indicator 2.2 Rating (Distinguished, Accomplished, Emerging, Unsatisfactory)
- Indicator 2.3 Rating (Distinguished, Accomplished, Emerging, Unsatisfactory)
- Standard 3 Rating (Distinguished, Accomplished, Emerging, Unsatisfactory)
- Indicator 3.1 Rating (Distinguished, Accomplished, Emerging, Unsatisfactory)
- Indicator 3.2 Rating (Distinguished, Accomplished, Emerging, Unsatisfactory)
- Indicator 3.3 Rating (Distinguished, Accomplished, Emerging, Unsatisfactory)
- Standard 4 Rating (Distinguished, Accomplished, Emerging, Unsatisfactory)
- Indicator 4.1 Rating (Distinguished, Accomplished, Emerging, Unsatisfactory)
- Indicator 4.2 Rating (Distinguished, Accomplished, Emerging, Unsatisfactory)
- Standard 5 Rating (Distinguished, Accomplished, Emerging, Unsatisfactory)
- Indicator 5.1 Rating (Distinguished, Accomplished, Emerging, Unsatisfactory)
- Indicator 5.2 Rating (Distinguished, Accomplished, Emerging, Unsatisfactory)
- Indicator 5.3 Rating (Distinguished, Accomplished, Emerging, Unsatisfactory)
- Standard 6 Rating (Distinguished, Accomplished, Emerging, Unsatisfactory)
- Indicator 6.1 – Student Growth Goal 1 Rating (Distinguished, Accomplished, Emerging, Unsatisfactory)
- Indicator 6.2 – Student Growth Goal 2 Rating (Distinguished, Accomplished, Emerging, Unsatisfactory)
- Indicator 6.3 – Reading Growth Rating (Distinguished, Accomplished, Emerging, Unsatisfactory)
- Indicator 6.4 – Mathematics Growth Rating (Distinguished, Accomplished, Emerging, Unsatisfactory)

Although no individual teachers will be identified from this data file, the file will be kept securely on the researcher's computer, which requires a password for sign in. The county and school names are being requested for descriptive analyses to describe the number of participating elementary, middle, and high schools during the demonstration year.

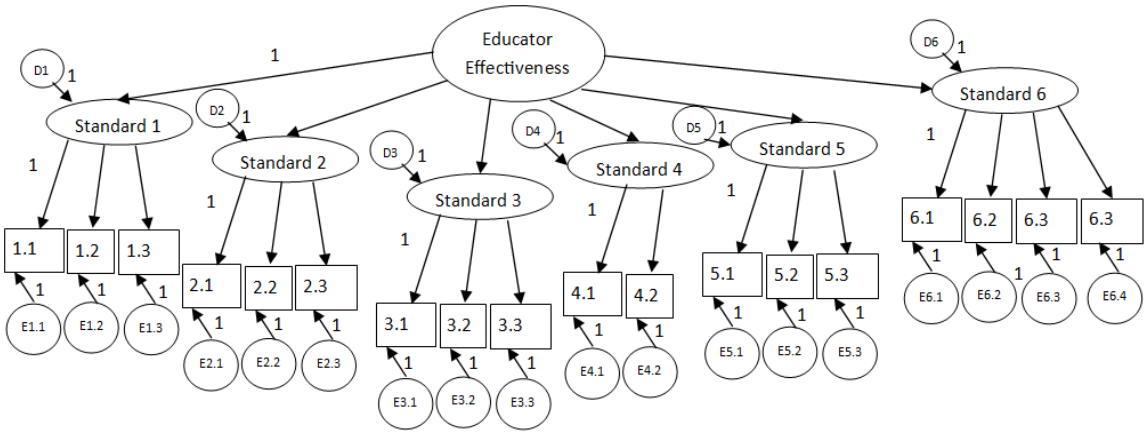
Data Analysis

All data analyses will be conducted using SPSS ver. 22 and SPSS Amos ver. 22. Upon receipt of the data, the file will first be examined to determine if there are any missing data. If data are missing within a de-identified, teacher-level record, the record will be flagged so as to not be included in data analyses. Additionally, the data will be reviewed for any information that may not fit the expected outcome. If there are any questions or concerns

about the output where an apparent anomaly may exist, there is a contact at the West Virginia Department of Education, through the Research Review Committee, who may be reached for additional information. Upon ensuring that the data are clean, descriptive statistics on the number excluded records and the scores (e.g. measures of location and variability) will be presented along with bar charts to show frequencies of scores for the combined results, as well as by teaching progression (i.e. Initial, Intermediate, and Advanced), and school type (i.e. elementary, middle, or high school). Additionally, a test of normality will be run to determine whether the data are normal. Whether the distribution is normal will determine the subsequent analyses described.

The confirmatory factor analysis and hierarchical confirmatory factor analysis models will be designed through SPSS AMOS version 22. Indicators 1.1-1.3 measure Standard 1, 2.1-2.3 measure Standard 2, 3.1-3.3 measure Standard 3, 4.1-4.2 measure Standard 4, 5.1-5.3 measure Standard 5, and 6.1-6.4 measure Standard 6. Each standard has at least two direct causes. All of these standards, or first-order factors, indirectly measure the second-order factor, which is *g*, Educator Effectiveness. Before running the HCFA, the CFA with indicators and standards will be run. The correlations between the standards will be examined to see if any are highly correlated, which means that there is little distinction between the standards. The model fit of the standard CFA will also be examined. For the HCFA, the other presumed cause of the first-order factors is a disturbance, which represents factor variance not explained by *g*, Educator Effectiveness. The disturbances and *g* are exogenous, but the first-order factors are endogenous.

Figure 2. Hierarchical Confirmatory Factor Analysis (HCFA) Model



For this study, a standard multi-group confirmatory factor analysis model is utilized because each indicator loads on only one factor. For this reason, the variance of *g* is fixed to 1.0 to standardize it, which leaves all six direct effects of *g* on the first-order factors as free parameters.

Hierarchical confirmatory factor analysis (HCFA) on multiple groups (teachers in each of the three progressions) will be run to determine if each of the factors is independent from one another. For a model to be identified in HCFA, there must be at least three-first order factors, and each first-order factor should have at least two indicators. In the case of the *Educator Evaluation Instrument*, the model is identified according to this rule (Klein, 2011).

HCFA will confirm the factorial structure of the instrument for a target population (Wang, 2012, p. 30) and will show how much influence the factor has on the indicator. For indicators loading on one factor, as is the case in this study, the standardized factor loadings are correlations between indicator and the factor. The factor loadings for the three groups of teachers will be examined to see if they are equal or nearly equal. In addition the following results will be examined to understand if the model fits the data: Chi square (CMIN), Comparative Fit Index(CFI), Root Mean Square Error of Approximation (RMSEA), and

confidence interval (CI). Where CMIN = 0, there is a perfect fit. Chi-square should be small, and it should not be significant; it is a badness of fit statistic in that, the higher the value, the worse the fit. However, chi square increases with larger samples and non-normal distributed data. If data were determined to be non-normal, a bootstrapping technique will be utilized through Amos for 2000 bootstrap samples. From those iterations, the mean chi-square from the bootstrapped samples will be analyzed. CMIN/DF should be less than 2. If RMSEA is < .06, there is a good fit; if it is < .09, it is adequate fit. RMSEA is more of a badness of fit test where the higher the number, the worse the fit . Additionally, RMSEA should fall between the confidence interval (CI). CFI should be 1.0 for a perfect fit. If the model doesn't fit, it challenges the theories that are the basis for the instrument. With HCFA, the researcher will determine whether the model supports or fails to support the theory behind it.

If the CFA and HCFA do not have good model fit, an exploratory factor analysis will be run for principal component analysis. With the results of the exploratory factor analysis, HCFA will be run based on the new model to determine if there is good fit.

Chapter 4

Results

All computations were obtained via SPSS ver. 22. Descriptive statistics were computed on variables in the data file to determine the number of cases in each rating category, as shown in Table 2 below.

Table 2

Effectiveness Rating Frequency

	Frequency	Percent	Valid Percent	Cumulative Percent
Unsatisfactory	43	1.1	1.1	1.1
Emerging	683	17.7	17.7	18.9
Accomplished	2964	77.0	77.0	95.9
Distinguished	158	4.1	4.1	100.0
Total	3848	100.0	100.0	

Next, the descriptives were run to determine the number of valid cases by progression, which will be the basis for the groupings utilized in the CFA and HCFA where Advanced refers to teachers with 6 or more years of experience, Intermediate refers to teachers with 4-5 years of experience, and Initial refers to teachers with 1-3 years of experience. The frequencies are shown in Table 3.

Table 3

Progression Frequency

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	Advanced	2574	66.9	66.9	66.9
	Intermediate	423	11.0	11.0	99.8
	Initial	844	21.9	21.9	88.8
	N/A	7	.2	.2	100.0
	Total	3848	100.0	100.0	

Nearly 67% of West Virginia's teachers in the demonstration year have more than six years of teaching experience. Seven cases with the progression of Not Applicable (N/A) were then deleted by hand as N/A was not an option for progression type rendering the results associated with these cases as flawed.

The cross tabulation in Table 4 shows the number of cases by progression by overall effectiveness. The highest number of cases across all three progressions was in the Accomplished category as shown below.

Table 4

Cases by Progression and Effectiveness Rating

		Effectiveness Rating				
		Distinguished	Accomplished	Emerging	Unsatis-factory	Total
Progression	Advanced	125	2124	295	30	2574
	Intermediate	17	306	95	5	423
	Initial	15	532	289	8	844
Total		157	2962	679	43	3841

Next, a test for normality was run. The Kolmogorov-Smirnov test of normality was examined because the sample size is greater than 2000. The results indicate that for each

component of the *Educator Evaluation Instrument*, the data are not normal as indicated by the results in Table 5. Because the data are not normal, a boostrap methodology is used in the CFA and HCFA processes. The chi-square result cannot be interpreted directly, because it will be speciously large due to non normality.

Table 5

Tests of Normality

	Kolmogorov-Smirnov		
	Statistic	df	Sig.
Rating1	.404	3841	.000
Rating2	.401	3841	.000
Rating3	.413	3841	.000
Rating4	.395	3841	.000
Rating5	.407	3841	.000
Rating6	.408	3841	.000
std1_1	.353	3841	.000
std1_2	.400	3841	.000
std1_3	.390	3841	.000
std2_1	.377	3841	.000
std2_2	.398	3841	.000
std2_3	.395	3841	.000
std3_1	.395	3841	.000
std3_2	.394	3841	.000
std3_3	.392	3841	.000
std4_1	.393	3841	.000
std4_2	.385	3841	.000
std5_1	.393	3841	.000
std5_2	.380	3841	.000
std5_3	.400	3841	.000
StudentGrowthGoal175	.390	3841	.000
StudentGrowthGoal275	.391	3841	.000
MathGRating	.211	3841	.000
RLAGRating	.221	3841	.000
OverallRating	.439	3841	.000

Notes: Lilliefors significance correction was applied. Rating*x* refers to the overall rating for the *x* teaching standard (1-6). Std*x_y* refers to the indicator associated with the standard where *x* = the teaching standard (1-6) and *y* = the indicator associated with the standard. StudentGrowthGoal175 and StudentGrowthGoal275 refer to the two scores associated with student growth goals that teachers established at the beginning of the year and the degree to which the goals were met. MathGRating and RLAGRating refer to the school-wide growth scores for mathematics and reading respectively. OverallRating refers to the overall rating received.

The raw data were then loaded into Amos ver. 22, and three groups were established based on the Progression – Advanced, Intermediate, and Initial. For each of these groups, the number of cases were as follows after listwise deletion was implemented for missing variables as shown in Table 6. Thirty-nine cases were deleted due to missing variables.

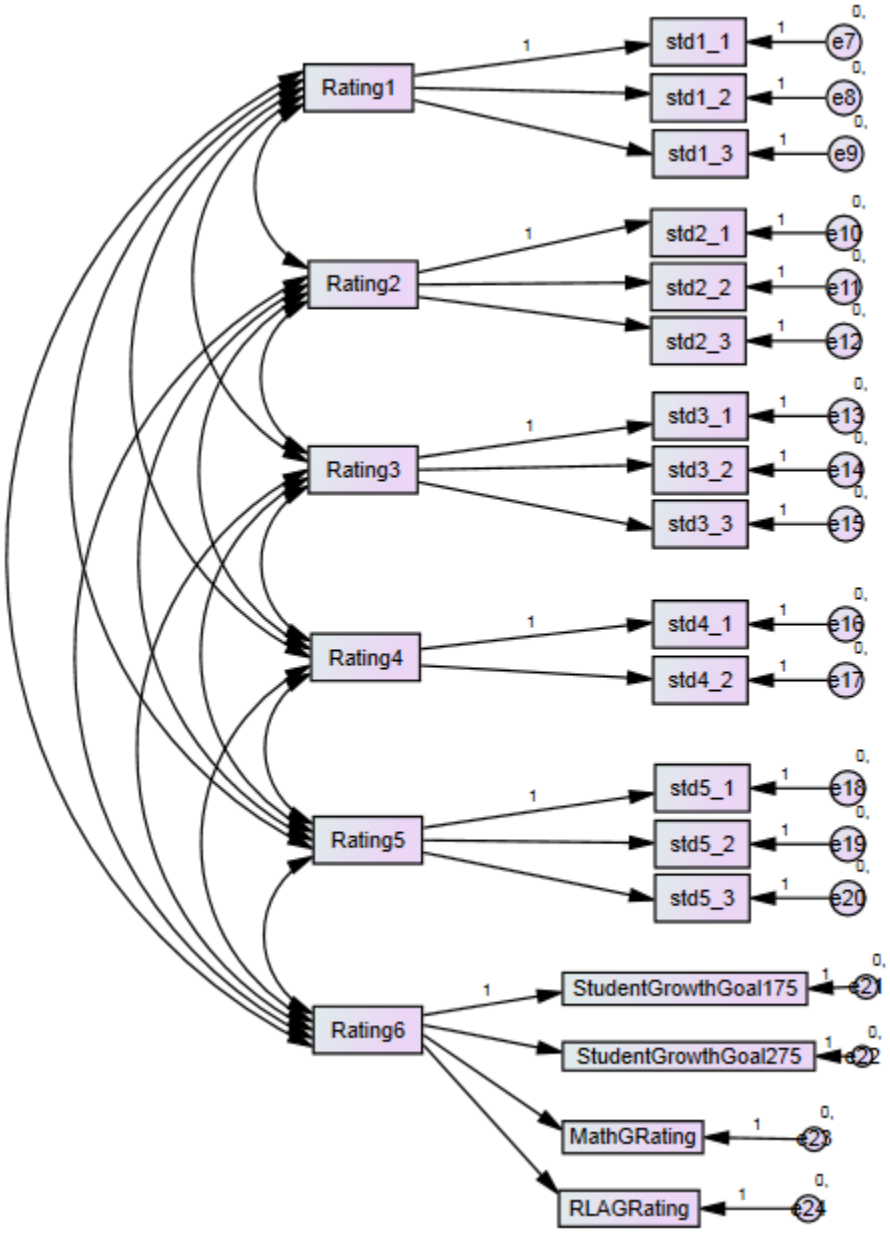
Table 6

Cases by Progression after Listwise Deletion

Progression	Frequency
Advanced	2545
Intermediate	421
Initial	837
Total	3802

For the CFA, the data were loaded and the bootstrap technique was applied for 2000 samples. The model created included observable variables for Ratings 1-6 (that correspond with each of the Teaching Standards 1-5 and the Growth component, 6), along with each of their indicators as shown in Figure 3.

Figure 3: CFA with Ratings and Indicators



The results show the correlations were moderate among standards 1-5, but low between standards 1-5 and 6. The moderate correlations greater than .6 among standards 1-5 indicate that there may be little distinguishability between the first five standards in the *Educator*

Evaluation Instrument in each of the three teaching progressions. As shown in Figures 4, 5, and 6.

Figure 4: CFA Advanced Sample

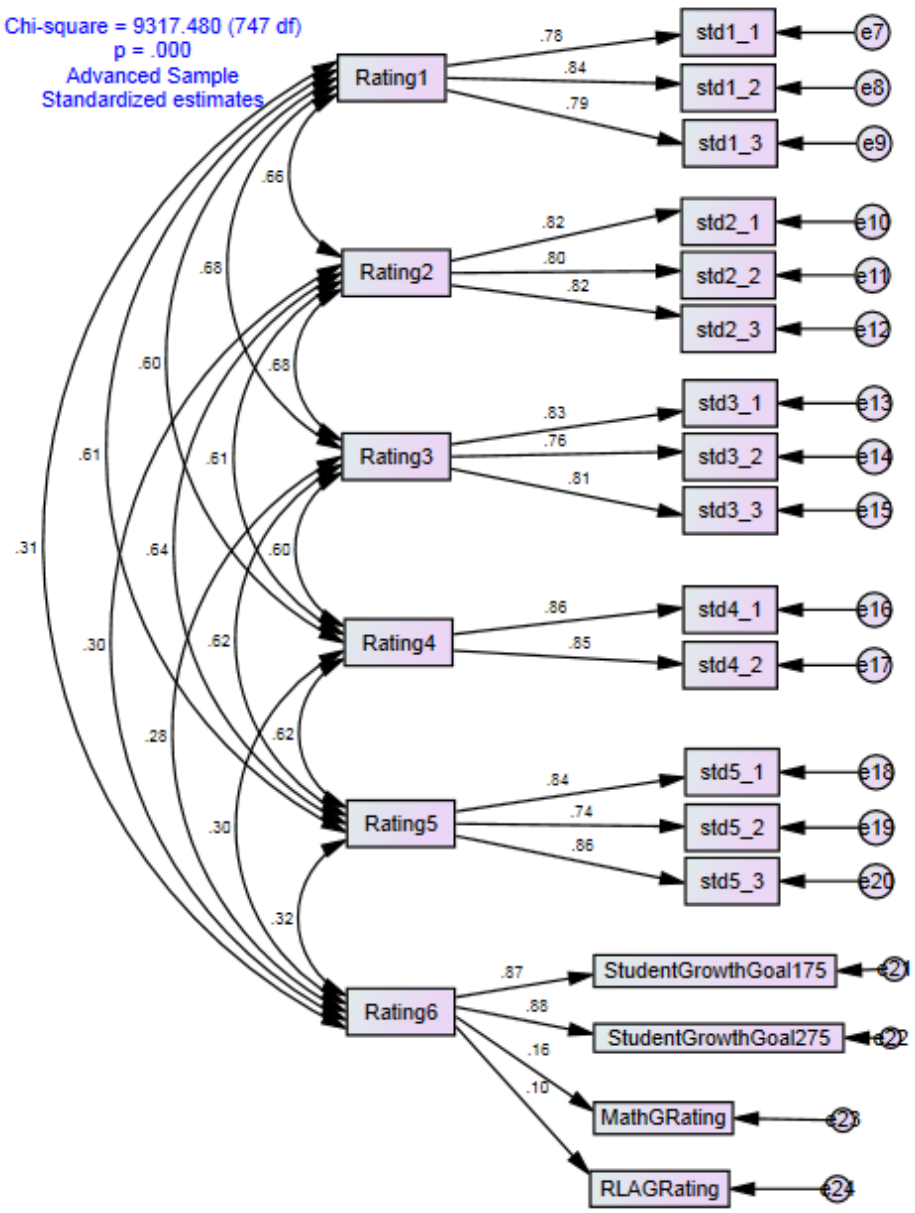


Figure 5: CFA Intermediate Sample

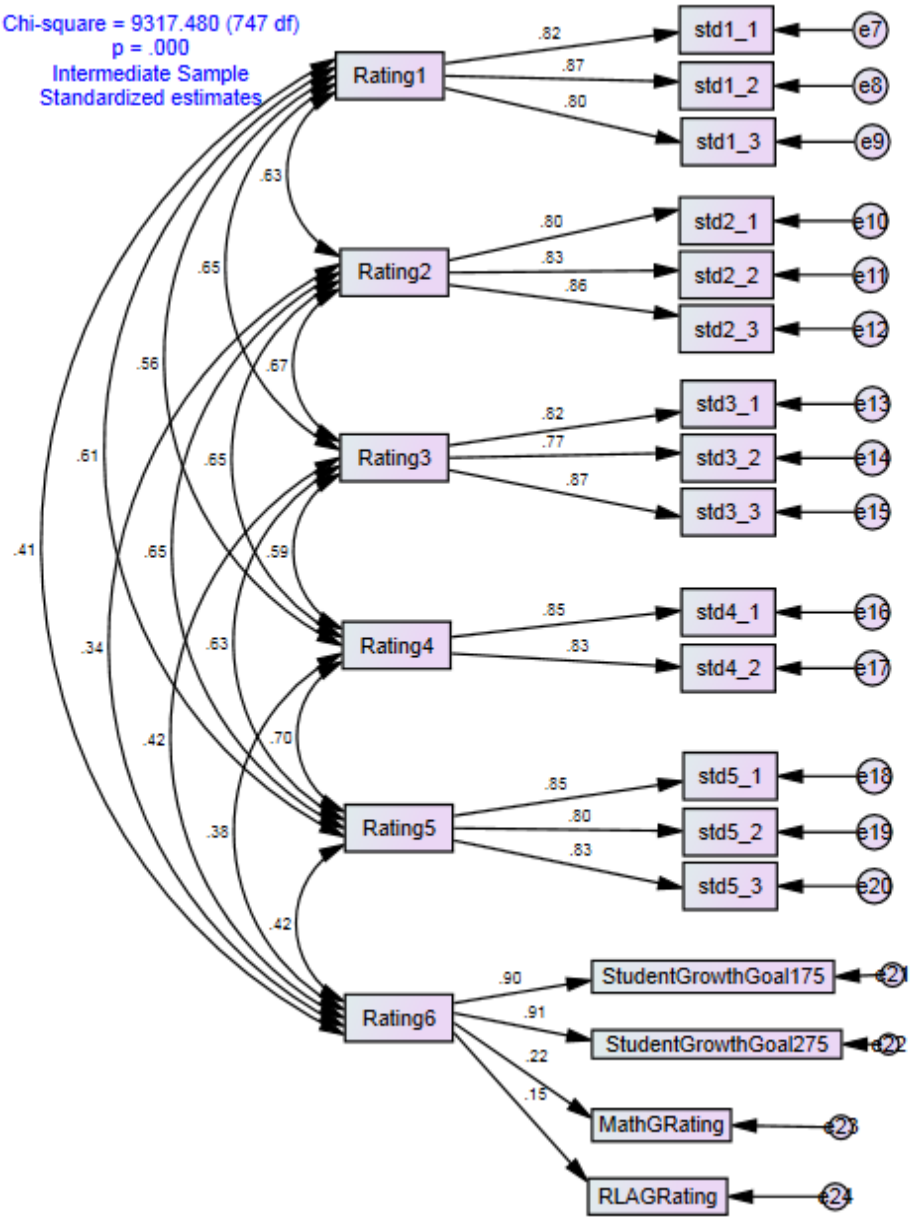
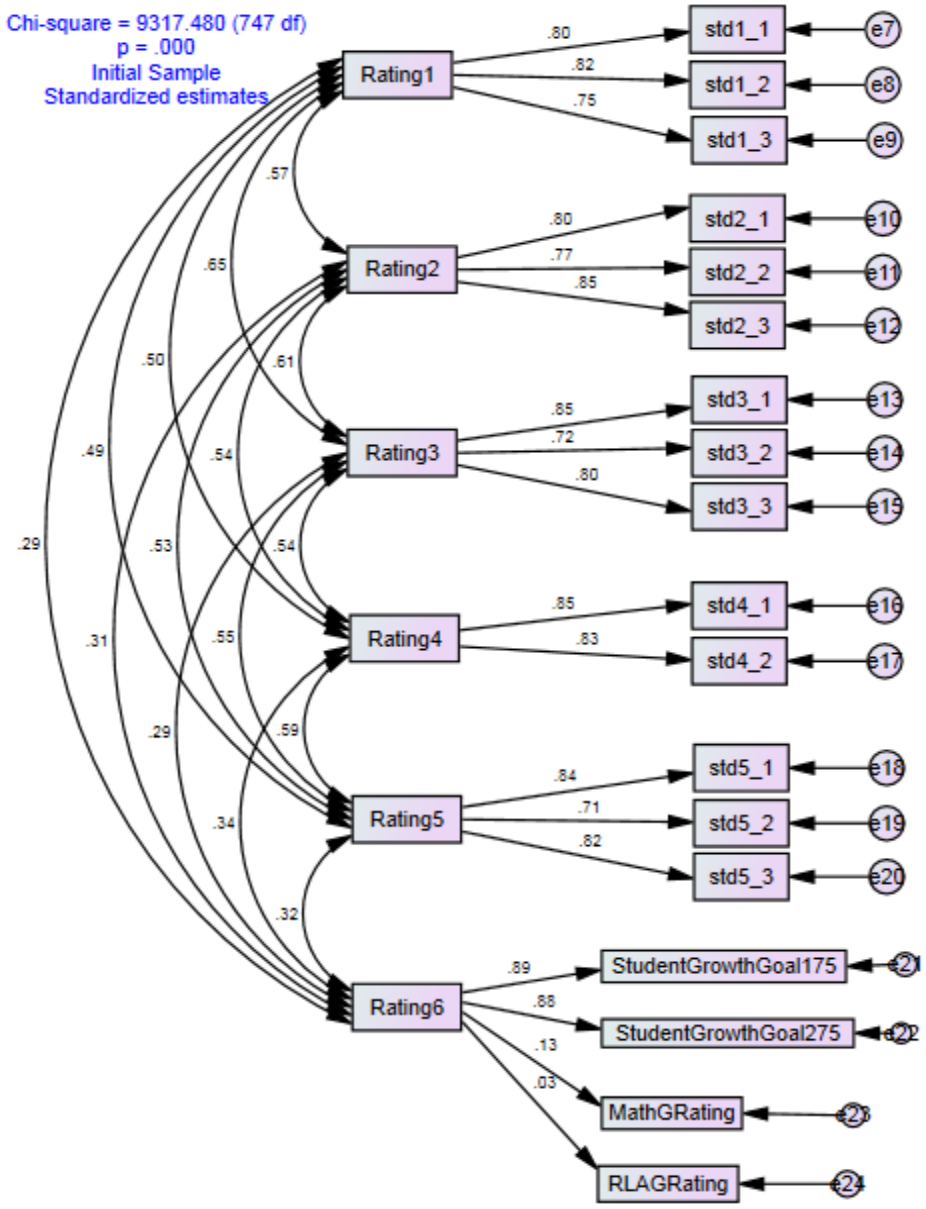


Figure 6: CFA Initial Sample



The model fit for the CFA with the six teaching Standards and the associated indicators was moderate or sometimes permissible, but there were no fit indices that suggest good model fit as shown by the results in Table 7.

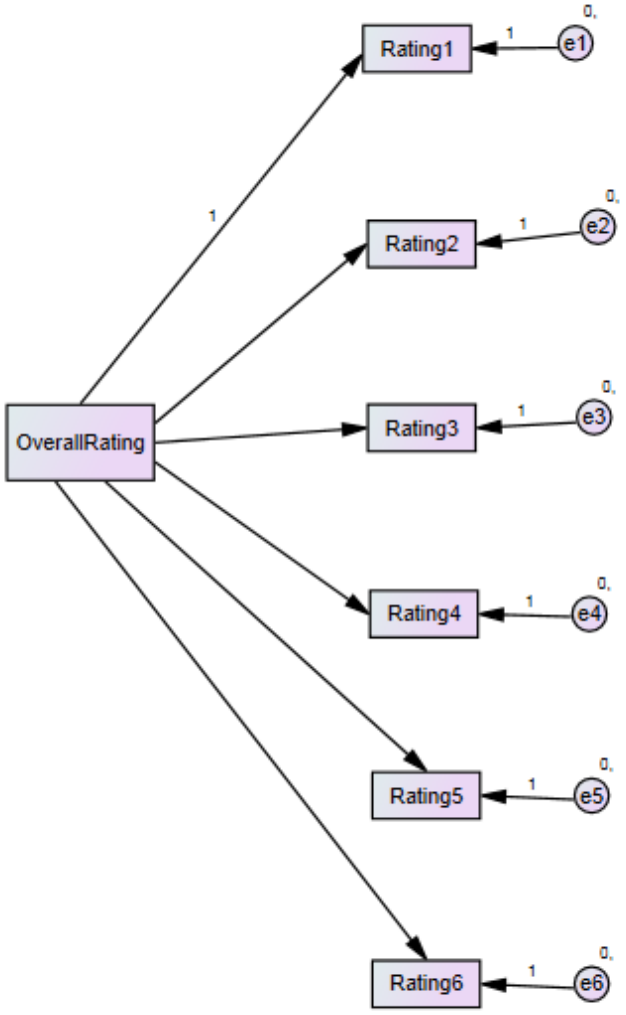
Table 7

Model Fit Summary Results of the CFA Model

Model	Metric	Result	Interpretation
CFA Model	CMIN/DF	12.473	1.0 = perfect fit; should be < 3 for good fit
	P	.000	should be > .05
	CFI	.899	>.80 is sometimes permissible
	RMSEA	.055	.05 - .10 = moderate fit
	PCLOSE	.000	should be > .05

Although there good fit could not be estimated from the model above with the ratings for the six teaching standards and their associated indicators, it was necessary to determine if there was good model fit with the Ratings for the six teaching standards on overall effectiveness as shown in the model in Figure 7.

Figure 7: CFA Model with Overall Effectiveness and Ratings



The model fit results do not confirm good fit as shown by the results in Table 8 and Figures 8, 9, and 10.

Table 8

Model Fit Summary Results for the CFA Model of Overall Effectiveness and Ratings

Model	Metric	Result	Interpretation
CFA Model	CMIN/DF	59.759	1.0 = perfect fit; should be < 3 for good fit
	P	.000	should be > .05
	CFI	.823	>.80 is sometimes permissible
	RMSEA	.124	> .10 = bad fit

Figure 8: CFA Advanced Sample – Overall Effectiveness and Ratings

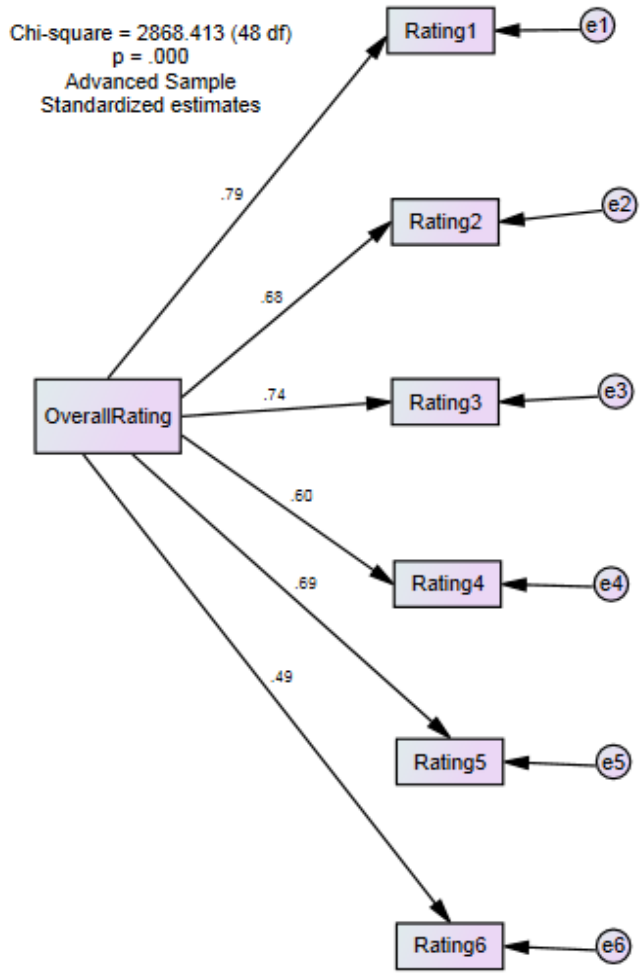


Figure 9: CFA Intermediate Sample – Overall Effectiveness and Ratings

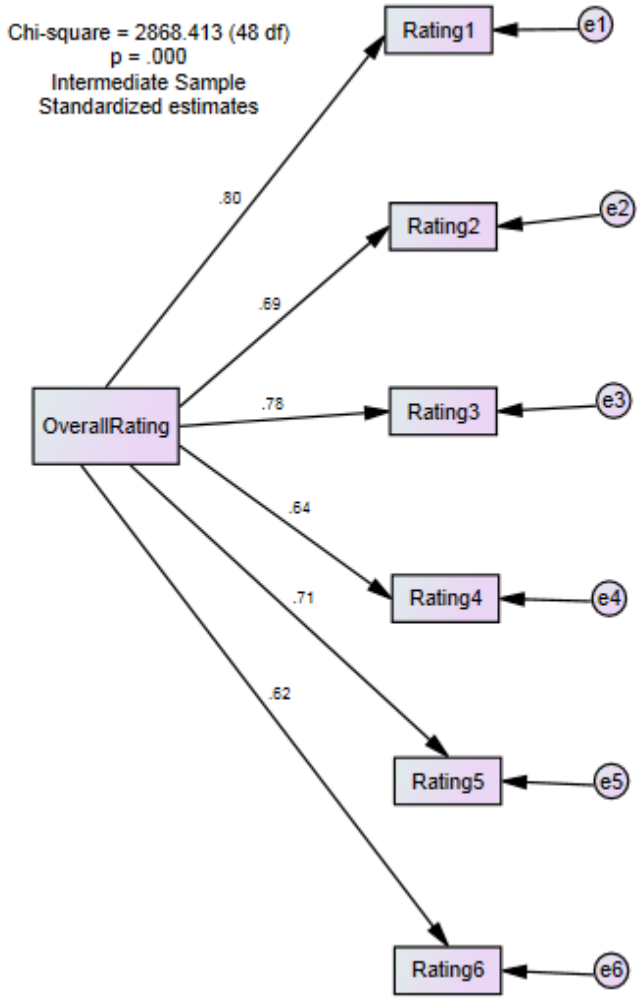
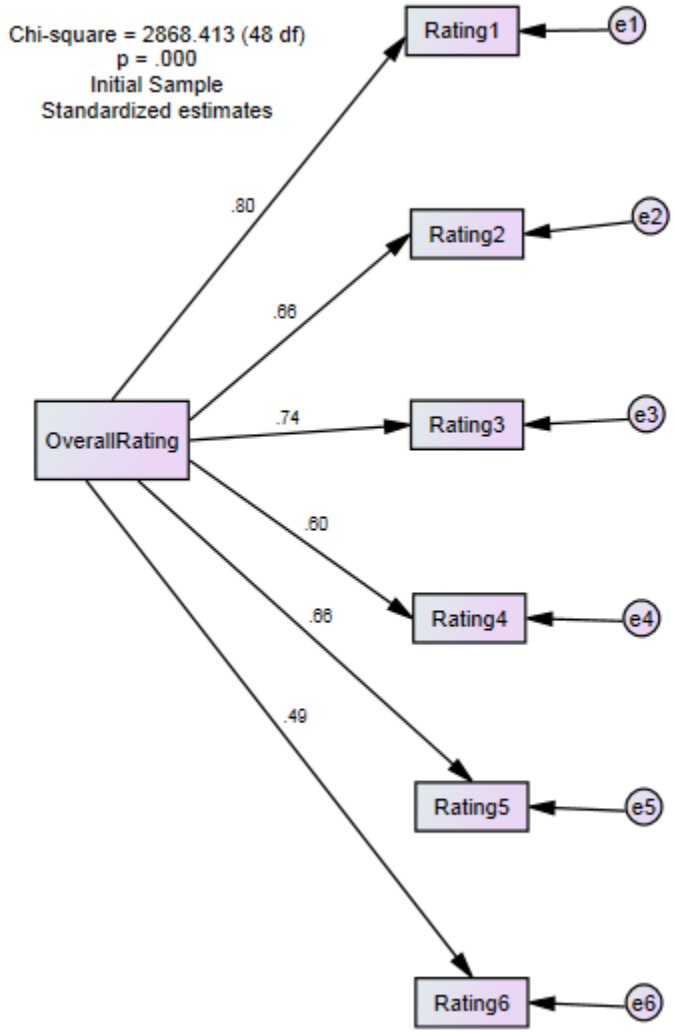
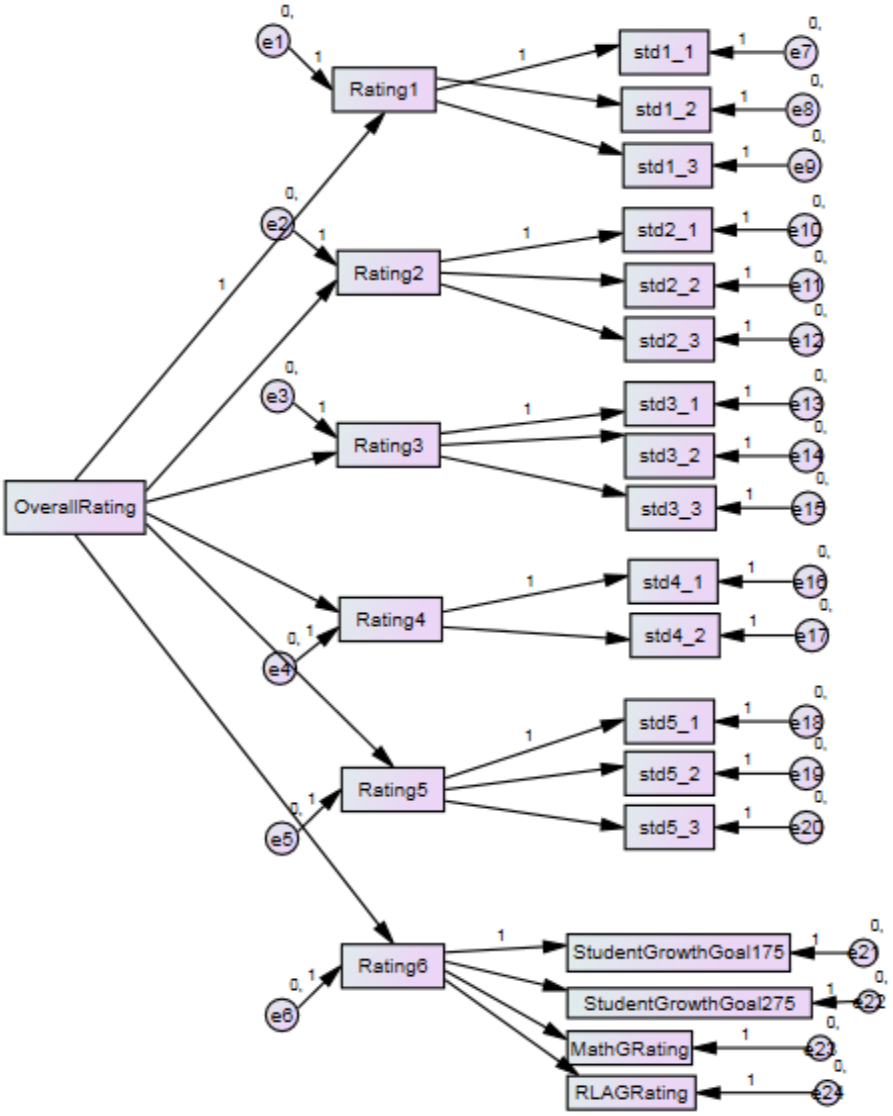


Figure 10: CFA Initial Sample – Overall Effectiveness and Ratings



Although good fit could not be established with the CFAs that utilize the components in a single-factor structure, and good fit is unlikely with the HCFA, the HCFA was developed next with the *g* factor as the overall effectiveness rating, the six ratings for the corresponding standards, and the indicators that correspond with each of the ratings. The model is shown in Figure 11. All are observed variables in this HCFA, and a bootstrap was applied at 2000 samples.

Figure 11: HCFA with Overall Effectiveness, Ratings, and Indicators



The results for the model fit do not indicate good fit as shown in Table 9.

Table 9

Model Fit Summary Results for the HCFA Model

Model	Metric	Result	Interpretation
HCFA Model	CMIN/DF	14.563	1.0 = perfect fit; should be < 3 for good fit
	P	.000	should be > .05
	CFI	.873	>.80 is sometimes permissible
	RMSEA	.059	.05 - .10 = moderate fit
	PCLOSE	.000	should be > .05

Because the model fit could not be described as a good fit on either of the CFA models or the HCFA, an exploratory factor analysis was conducted to determine if the model should be constructed in a different fashion. When the standard ratings and indicators were entered into the exploratory factor analysis, the results showed extreme high and low loadings, but that the mode was reduced to four components as shown in Table 10.

Table 10

Component Matrix

	Component			
	1	2	3	4
Rating1	.826	-.095	.016	.271
Rating2	.835	-.115	-.019	.160
Rating3	.835	-.129	-.020	.206
Rating4	.795	-.041	.028	-.358
Rating5	.817	-.049	.040	-.360
Rating6	.480	.844	-.013	.045
std1_1	.732	-.087	.012	.166
std1_2	.777	-.082	.022	.230
std1_3	.774	-.087	-.008	.238
std2_1	.777	-.116	-.056	.068
std2_2	.746	-.071	.032	.137
std2_3	.799	-.098	-.013	.194
std3_1	.772	-.107	-.024	.178
std3_2	.764	-.111	-.021	.139
std3_3	.792	-.101	-.039	.156
std4_1	.730	-.056	.019	-.333
std4_2	.769	-.034	.006	-.336
std5_1	.747	-.038	.043	-.377
std5_2	.722	-.017	.025	-.269
std5_3	.790	-.056	.013	-.253
StudentGrowthGoal175	.456	.807	-.142	.046
StudentGrowthGoal275	.475	.800	-.131	.036
MathGRating	.085	.156	.828	.037
RLAGRating	.030	.110	.840	.063

Notes: The extraction method utilized was Principal Component Analysis where four components were extracted.

When this occurs, it is necessary to utilize a varimax extraction method for the principal component analysis. With the varimax method, four components were extracted as shown in Table 11.

Table 11

Rotated Component Matrix

	Component			
	1	2	3	4
Rating1	.817	.274	.143	.045
Rating2	.765	.369	.124	.001
Rating3	.796	.333	.114	.000
Rating4	.411	.758	.142	.021
Rating5	.428	.733	.138	.032
Rating6	.177	.163	.935	.115
std1_1	.681	.302	.119	.032
std1_2	.752	.277	.139	.048
std1_3	.757	.268	.138	.018
std2_1	.668	.407	.106	-.044
std2_2	.671	.333	.134	.052
std2_3	.753	.318	.132	.010
std3_1	.725	.316	.117	-.004
std3_2	.698	.343	.108	-.004
std3_3	.727	.344	.108	-.004
std4_1	.379	.700	.112	.010
std4_2	.403	.182	.146	.001
std5_1	.362	.744	.128	-.019
std5_2	.401	.640	.151	.025
std5_3	.473	.671	.135	.011
StudentGrowthGoal175	.173	.145	.911	-.019
StudentGrowthGoal275	.183	.166	.907	-.008
MathGRating	.025	.046	.059	.844
RLAGRating	.009	-.002	.000	.850

Notes: The extraction method utilized was Principal Component Analysis. The rotation method utilized was Varimax with Kaiser Normalization, which converged in five iterations.

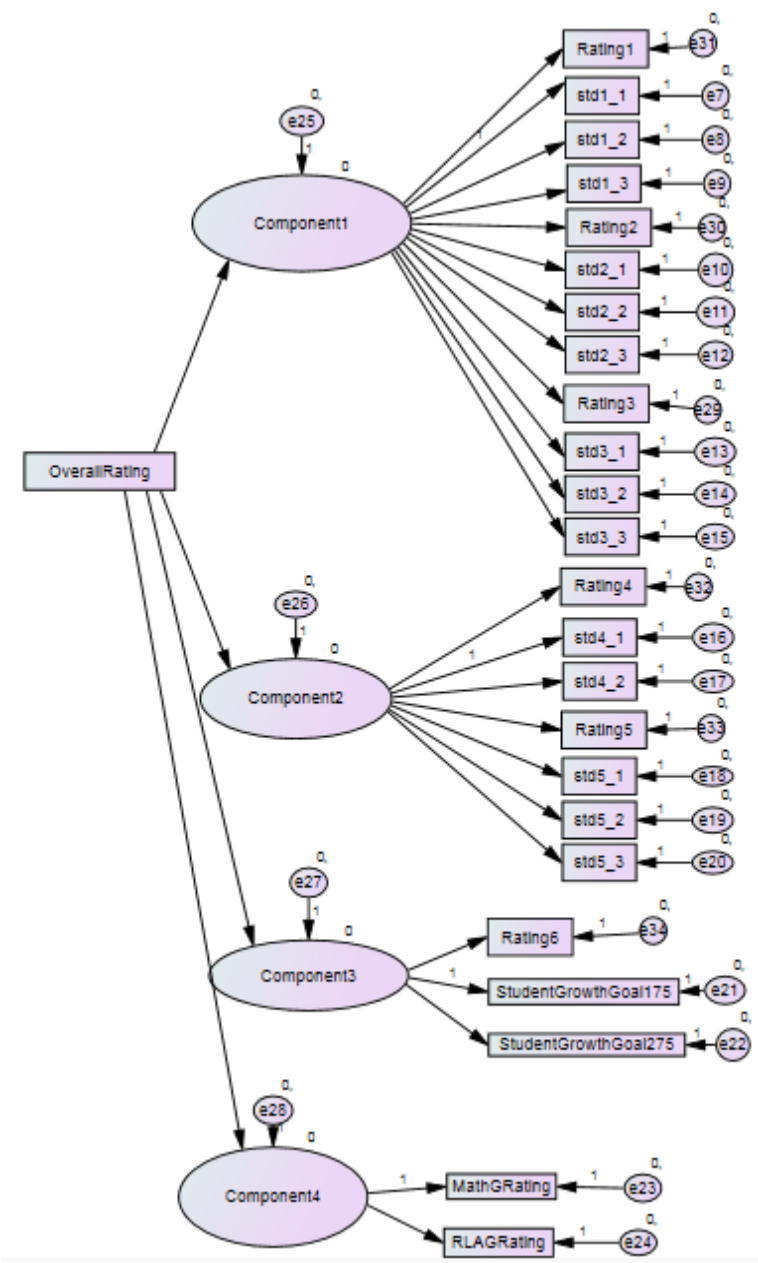
The extraction shows that Component 1 has the highest loadings for Ratings 1-3 and their accompanying indicators. Component 2 includes Ratings 4-5 and their accompanying indicators. Component 3 contains Rating 6 and both student growth goals. Component 4

includes the math school-wide growth score and the reading/language arts school wide growth score.

Given the results of the exploratory factor analysis with the reduction of indicators, a model was built in Amos with the four components as the first-order factors.

This model was constructed based on the results of the varimax extraction as shown in Figure 12.

Figure 12: First HCFA Model Based on Exploratory Results



The model fit results do not show good model fit as depicted in Table 12.

Table 12

Model Fit Summary Results for HCFA

Model	Metric	Result	Interpretation
HCFA Model	CMIN/DF	23.605	1.0 = perfect fit; should be < 3 for good fit
	P	.000	should be > .05
	CFI	.796	>.80 is sometimes permissible
	RMSEA	.077	.05 - .10 = moderate fit
	PCLOSE	.000	should be > .05

The standardized regression weights were also examined to understand to contribution of each component on overall rating of effectiveness. Although the estimates as shown in Table 13 do reflect the theoretical weights developed by the Teacher Effectiveness Task Force of the first-order factors onto the second-order factor, it is always most permissible to get to a simpler model when possible.

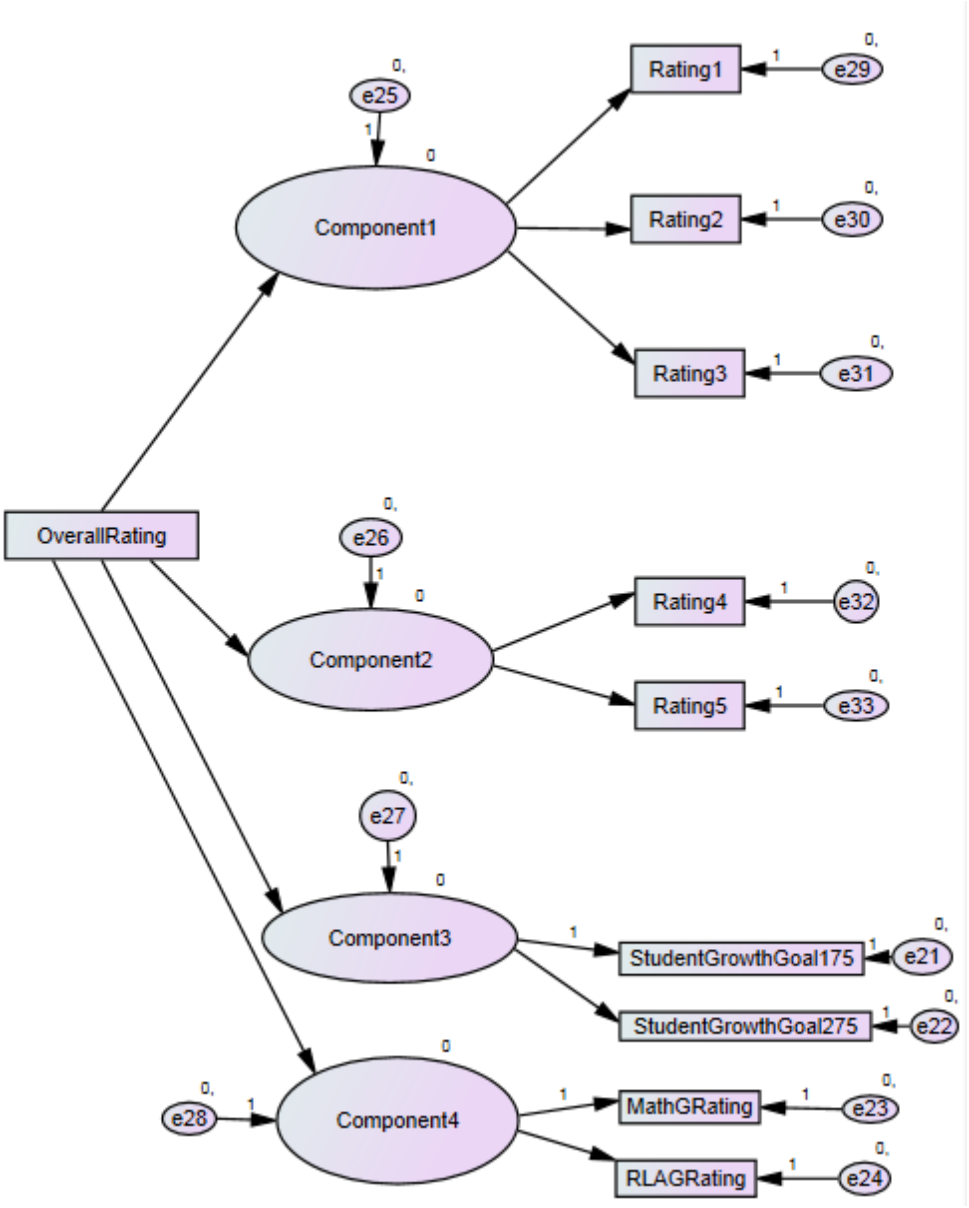
Table 13

Standardized Regression Weights by Progression for HCFA

			Advanced Group Estimate	Intermediate Group Estimate	Initial Group Estimate
Component 1	←	Overall Rating	.817	.852	.841
Component 2	←	Overall Rating	.760	.764	.743
Component 3	←	Overall Rating	.513	.625	.497
Component 4	←	Overall Rating	.066	.261	.120

Next, a model was constructed with the second-order factor, the four first-order factors (new components 1- 4) as shown in Figure 13.

Figure 13: Second HCFA Model based on Exploratory Results

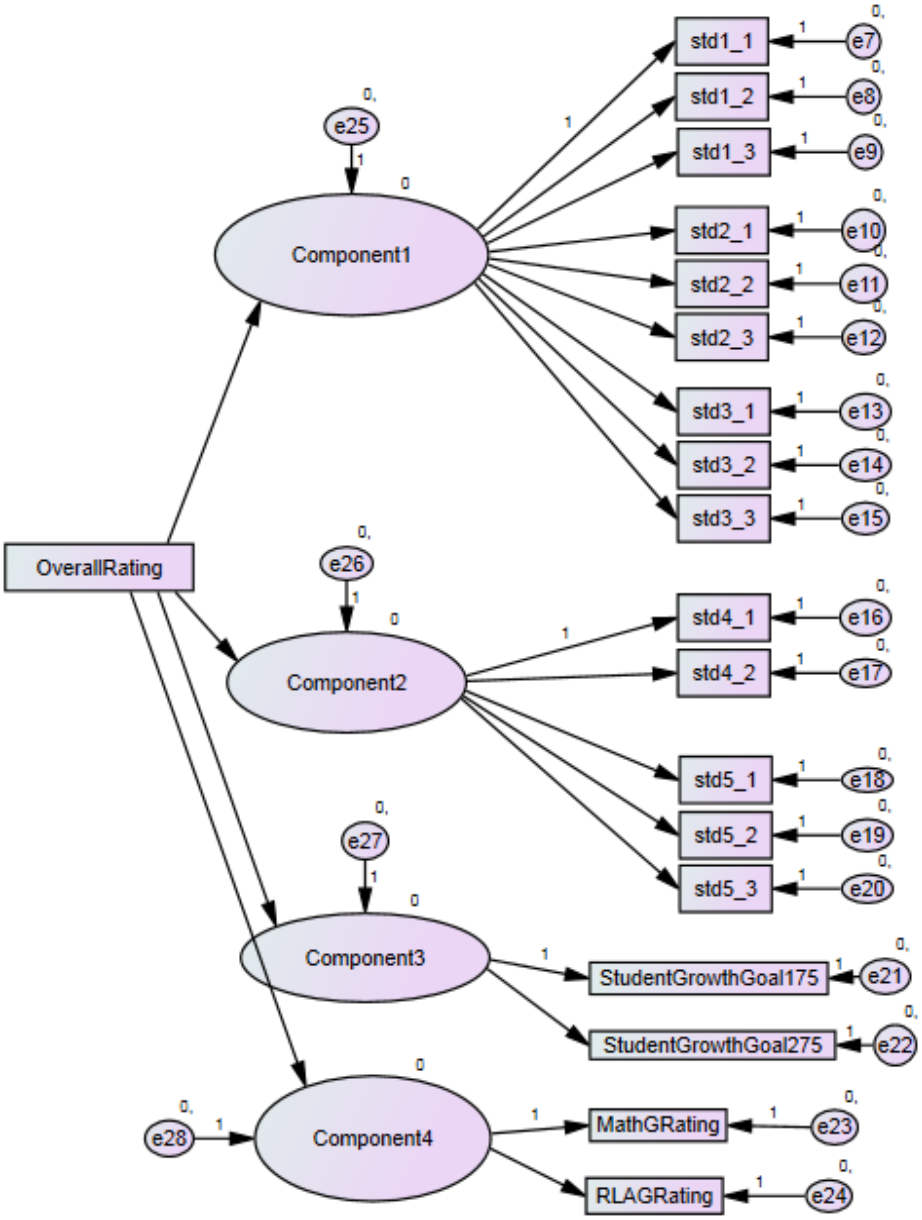


With this model that utilized only the ratings as indicators, the model was unidentified, making this an implausible option for a new model that would be valid for use in determining effectiveness. It requires the addition of at least six additional constraints to determine whether there is good model fit.

For this reason, an additional model was constructed, eliminating the overall ratings for each of the six teaching standards, so that just the indicator values were put into the

model. With the elimination of these ratings' values, the will be less redundancy, but sufficient parameters for an identified model. Therefore, the next model was developed with the same four first-order factors as determined from the principal component analysis. In this model, however, the indicators were place in the model instead of the ratings as shown in Figure 14 below.

Figure 14: Third HFCA Model Based on Exploratory Results



In the new confirmatory model, the standard ratings were eliminated, and only indicators and the overall effectiveness were included as observed variables. The new components were included as latent variables in the model.

Upon conducting the HCFA on the new, four-component model, the model fit results were as shown in Table 14, with several of the model fit specifications indicating good fit.

Table 14

Model Fit Summary Results for the Four-Component HCFA

Model	Metric	Result	Interpretation
HCFA Model	CMIN/DF	7.053	1.0 = perfect fit; should be < 3 for good fit
	P	.000	should be > .05
	CFI	.939	>.90 traditional fit
	RMSEA	.040	<.05 = good fit
	PCLOSE	1.000	should be > .05

It is important to note that the greater than normal CMIN/DF was to be expected because the data were not normal. Even with the bootstrapping technique applied, the chi-square value will be inflated as it is here.

For each of the categories of teachers—advanced, advanced, intermediate, and initial—there were some similarities and differences in how the first-order factors loaded onto the second-order factor of overall effectiveness. As shown below in Figures 15, 16, and 17, the results showing the greatest similarity were the Advanced progression of teachers and the initial progression.

Figure 15: HCFA Advanced Sample - Components with Indicators

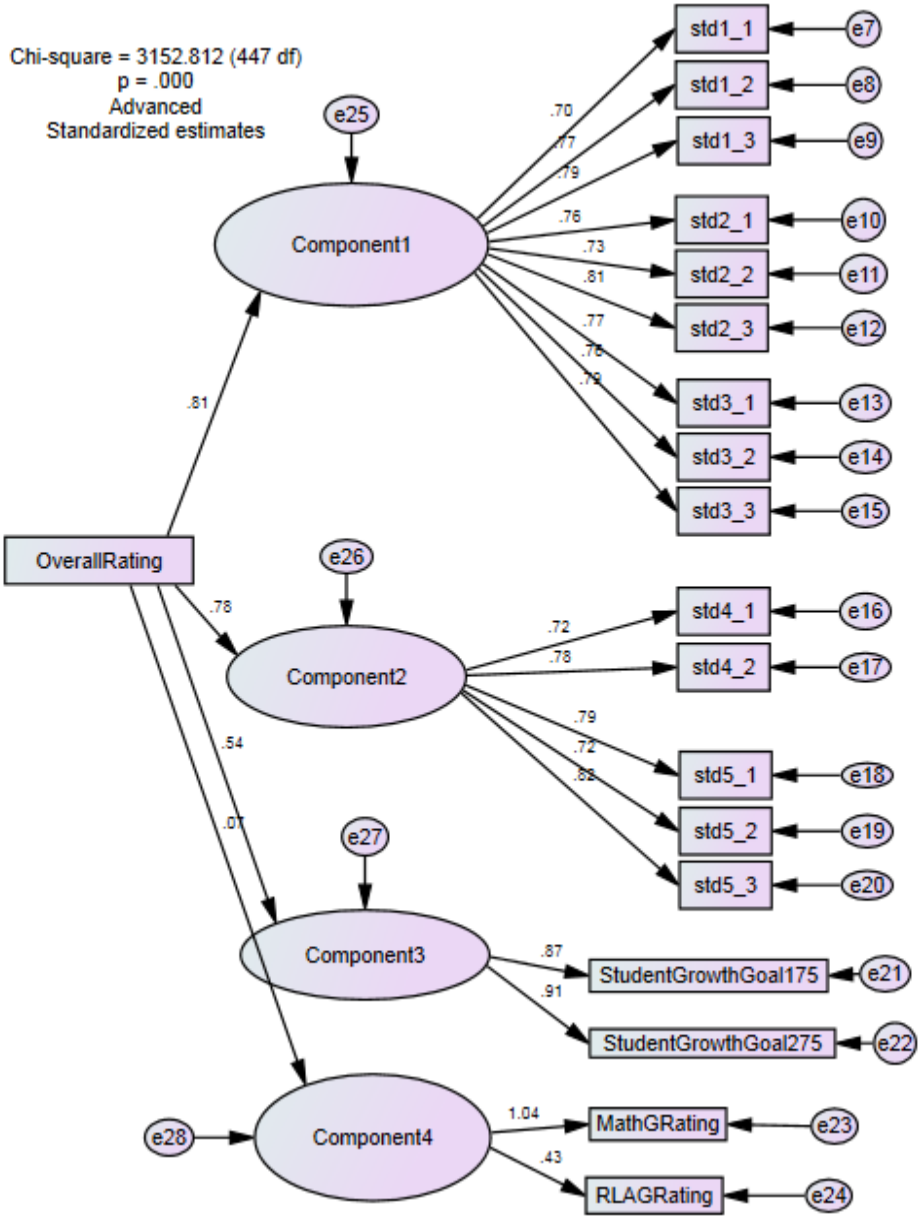


Figure 16: HCFA Intermediate Sample - Components with Indicators

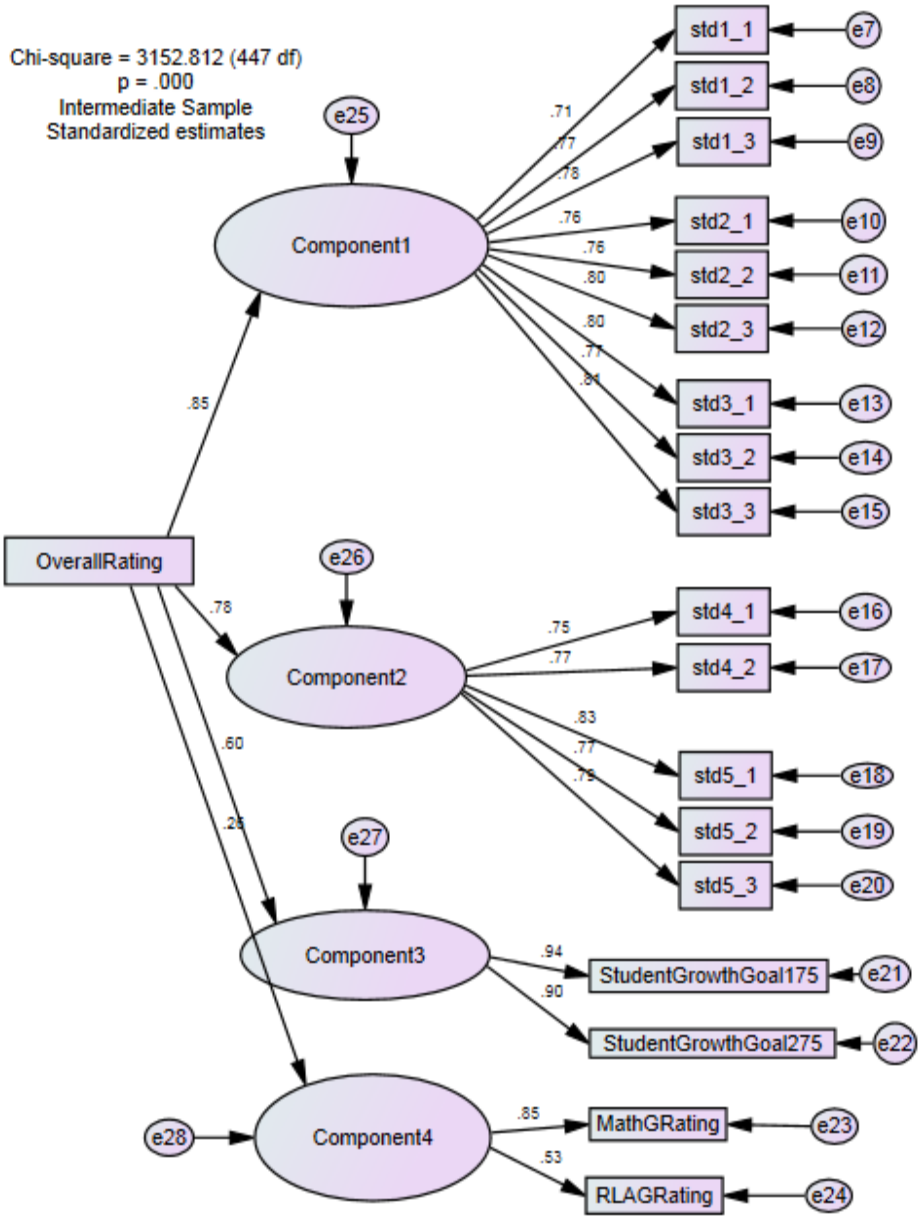
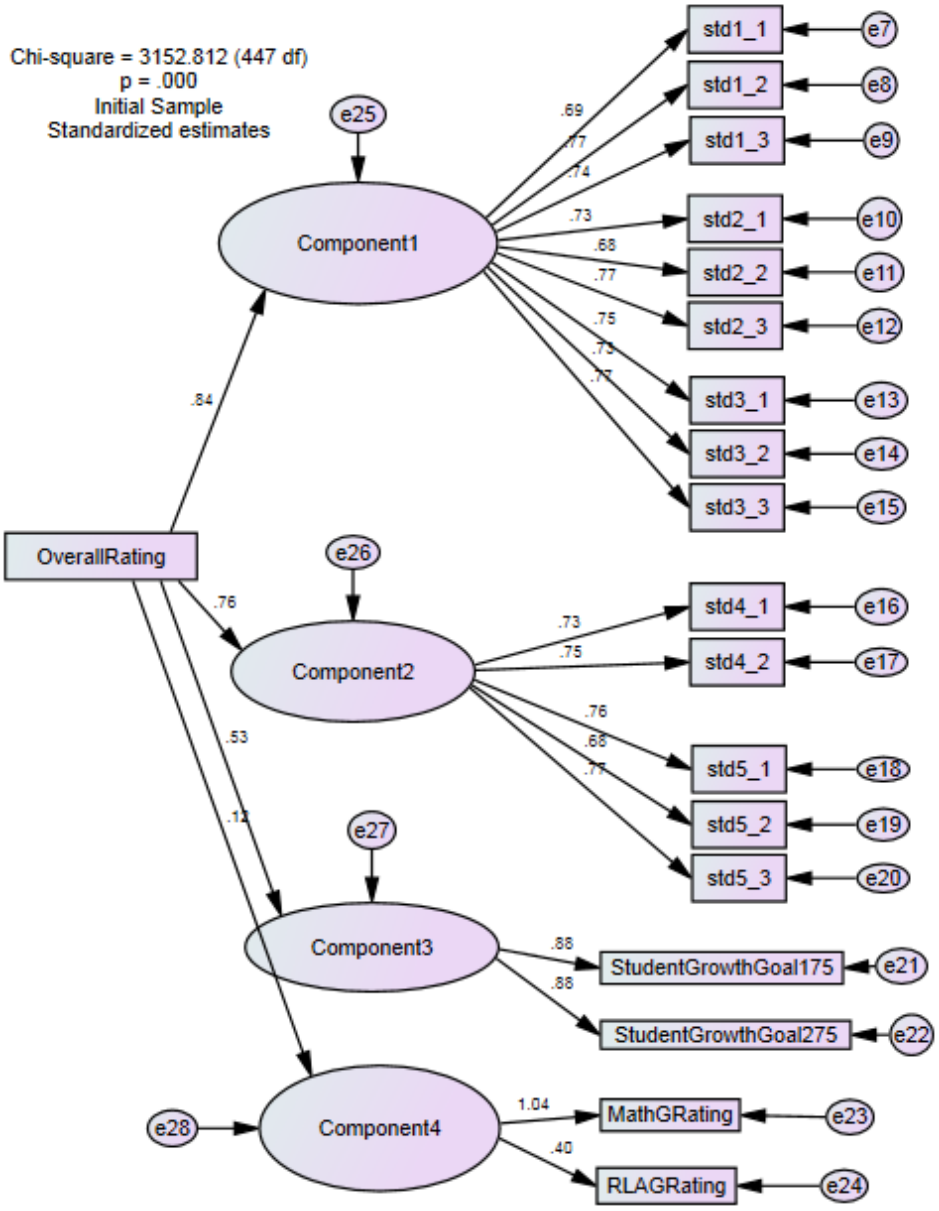


Figure 17: HCFA Initial Sample - Components with Indicators



The standardized regression weights for the new HCFA model for each of the progressions is as shown in Table 15.

Table 15

Standardized Regression Weights by Progression for Four-Component HCFA

			Advanced Group Estimate	Intermediate Group Estimate	Initial Group Estimate
Component 1	←	Overall Rating	.814	.851	.838
Component 2	←	Overall Rating	.778	.778	.761
Component 3	←	Overall Rating	.538	.604	.529
Component 4	←	Overall Rating	.066	.261	.120

Chapter 5

Conclusions & Recommendations

The results of the CFAs and HCFAs for the *Educator Evaluation Instrument* indicate it does not meet the requirements to establish construct validity through good model fit indices. However, an exploratory factor analysis was conducted, and a new model was developed with good model fit. It was subsequently examined via hierarchical confirmatory factor analysis (HCFA) through Amos software. The results indicated the need for the indicators from the current model to roll up to four first-order components, rather than the original six. The grouping of these indicators to four components as shown in Figure 14 did have good model fit and met criteria to indicate construct validity of use of the *Educator Evaluation Instrument* for determining teachers' effectiveness.

Additionally, the *Educator Evaluation Instrument* does meet criteria as described as important components of an evaluation instrument in the literature:

- Multiple measures
- Teacher experience
- Observation and artifacts
- Teacher self-reflection
- Student growth measures
- Student growth models

The educator evaluation system does not yet use a teacher-student rostering mechanism, but no individual student growth measures are contributed the teacher systematically. In the future, if a teacher's growth score is to be calculated based on the students that he/she instructs (as opposed to a school-wide score, or a score derived from the

teacher-developed student growth goals), a teacher-student roster mechanism is recommended to be implemented. The new model, which will be organized differently, truly only requires a shift in the indicators them themselves to the four new components.

Recommendations

Based on the results of the HCFAs and the exploratory factor analysis for the data associated with the West Virginia Department of Education (WVDE) should reconsider use of the *Educator Evaluation Instrument* in its current form changes to the *Educator Evaluation Instrument* are necessary to make this instrument valid for use in determining the overall effectiveness of teachers. Assuming the professional teaching standards will continue to be utilized as part of the *Educator Evaluation Instrument*, the instrument should be modified to support good model fit.

In this new HCFA model, the summative ratings 1-6 for each of the teaching standards 1-6 are actually eliminated, and replaced with four different components, or first-order factors. The indicators associated with each of the six teaching standards remain as indicators, or sub elements in the model, but are organized differently than they were in the original model onto the four new components. These four components should be labeled to align with the standards that they represent and the meaning of the components.

Teaching standards 1-3 are related to teaching practices by nature:

- Standard 1 – Curriculum and Planning
- Standard 2 – The Learner and the Learning Environment
- Standard 3 – Teaching

Therefore, it is recommended that Component 1 be labeled as Teaching Practices in the *Educator Evaluation Instrument*, and that it be based on the indicators associated with each of the three teaching standards: 1.1, 1.2. 1.3, 2.1, 2.2, 2.3, 2.4, 3.1, 3.2, and 3.3.

Component 2 was composed of indicators associated with teaching standards 4 and 5:

- Standard 4 – Professional Responsibilities for Self-Renewal
- Standard 5 – Professional Responsibilities for School and Community

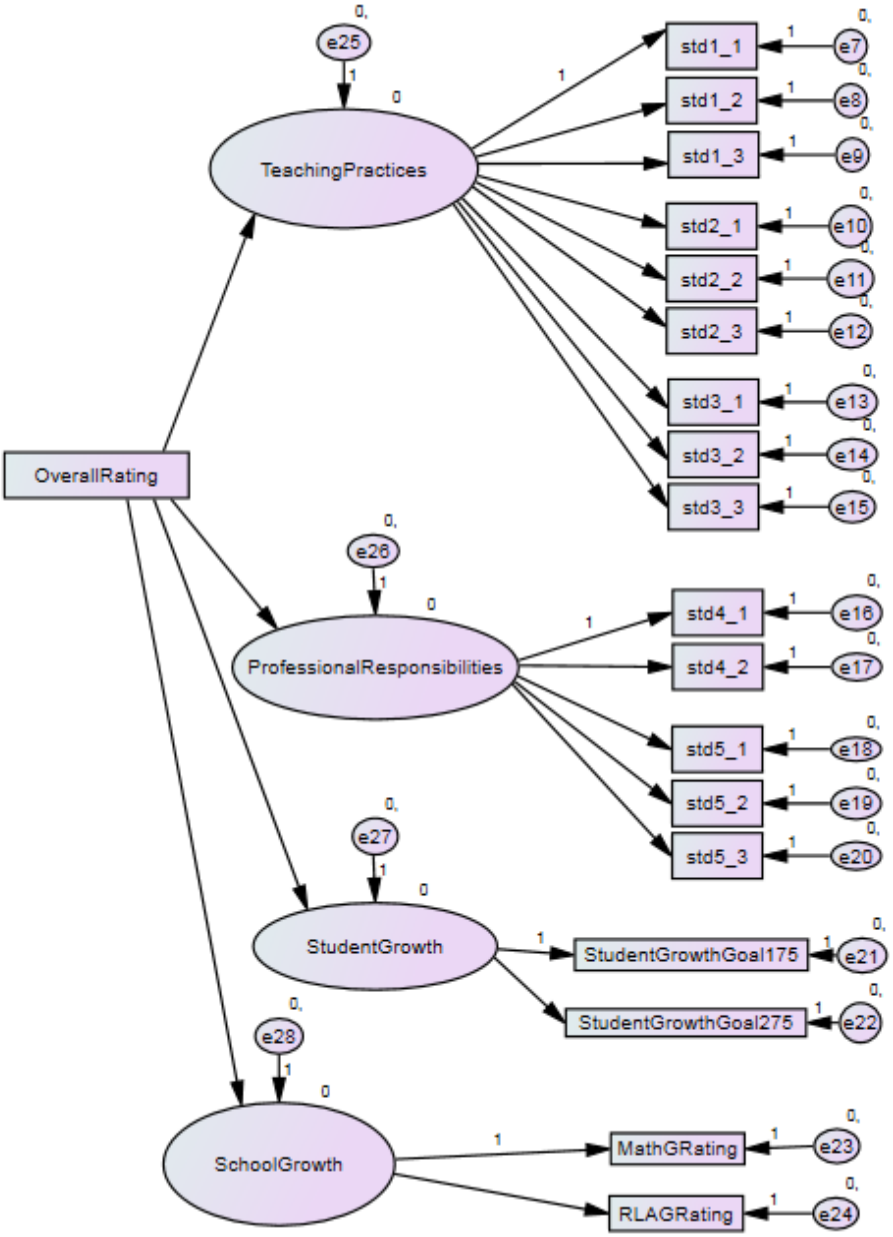
Both of these standards related to Professional Responsibilities of the teacher, so it is therefore recommended that Component 2 be labeled at Professional Responsibilities in the *Educator Evaluation Instrument* as component two is associated with the indicators 4.1, 4.2, 5.1, 5.2, and 5.3.

Component 3 is composed of a portion of the indicators for teaching standard 6. The portion consisted of the results from goals established at the beginning of the year by the teacher and then measured at a second point later in the year to determine growth. Each teacher is required to have these two goals associated with his/her students. Although teaching standard 6 is labeled as Student Learning, it is recommended that Component 3, which is composed of two indicators from standard 6, be labeled as Student Growth, because the two indicators that compose this component are related to the degree to which students show growth. It will also help to distinguish it from Component 4.

The last component, Component 4, consists of the mathematics school-wide growth score and the reading/language arts school-wide growth score. It is recommended that Component 4 be labeled School Growth, which represents the indicators of both of these scores.

The new labels as part of the new model are shown in Figure 18. This figure shows the overall shift in how the indicators should be organized to achieve good model fit, and construct validity for determining educator effectiveness.

Figure 18: Recommended Labels for Recommended HCFA Model



The weighting that was established for the ratings and indicators in the original *Educator Evaluation Instrument* are very closely retained based on the loadings of the four components on the overall rating of effectiveness. The weighting recommendations of the components takes into consideration the standardized regression weights of each of the

components on the overall rating, and the individual progressions, along with considering the law, and balancing practicality of the instruments use. The law is clear with respect to the use of growth scores as part of the educator evaluation instrument that 20% of the overall evaluation is to be based on growth scores. The standardized regression weights support the manner in which those were attributed previously and so can remain. The student growth component is based on goals set by the teacher regarding students' performance. In the original educator evaluation system, the student growth component was weighted at 15% of the overall score, and it is recommended that this weight of 15% remain in place in the new model. In the original model, school-wide growth was at weighted 5%, and it is recommended that the weighting of 5% remain intact in the new model. By maintaining the weights for the student growth component and the school growth component, the transition to the newly organized model by users will be smoother by retaining some aspects of the instrument as they were utilized.

As for the remaining 80% of the weight of the model, components 1 and 2, or the Teaching Practices and Professional Responsibilities components, show an almost even split with the standardized regression weights as shown in Table 15. Even though the standardized regression weights were not initially intended for examination as part of this study, they are part of the recommendation for weighting because they offer a solution based on the data. There are nine indicators that contribute to Teaching Practices and five that contribute to the Professional Responsibilities components; the weighting recommendations take into account the greater number of indicators within Teaching Practices while also considering the standardized regression weights.

It is recommended that the WVDE consider the following revisions to the weighting of the instrument as shown in Table 16.

Table 16

Recommended Components and Weights of the New Educator Evaluation Instrument

Component	Weight	% of Score
Teaching Practices	45%	80%
Professional Learning	35%	
Student Growth	student learning goal 1: 7.5% student learning goal 2: 7.5%	15%
School Growth	2.5% mathematics 2.5% reading	5%
Standard 7: Professional Conduct		Required, but does not count in the overall score
Total		100%

It is also recommended that once there are state-wide results, the results should be tested for normality and then run through the HCFA to confirm good model fit. With the inclusion of data from all of West Virginia's approximately 20,000 teachers, the West Virginia Department of Education can ensure that the *Educator Evaluation Instrument* is valid for use in determining an educator's effectiveness.

Lastly, it is also recommended that the progression levels be examined to ensure that the differences among these progressions are different enough to constitute different rules regarding observations, an important and critical component as described in the literature. Understanding more about the performance in these groups can help to inform the team at the

WVDE as to whether the progressions, and years of teaching that they are associated, truly represent the appropriate years of experience spans as they currently are defined.

To further the work of the WVDE to ensure that the *Educator Evaluation Instrument* is part of a comprehensive system of support that

- sets high standards of performance for both veteran and new teachers;
- ensures high-quality instruction focused on increasing student achievement;
- encourages continuous growth and improvement over time,

it is imperative that the newly configured model be adopted to ensure the valid use of the instrument as part of the system.

APPENDIX A: EVALUATION RUBRIC FOR TEACHERS

Evaluation Rubrics for Teachers

STANDARD 1: CURRICULUM AND PLANNING			
Element 1.1: The teacher demonstrates a deep and extensive knowledge of the subject matter.			
Distinguished	**Accomplished**	**Emerging**	**Unsatisfactory**
The teacher • demonstrates expert, specialized content knowledge • collaborates with teachers from other grades and subjects to extend and connect student learning to other content areas	The teacher • demonstrates extensive content knowledge • connects student learning to other content areas	The teacher • demonstrates content knowledge • attempts to connect student learning to other content areas	The teacher • does not demonstrate sufficient content knowledge • does not attempt to connect student learning to other content areas
Element 1.2: The teacher designs standards-driven instruction using state-approved curricula.			
Distinguished	**Accomplished**	**Emerging**	**Unsatisfactory**
The teacher • collaborates with others, including students, to design instruction and assessment aligned to the state-approved curricula • collaborates with students to design sequential learning activities that provide for varied student abilities and interests • collaborates with others, including students, to design learning activities that promote student collaboration, critical thinking and problem solving	The teacher • designs written instructional plans that align instruction and assessment to the state-approved curricula • designs sequential learning activities that provide for varied student abilities and interests • designs activities that promote student collaboration, critical thinking, and problem solving	The teacher • designs written instructional plans aligned to the state-approved curricula • designs sequential learning activities at appropriate developmental levels • designs activities that promote student collaboration	The teacher • does not design written instructional plans • does not design instructional plans and/or units that are driven by state-approved curricula • does not design sequential learning activities at appropriate developmental levels • does not design activities that promote student collaboration
Element 1.3: The teacher uses a balanced assessment approach to guide student learning.			
Distinguished	**Accomplished**	**Emerging**	**Unsatisfactory**
The teacher • collaborates with students to design and use a variety of assessments, including peer and student self-reflections, to monitor student progress and set learning goals • collaborates with students and others to clearly define and communicate assessment criteria • shares assessment data and provides timely feedback to students and other stakeholders	The teacher • designs and uses formative and summative assessments to monitor student progress and set learning goals • clearly defines and communicates assessment criteria • shares assessment data and provides timely feedback to students	The teacher • designs and uses formative and summative assessments • communicates assessment criteria • shares assessment data with students	The teacher • does not use formative and summative assessments • does not communicate assessment criteria • does not share assessment data or provide feedback to students

STANDARD 2: THE LEARNER AND THE LEARNING ENVIRONMENT			
Element 2.1: The teacher understands and responds to the unique characteristics of learners.			
Distinguished	**Accomplished**	**Emerging**	**Unsatisfactory**
The teacher • demonstrates extensive knowledge of students' social, emotional and academic needs, interests, learning styles, cultural heritage, and gender • plans and implements differentiated learning activities with students • helps colleagues understand the unique characteristics of all learners	The teacher • demonstrates thorough knowledge of students' social, emotional and academic needs, interests, learning styles, cultural heritage, and gender • plans and implements differentiated learning activities for students	The teacher • demonstrates adequate knowledge of students' social, emotional and academic needs, interests, learning styles, cultural heritage, and gender • plans and implements differentiated learning activities for some students	The teacher • does not demonstrate knowledge of students' social, emotional and academic needs, interests, learning styles, cultural heritage, and gender • does not plan and implement appropriate learning activities
Element 2.2: The teacher establishes and maintains a safe and appropriate learning environment.			
Distinguished	**Accomplished**	**Emerging**	**Unsatisfactory**
The teacher • collaborates with students to establish an effective classroom management system • collaborates with students to ensure appropriate behavior as defined by the code of conduct • organizes space and materials in a safe, highly efficient and well-designed learning environment	The teacher • establishes an effective classroom management system • responds appropriately and respectfully to student behavior as defined by the code of conduct • organizes space and materials to ensure safety and efficiency	The teacher • establishes a classroom management system • responds inadequately to student behavior as defined by the code of conduct • organizes space and materials to ensure safety	The teacher • does not implement an effective classroom management system • does not respond to student behavior as defined by the code of conduct • does not organize space and materials to ensure safety
Element 2.3: The teacher establishes and maintains a learner-centered culture.			
Distinguished	**Accomplished**	**Emerging**	**Unsatisfactory**
The teacher •establishes with students clear criteria for high-quality work •collaborates with students to maximize instructional time •engages students in active, self-directed learning as part of a community of learners •provides extensive opportunities for students to collaborate in learning	The teacher •sets and communicates clear criteria for high-quality work •uses instructional time efficiently •engages students in active learning •provides adequate opportunities for students to collaborate in learning	The teacher •sets criteria for high-quality work •uses instructional time with limited efficiency •engages students in learning •provides limited opportunities for students to collaborate in learning	The teacher •does not establish criteria for quality work •does not use instructional time efficiently •does not engage students in learning •does not provide opportunities for students to collaborate in learning

STANDARD 3: TEACHING			
Element 3.1: The teacher utilizes a variety of research-based instructional strategies.			
Distinguished	**Accomplished**	**Emerging**	**Unsatisfactory**
The teacher • collaborates with students to use an extensive variety of effective instructional strategies to deliver content • collaborates with students to provide scaffolding and differentiated instruction • extensively uses appropriate technology to deliver content	The teacher • uses a variety of effective instructional strategies to deliver content • demonstrates adequate use of scaffolding and differentiated instruction • adequately uses technology to deliver content	The teacher • uses a limited variety of effective instructional strategies to deliver content • demonstrates limited use of scaffolding or differentiated instruction • demonstrates limited use of appropriate technology to deliver content	The teacher • does not use effective instructional strategies to deliver content • does not scaffold or differentiate instruction • does not use appropriate technology to deliver content
Element 3.2: The teacher motivates and engages students in learning, problem solving and collaboration.			
Distinguished	**Accomplished**	**Emerging**	**Unsatisfactory**
The teacher • facilitates student-led learning activities leading to deep understanding of the content • encourages students to initiate or adapt learning activities to deepen understanding • provides students with extensive opportunities to collaborate and peer assess using appropriate technologies to gather information, problem solve and share learning	The teacher • provides learning activities relevant to the content that involve meaningful real-world experiences leading to deep understanding • explains directions and procedures clearly and models them when necessary • provides students with adequate opportunities to collaborate and peer assess using appropriate technologies to gather information, problem solve and share learning	The teacher • provides learning activities relevant to the content • explains directions and procedures • provides students with limited opportunities to collaborate using appropriate technologies	The teacher • does not provide learning activities that are relevant to the content • does not provide meaningful activities • does not explain directions and procedures • does not provide students opportunities to collaborate
Element 3.3: The teacher adjusts instruction based on a variety of assessments and student responses.			
Distinguished	**Accomplished**	**Emerging**	**Unsatisfactory**
The teacher •effectively modifies instruction to meet the needs of all students •extensively monitors student progress using a variety of assessments •collaborates with students and others to make instructional decisions •extensively analyzes and uses student data to make instructional decisions •uses a variety of formative assessments to differentiate instruction and provide effective interventions	The teacher •modifies instruction when need is apparent •consistently monitors student progress using a variety of assessments •uses student feedback to make instructional decisions •analyzes student data to make instructional decisions •uses a variety of formative assessments to differentiate instruction and provide appropriate interventions	The teacher •recognizes missed opportunities to modify instruction •inconsistently monitors student progress using a variety of assessments •examines student data •uses formative assessments to provide whole-group interventions	The teacher •does not modify instruction •does not monitor student progress •does not base instruction on a variety of assessments •does not provide interventions based on student data

STANDARD 4: PROFESSIONAL RESPONSIBILITIES FOR SELF-RENEWAL			
Element 4.1: The teacher engages in professional development for self-renewal that guides continuous examination and improvement of professional practice.			
Distinguished	**Accomplished**	**Emerging**	**Unsatisfactory**
The teacher • initiates the investigation that leads to the development of best practices • extensively implements best practices • mentors others in implementation of best practices • shares results of investigation at the local, state, or national level	The teacher • engages in professional learning to investigate best practices • consistently implements best practices • shares best practices within the school community	The teacher • participates in opportunities to investigate best practices when invited to do so • inconsistently implements best practices	The teacher • does not participate in professional development of best practices as required for self-renewal • does not implement best practices • does not implement best practices acquired through professional development to improve unsatisfactory performance rating
Element 4.2: The teacher actively engages in collaborative learning opportunities for self-renewal with colleagues.			
Distinguished	**Accomplished**	**Emerging**	**Unsatisfactory**
The teacher • initiates or advances the development of a collaborative team • contributes consistently to group learning • mentors others in utilizing the knowledge and skills gained	The teacher • participates actively in and/or facilitates a collaborative team • contributes to group learning • utilizes the knowledge and skills gained	The teacher • participates in a collaborative team when invited to do so • attempts to utilize the knowledge and skills gained	The teacher • works in isolation • does not contribute productively to work of collaborative teams as required for self-renewal • does not utilize knowledge and skills gained • does not utilize knowledge and skills gained to improve unsatisfactory performance rating

STANDARD 5: PROFESSIONAL RESPONSIBILITIES FOR SCHOOL AND COMMUNITY			
Element 5.1: The teacher participates in school-wide collaborative efforts to support the success of all students.			
Distinguished	**Accomplished**	**Emerging**	**Unsatisfactory**
The teacher • leads the ongoing development of school-wide initiatives based on school and student data • participates in the design and delivery of professional development for the implementation of school-wide initiatives	The teacher • collaborates in the development of school-wide initiatives based on school and student data • participates in the implementation of school-wide initiatives	The teacher • participates in school-wide initiatives	The teacher • does not participate in school-wide initiatives
Element 5.2: The teacher works with parents, guardians, families and community entities to support student learning and well-being.			
Distinguished	**Accomplished**	**Emerging**	**Unsatisfactory**
The teacher • develops ongoing opportunities for families to participate in classroom activities based on needs assessment • interacts appropriately with families within the school and community • utilizes theory and current research to facilitate meaningful connections between the school and family • develops and promotes meaningful school activities by utilizing community expertise and resources	The teacher • offers ongoing opportunities for families to participate in classroom activities • interacts appropriately with families within the school setting • seeks relevant knowledge of the family in order to provide meaningful connections between the school and family • creates positive connections between the school and the community	The teacher • participates in school-wide family activities • has minimal interaction with families • responds appropriately to contact from families • occasionally connects school activities with community resources	The teacher • does not attend school-wide family activities • does not respond or inappropriately responds to contact from families • does not positively contribute to the relationship between school and community
Element 5.3: The teacher promotes practices and policies that improve school environment and student learning.			
Distinguished	**Accomplished**	**Emerging**	**Unsatisfactory**
The teacher • involves and coaches others to implement and sustain teacher-identified change • takes a leadership role in growth initiatives that affect practice and policy throughout the school community	The teacher • identifies possible areas of growth within the classroom and school • recommends and facilitates opportunities for change and growth in the classroom and school	The teacher • participates in required initiatives leading to change in practice and policy in the classroom and school	The teacher • does not participate in available opportunities for change and growth that affect practice and policy

Student Learning

Student Learning is the single, most important goal of education. Many things affect students' quality of life and readiness to learn. The quality of teaching, however, is the most important school-related factor with the potential to impact student learning. The work of the teacher constitutes multiple dimensions that contribute to student achievement. This performance standard requires educators to demonstrate their students' success through multiple measures. The educator evaluation recognizes the professional commitment and hard work necessary for West Virginia students to achieve at high levels. It recognizes student growth in a variety of classrooms across diverse social and academic contexts.

STANDARD 6: STUDENT LEARNING			
Element 6.1: The work of the teacher results in measurable progress of student learning of state-approved curricula.			
Distinguished	**Accomplished**	**Emerging**	**Unsatisfactory**
Evidence from multiple measures consistently validates progress of student learning of appropriate state-approved curricula. The teacher accomplishes a student learning goal that involves collaborative efforts across classrooms.	Evidence from multiple measures consistently validates progress of student learning of the appropriate state-approved curricula.	Evidence from multiple measures does not consistently validate progress of student learning of the appropriate state-approved curricula.	Evidence from multiple measures does not validate progress of student learning of appropriate state-approved curricula.

Professional Conduct

Professional Conduct reflects the understanding that teaching is both a demanding and rewarding profession that involves a serious commitment to the highest standards of public service. This performance standard sets clear criteria for those competencies and habits of mind without which professional teaching simply cannot occur. The Professional Conduct standard allows educators to address areas of concern without necessitating an improvement plan. The Professional Conduct performance standard does not, however, supplant code and policy to which educators remain fully accountable and is not determinative of whether behavior is correctable. Certain violations may be cured by implementation of an improvement plan; others will require immediate action.

STANDARD 7: PROFESSIONAL CONDUCT			
Element 7.1: The teacher demonstrates professional conduct as defined in law, policy and procedure at the state, district, and school level.			
	Meets Standard	**Below Standard**	**Unsatisfactory**
Policy and Procedure	Adheres to state, district and school policy and procedure	Adheres to state, district and school policy and procedure with few exceptions	Demonstrates a pattern of violating state, district or school policy and procedure
Attendance	Adheres to state, district and school attendance policy and procedure	Adheres to state, district and school attendance policy and procedure with few exceptions	Demonstrates a pattern of absences that violate state, district or school attendance policy and procedure
Schedule	Adheres to state, district and school work schedule policy and procedure	Adheres to state, district and school work schedule policy and procedure with few exceptions	Demonstrates a pattern of failure to adhere to the work schedule defined by state, district, or school policy and procedure
Respect	Interacts professionally with students, parents/ guardians, colleagues and community	Interacts professionally with students, parents/ guardians, colleagues and community with few exceptions	Demonstrates a pattern of behavior with students, parents/guardians, colleagues and/or community which is unprofessional

REFERENCES

Ball, D. L., & Bass, H. (2000). Interweaving content and pedagogy in teaching and learning to teach: Knowing and using mathematics. In J. Boaler (Ed.) *Multiple perspectives on the teaching and learning of mathematics* (p. 83-104). West port, CT: Ablex.

Betebenner, D. (2009), Norm- and Criterion-Referenced Student Growth. Educational Measurement: Issues and Practice, 28: 42–51. doi: 10.1111/j.1745-3992.2009.00161.x

Betebenner, D.W., VanIwaarden, A., Domingue, B., & Shang, Y. (2013). SGP: An R Package for the Calculation and Visualization of Student Growth Percentiles & Percentile Growth Trajectories. (R package version 1.1-0.0. URL http://schoolview.github.com/SGP/

Bushaw, W. J., & Lopez, S. J. (2013). The 45th annual PDK/Gallup poll of the public's attitudes toward the public schools. *Phi Delta Kappan*. Retrieved December 6, 2013, from http://pdkintl.org/noindex/2013_PDKGallup.pdf

Cook, T.D. and Campbell, D.T. (1979). Quasi-experimentation: Design and analysis issues for field settings. Rand McNally College Publishing Company, Chicago, IL.

Cullingford, C. (1995). The Characteristics of an Effective Teacher. In *The effective teacher* (p. 25). London: Cassell.

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

Danielson, C. (2008). *The handbook for enhancing professional practice: Using the framework for teaching in your school*. Alexandria, VA: Association for Supervision and Curriculum Development.

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

Databases on State Teacher and Principal Evaluation Policies. (2013). *(STEP Database and SPEP Database)*. Retrieved October 27, 2013, from http://resource.tqsource.org/stateevaldb/Compare50States.aspx

De Jong, Z., Van der Heijde, D., McKenna, S. P., & Whalley, D. (1997). The reliability and construct validity of the RAQoL: Rheumatoid arthritis specific quality of life instrument. *British Journal of Rheumatology*, *36*(8), 878-883. Retrieved October 14, 2013, from http://www.swetswise.com.proxy.lib.wayne.edu/FullTextProxy/swproxy?url=http%3A%2F%2Frheumatology.oxfordjournals.org%2Fcgi%2Freprint%2F36%2F8%2F878.pdf&ts=1381771400506&cs=354873911&userName=24481.ipdirect&emCondId=24481&articleID=35171229&yevoID=2166285&titleID=177891&remoteAddr=141.217.20.120&hostType=PRO

DPAS II - Delaware Performance Appraisal System. (2013). *State of Delaware*. Retrieved December 18, 2013, from http://www.doe.k12.de.us/csa/dpasii/

Feedback for better teaching - nine principles for using measures of effective teaching. (2013). *MET Project*. Retrieved October 19, 2013, from http://www.metproject.org/downloads/MET_Feedback%20for%20Better%20Teaching_Principles%20Paper.pdf

Florida Department of Education. (2013). *District Performance Evaluation Systems*. Retrieved December 21, 2013, from http://www.fldoe.org/profdev/pa.asp

Gallagher, C. (2012). *Presenting findings from measures of teacher effectiveness: Key reporting considerations for states and districts* (Rep.). WestEd.

Goe, Laura. *The Link between Teacher Quality and Student Outcomes: A Research Synthesis*. Tech. Washington, DC: ETS, 2007. Print.

Goldhaber, D. D., & Brewer, D. J. (2000). Does teacher certification matter? High school teacher certification status and student achievement. *JSTOR*. Retrieved November 4, 2013, from http://www.jstor.org/stable/1164392

Hamilton, J. (2010). Delaware and Tennessee Win First Race to The Top Grants. U.S. Department of Education. Retrieved September 05, 2013, from http://www.ed.gov/news/press-releases/delaware-and-tennessee-win-first-race-top-grants

Hawaii DOE | Educator Effectiveness. (2013). *Hawaii DOE Educator Effectiveness*. Retrieved December 21, 2013, from http://www.hawaiipublicschools.org/TeachingAndLearning/EducatorEffectiveness/Pages/Home.aspx

House Bill 4627 of 2011, 96th Legislature, Regular Session. (2011).

IMPACT - The DCPS Effectiveness Assessment System for School-Based Personnel. (2011). DC Public Schools, Washington, DC. Retrieved October 18, 2013, from http://www.dc.gov/DCPS/impact

Index Page for the ESEA Flexibility Page. (n.d.). Index Page for the ESEA Flexibility Page. Retrieved September 5, 2013, from http://www2.ed.gov/policy/elsec/guid/esea-flexibility/index.html

Kline, R. B. (2011). *Principles and practice of structural equation modeling*. New York: Guilford Press.

MDE - Michigan Educational Assessment Program (MEAP). (2013). *MDE - Michigan Educational Assessment Program (MEAP)*. Retrieved December 6, 2013, from http://www.michigan.gov/mde/0,1607,7-140-22709_31168---,00.html

Michigan Department of Education, Office of Professional Preparation Services. (2013). *2013 Facts on Educator Certification*. Retrieved December 6, 2013, from http://www.michigan.gov/documents/mde/Facts_About_Teacher_Certification_In_Michigan_230612_7.pdf

Meharie A., & Hixson, N. (2013). West Virginia revised educator evaluation system for teachers 2011-2012: First year pilot report. Charleston, WV: West Virginia Department of Education, Division of Teaching and Learning, Office of Research.

Murphy, E. (2004). Promoting construct validity in instruments for the analysis of transcripts of online asynchronous discussions. Educational Media International, 41:4, 347-354, DOI: 10.1080/0952398042000314848

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. doi: 10.1111/j.1745-3984.2007.00026.x

NCLB Not Closing Test Score Gaps. (2006). *The National Center for Fair & Open Testing*. Retrieved October 27, 2013, from http://www.fairtest.org/nclb-not-closing-test-score-gaps

New No Child Left Behind Flexibility: Highly Qualified Teachers. (2004). *New No Child Left Behind Flexibility: Highly Qualified Teachers*. Retrieved December 21, 2013, from http://www2.ed.gov/nclb/methods/teachers/hqtflexibility.html

Nye, B., Konstantopoulos, S., & Hedges, L. V. (2004). How Large Are Teacher Effects? Educational Evaluation and Policy Analysis, 26(3), 237-257. doi: 10.3102/01623737026003237

Odden, A. (2004). Lessons learned about standards-based teacher evaluation systems.*JSTOR*. Retrieved November 4, 2013, from http://www.jstor.org/stable/1493311

Palardy, G. J., & Rumberger, R. W. (2008). Teacher Effectiveness in First Grade: The Importance of Background Qualifications, Attitudes, and Instructional Practices for Student Learning. *Educational Evaluation and Policy Analysis*, *30*(2), 111-140. doi: 10.3102/0162373708317680

Peterson, K. D., & Peterson, C. A. (2006). *Effective teacher evaluation: A guide for principals*. Thousand Oaks, CA: Corwin Press.

Rockoff, J. E. (2004). The Impact of Individual Teachers on Student Achievement: Evidence from Panel Data. American Economic Review, 94(2), 247-252. doi: 10.1257/0002828041302244

Roster verification. (2013). *Public Schools of North Carolina Educator Effectiveness Model*. Retrieved November 21, 2013, from http://www.dpi.state.nc.us/effectiveness-model/roster/

Rothstein, J. (2009). Student sorting and bias in value added estimation: Selection on observables and unobservables. *NBER*. Retrieved January 1, 2014, from http://www.nber.org/papers/w14666

Rothstein, R., Ladd, H. F., Ravitch, D., Baker, E. L., Barton, P. E., Darling-Hammond, L., … Shepard, L. A. (2010). *Problems with the use of student test scores to evaluate teachers* (Rep. No. 278). Retrieved November 21, 2013, from http://www.epi.org/publication/bp278/

Runyon, R.P., Coleman, K.A., & Pittenger, D.J. (2000). *Fundamentals of Behavioral Statistics* (9th ed.) Boston, MA: McGraw-Hill.

SAS Institute Inc. (2013). NC DPI Login. *NC DPI Login*. Retrieved December 6, 2013, from https://ncdpi.sas.com/

Sawilowsky, S. S. (2007). Construct validity. In (Neil J. Salkind & Kristin Rasmussen, Eds.) *Encyclopedia of Measurement and Statistics, 1*, 178-181. Thousand Oaks, CA: Sage. DOI: http://dx.doi.org/10.4135/9781412952644.

Student Growth Percentiles. (2013). *Office of Superintendent of Public Instruction*. Retrieved December 23, 2013, from https://www.k12.wa.us/assessment/studentgrowth.aspx

Student Growth Percentiles – Georgia's Student Growth Model. (2013). *Georgia Department of Education*. Retrieved December 23, 2013, from http://www.gadoe.org/School-Improvement/Teacher-and-Leader-Effectiveness/Pages/Student-Growth-Percentiles.aspx

Teacher and Leader Effectiveness (TLE). (2013). *Oklahoma State Department of Education*. Retrieved December 21, 2013, from http://ok.gov/sde/tle

Tennessee Department of Education. (2013). *TCAP and TVAAS Scores*. Retrieved December 6, 2013, from http://www.tn.gov/education/assessment/test_results.shtml

TLE Roster Verification. (2013). *Oklahoma State Department of Education*. Retrieved November 21, 2013, from http://ok.gov/sde/tle-roster-verification

Trochim, W. M. (2006). Threats to Construct Validity. *Threats to Construct Validity*. Retrieved October 14, 2013, from http://www.socialresearchmethods.net/kb/consthre.php

Tulsa Public Schools TLE Observation and Evaluation Rubric Teachers. (2012). Retrieved December 18, 2013, from http://ok.gov/sde/sites/ok.gov.sde/files/TLE-TPSFramework12.pdf

Wang, J. (2012). *Wiley series in probability and statistics: Structural equation modeling with MPlus : Methods and applications* (3rd ed.). Somerset, NJ: Wiley.

What Matters Most: Teaching for America's Future. (1996). Retrieved November 11, 2013, from http://nctaf.org/wp-content/uploads/WhatMattersMost.pdf

West Virginia Educator Evaluation System for Teachers. West Virginia Department of Education, n.d. Web. 25 May 2014. http://wvde.state.wv.us/evalwv/documents/TeacherEvaluationGuidebook2013_001.pdf

West Virginia Professional Teaching Standards. West Virginia Department of Education, n.d. Web. 30 Dec. 2013. <http://wvde.state.wv.us/teachwv/profstandards.html>.

WV Principal & Teacher Evaluation. West Virginia Department of Education, n.d. Web. 10 Aug. 2013. <http://wvde.state.wv.us/evalwv/>.

ABSTRACT

VALIDITY OF THE *EDUCATOR EVALUATION INSTRUMENTI* IN THE STATE OF WEST VIRGINIA

by

CARLA HOWE

August 2014

Advisor: Dr. Shlomo Sawilowsky

Major: Educational Evaluation and Research

Degree: Doctor of Philosophy

In the state of West Virginia, the educator evaluation system was implemented in 2010 as part of a comprehensive system of support to increase teacher effectiveness and student learning. As part of the system, the *Educator Evaluation Instrument* was developed to measure teachers' effectiveness. This study was conducted to determine whether the *Educator Evaluation Instrument* was valid for use in measuring effectiveness.

A hierarchical confirmatory factor analysis (HCFA) was conducted on the scores from the demonstration year. The data were not normal, nor was good model fit established based on the current model. Because good model fit could not be established, an exploratory factor analysis was conducted, and four components were extracted and utilized as the first-order factors in the HCFA through principal component analysis. With the new model, good fit was established, and therefore redesigning the *Educator Evaluation Instrument* to align with the new components is recommended to ensure validity of use.

AUTOBIOGRAPHICAL STATEMENT

Carla Howe began her professional work as a teacher of middle school students in Spanish and English. While teaching, she began her graduate work in the area of Educational Evaluation and Research at Wayne State University, and obtained a Master's of Education in 2006.

After teaching for nearly a decade, she transitioned her professional work to the State Education Agency as an Education Research Consultant at the Michigan Department of Education (MDE) in Lansing, Michigan. In the Bureau of Assessment and Accountability, Carla worked on issues related to school accountability and accreditation, statewide assessments, and educator evaluations.

Most recently, Carla lives and works in West Virginia at the State Education Agency as the Data Governance Manager.