3-15-2016

A Comparison of Educational "Value-Added" Methodologies for Classifying Teacher Effectiveness: Value Tables vs. Covariate Regression

Theodore J. Dwyer
University of South Florida, ted.dwyer@gmail.com

A Comparison of Educational "Value-Added" Methodologies for Classifying Teacher Effectiveness:

Value Tables vs. Covariate Regression

by

Theodore J. Dwyer

A dissertation submitted in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
in Curriculum and Instruction
with an emphasis in Measurement and Evaluation
College of Education
University of South Florida

Major Professor: John M. Ferron, Ph.D.
Yi-Hsin Chen, Ph.D.
Robert F. Dedrick, Ph.D.
Donald A. Dellow, Ed.D.

Date of Approval:
March 04, 2016

Keywords: VAM, Transition Tables, Teacher Evaluation, Student Achievement

Dedication

It truly takes a village. This dissertation is dedicated to those who have provided support for its completion: my Loved Ones, Friends, and Colleagues.

Acknowledgements

I would like to acknowledge the academic support that I received from my dissertation committee, most especially Dr. John Ferron – his patience and support through all of the years, edits, changes, and discussions was invaluable and greatly appreciated.

Table of Contents

List of Tables

List of Figures

Abstract

There is a great deal of concern regarding teacher impacts on student achievement being used as a substantial portion of a teacher's performance evaluation. This study investigated the degree of concordance and discordance between mathematics teacher ranking using value tables and covariate regression, which have both been used as measures for teacher effectiveness. The researcher examined teacher rankings, before and after the state recommended classification, using correlational techniques, comparison matrices, and visual examination for value-added scores derived from the value table versus the covariate regression approach. Examination demonstrated strong correlations between the initial rankings (r = .77 to .98) and a high concordance (γ = .96 to 1.0) once the recommended classifications were applied to the teachers rankings. The overall implications of this project are that more complex methods may parse the impact information out with higher statistical accuracy, however, once the recommended classification is applied to the methods there may be very little difference in the classification of teachers.

Chapter 1

Introduction

Value Add is a process used in business to look at the value or outcome of a product or service. The overarching rationale for value-added analyses is relatively simple –the value added is final value/score of a product/outcome impacted once a process/treatment is applied to the business/individual. The use of a Value Add approach that examines what value some process adds to a product can be used in many areas, including education. When applied to education, Value Added Systems are used to identify the impact of a teacher or program on a student based on that student's performance. This study examines the convergence and divergence between two Value Added methodologies used in teacher education evaluation systems. For this study "Value Added" refers to the process of using student achievement data to make evaluative statements about teacher effects on student achievement.

Value-added systems in education are methods of examining student achievement data to determine the extent to which students have demonstrated gains or losses over time. These gains and/or losses are then attributed to the teachers and schools responsible for those students. This methodology can be instrumental in examining pedagogical and curricular processes, and is often used to rate or rank individual teachers based on the academic growth of students in her or his classroom. In educator evaluation systems, Value Added processes are used to quantify the impact teachers have on their students' outcomes. The resulting evaluation is often used in teacher retention and compensation decisions.

Florida Policy for Teacher Evaluation Linked to Student Achievement in Florida

Public education institutions in Florida have a constitutionally mandated responsibility to ensure that:

> Adequate provision shall be made by law for a uniform, efficient, safe, secure, and high quality system of free public schools that allows students to obtain a high quality education and for the establishment, maintenance, and operation of institutions of higher learning and other public education programs that the needs of the people may require. [*FL Constitution Article XI, Section 1(a)*].

School districts in Florida seek to recruit and retain effective teachers to provide high quality education and fulfill this mandate.

One component of the teacher retention process is the evaluation of teachers. Teachers who appear to be struggling based on data reviews and observations by district administrative personnel are targeted for professional development to improve their performance. As part of the personnel evaluation system, teachers create individual professional development plans based on their own perception of their professional needs and goals, which include their perceived areas of improvement. Those teachers who are consistently identified as not meeting the needs of the students and conforming to the educational requirements of the district and state are provided due process to demonstrate improvement and then may be invited to leave the profession. In order for the culmination of this process to occur (i.e., ineffective teachers being asked to leave) there must be clear evidence that the teacher is actually performing at a substandard level. In reality, this rarely comes to fruition. Whereas there are mechanisms to identify teachers who are struggling, educational organizations continue to grapple with ways to quantify and utilize more objective methods to consistently identify teacher performance, more importantly, before the tenure process contributes to the retention of ineffectual teachers.

Legislative attempts, at both the federal and state levels, to assist educators in identifying and retaining highly effective teachers, and simultaneously identifying ineffective teachers, include performance pay. Performance pay systems combine teacher personnel evaluations with various value-added systems that link student achievement outcomes to specific teachers to assist in quantifying teacher quality/value. Some of the value-added methods currently employed in public education districts in the United States include pre-test and post-test comparisons, value tables, and multilevel modeling. State legislative actions have paralleled the Federal initiatives for performance pay.

Florida Performance Pay

The genesis for the performance pay movement in Florida began in 1998 with Florida Statute (Title XVI, 231.29) adding a requirement that student achievement be used to evaluate teachers, combined with other legislation [Title XVI, 230.23 (5) (c)] which required that a portion of a teacher's salary be linked to an annual performance appraisal. The performance pay plans instituted by the Florida districts often required teachers to apply annually to be considered for the bonus, and required extra work on the part of the teacher to be eligible (Office of Program Policy Analysis and Government Accountability, 2007). Moreover, the initial pay for performance legislation required that districts create a special fund from their operating budget to pay the bonuses.

In 2006, the Florida legislature attempted to institute a more equitable and less cumbersome performance pay plan within the K-12 school districts. The first of these attempts was known as "Effective Compensation" (E-Comp). The intent was to create a salary incentive that would reward teachers based on their students' academic achievement. Like the earlier legislation, the E-Comp program required that the districts fund the performance pay from their

existing budget. Over the next two years, the legislature proposed two other pay for performance plans: the Special Teachers Are Rewarded (STAR) program and the Merit Award Program (MAP F. S. 1012.225 Merit Award Program for Instructional Personnel and School-Based Administrators, 2007, 2008, 2009, 2010).

Whereas the legislature wanted districts to fund the 1998 and the 2006 E-Comp performance pay plan from their existing budgets, the STAR and MAP salary incentive programs (which have since been repealed during the 2011 legislative session) each provided for $147.5 million in state funding annually. The MAP performance pay system for the state of Florida provided K-12 institutions with funding to provide merit pay to their teachers. School districts who desired to implement MAP were required to submit a plan outlining their performance pay system annually to the state for approval. Only seven of the 67 Florida school districts elected to adopt the MAP performance pay system. The number of participating districts decreased each year until only three of the 67 districts participated in the program by 2011. This was identified as one of the reasons that in 2011 the legislature repealed the MAP law (HB 7087, 2011).

Seven Florida districts originally adopted MAP as a performance pay system. The plan approved for Hillsborough County Public Schools (HCPS) utilized a value table approach for identifying the highest performing teachers in the district. HCPS began implementing MAP in the 2006 – 2007 school year. This state-approved plan utilized a weighted combination of student achievement gains (60%) and performance appraisals (40%) which were converted to a percentile rank. Between the 2006-2007 and the 2008-2009 school years, the percentage of personnel receiving "perfect" ratings on their performance appraisals increased to nearly 90% of eligible individuals in the district, which when placed in juxtaposition with the achievement

data for students, revealed a stark inconsistency of lower student achievement instead of increased achievement. While the intent was to build a process that used multiple vectors of information, the consistently high evaluations across all teachers essentially removed any information that could have been included fairly to differentiate the effective teachers from the ineffective teachers. The HCPS value-added process had become the de facto arbiter of identification of high-quality teachers because there was extremely low variance in the performance appraisal ratings.

National Policy

The federal No Child Left Behind (NCLB, 2001) legislation established a requirement for an accountability system for any federally funded educational system. The requirements of NCLB were that school level achievement would be reported and made available to parents on state-wide testing aggregated by race, free and reduced lunch status, English Language Learners (ELL), and Students with Disabilities. As part of the federal accountability system established by NCLB, the U.S. Department of Education extended the option of using growth models to all states in 2007. NCLB linked federal funding to the implementation of the accountability system and formalized a set of sanctions for those federally funded schools who failed to meet the state targets for the year. For many states with pre-existing accountability systems, like Florida and Illinois, the requirements of NCLB created dual accountability systems. It also set the standard for all states to implement a state-wide testing system for students in 3rd through 8th grade. In July of 2009, President Obama and Secretary of Education Arne Duncan announced the creation of Race To The Top (RTTT), a $4.35 billion funded competitive grant program. The requirements of RTTT included integration of a value-added system into the teacher evaluation system, the adoption of common standards across states, increased use of

computers, and increased support for charter schools. States responded by instituting legislation that required their teacher evaluation systems to comply with the RTTT requirements (e.g., Florida: Student Success Act, SB 736, 2011). The changes to Florida state law require a substantial portion of the teacher evaluation system to be based on student achievement. In response to the requirement of the 2011 Florida: Student Success Act, the Florida Department of Education (FLDOE) selected a covariate regression approach for the calculation of teacher effects to be included in the overall evaluation.

By statute (Florida Statute 1012.34, Personnel Evaluation Procedures and Criteria), the Florida state evaluation system currently has four categories: 1. *Highly Effective*; 2. *Effective*; 3. *Needs Improvement* or, for instructional personnel in the first three years of employment who need improvement, *Developing*; and 4. *Unsatisfactory*. The FLDOE has provided guidance for classifying the value add scores of teachers into the evaluation categories as follows: two standard deviations (SD) above the mean is Highly Effective; Less than two SD above the mean and more than one SD below the mean is Effective; one SD below the mean is Needs Improvement; and two SD below the mean is Unsatisfactory (Copa, 2012).

Rationale for the Study

Florida's current covariate regression approach uses complex statistical modeling of multi-year student, classroom and school data to calculate a point value at the individual student level that is then aggregated at the teacher level with additional information from the overall school data. The covariate regression approach used for Florida's adopted value-added model is complex and not replicable or verifiable by educational stakeholders (teachers and principals) and difficult to understand (confusing, ambiguous) by other stakeholders outside the education system (community and parents). Further, replicating it within a district or organization is

unrealistic without access to the entire state's individual student data and all the teacher/class records. A value table is a value-added approach that utilizes pre-test and post-test data and assigns a value to a change in achievement level from the pre-test to the post-test. This provides a point value assigned to the change between a pre-test and a post-test (Dougherty, 2007, 2008). It is simpler and may provide the level of information necessary to achieve the statutory requirement. If such a model is found consistent with the more complex model, it has the added advantages of being easier to replicate and verify by educational stakeholders (teachers and principals) and being more understandable to other stakeholders (parents and the community). Furthermore, this would provide some validation evidence of value tables in relation to the covariate model for identifying high quality teachers.

The primary goal of this study was to investigate the consistency of teacher evaluation classifications using two value-added procedures: value tables and the covariate regression model currently mandated by the state of Florida for use in the state teacher evaluation system. This study was designed to investigate a question of parsimony and is directly linked to the current statutory requirement in Florida that student achievement counts as a substantial portion of a teacher's performance evaluation. This is important, given that the state requires district administrators to make retention decisions based on teachers' evaluations. This study did not examine the policy implications of this legislation; it only examined if there was a differential effect by procedure on this variable. This project was designed as a comparison of two value-added approaches without delving into the possible differences between tests (e.g., mathematics versus Language Arts), therefore, the analysis was restricted to the mathematics test scores and mathematics teachers.

Research Questions

1. What is the degree of concordance and discordance between the mathematics teachers' ranking using value-added scores derived from the value table approach versus the covariate regression approach?
2. What is the degree of concordance and discordance of the categories to which mathematics teachers are assigned when the state's recommendations for the classification of teachers into the four evaluation categories are applied to their value-added scores by the value table approach versus the covariate regression approach?

Overview of the Study

The study was a comparison of the application of two methodological approaches to deriving teacher value-added scores. The study compared the teacher scores derived from each of the two value-added approaches to examine if teachers would be classified differently based on the different procedures. The two models are the state method, which uses teachers' aggregated student residuals from the state adopted covariate regression model, and the value table method, which used the teachers' aggregated student values from a value table derived from the individual teacher's assigned students' achievement.

Value-added scores were derived or obtained for each student in the cohort of a large school district in Florida. The file containing the state-adopted student level residuals from the covariate regression was requested from the district. The student values were also computed using the value table approach. The value-added score for each mathematics teacher was computed using the two separate methods. A comparison was conducted to look at the consistency between the scores. The state recommendations for classification of teacher value-

added scores into evaluation categories was applied and a further comparison between the methods was conducted.

Data Source

Student and teacher data was obtained from a school district in Florida. These data were anonymized and linked through the use of an encoded student number and an encoded teacher number. The data was requested for the following school years: 2010-2011 and 2011-2012 to examine the consistencies for the data derived over two years. The student-level data requested included achievement on the Florida Comprehensive Assessment Test (FCAT) using the Mathematics Developmental Scale Scores (MDSS), mathematics course enrollment, and the student-level residual information provided by the state. The teacher data included mathematics course of instruction, school of instruction, and state teacher value-added score.

Significance

This study investigated differences between methodological approaches for measuring teacher impact on students. Most comparisons of value-added measures focus on the accuracy of a methodological approach or examining the effects of changes to a specific approach. Those investigations and projects are extremely important to the research and the practitioner communities as they search for the most accurate methodological approach for using student achievement information to reflect teacher impact in the classroom. This study did not focus on the granularity of differences between changing one aspect of a methodological approach. Rather, it focused on the question of parsimony as it relates to the classification of teachers into categories for the teacher evaluations. This is important because, in Florida student outcomes are required in the evaluations, and teacher evaluations are tied directly to financial implications for both the teachers and for the local education agencies. In order to ensure equity and parity

for teachers and to maintain a defensible approach to teacher evaluation it is important to use a method that is accurate, understandable, and, if possible, replicable. This does not argue that the more complex approach is not providing more information concerning the difference between individual teachers' impact on their students' achievement. The argument is that the information from the more complex model is collapsed into a classification system that ultimately removes much of the granularity of that information. The complex regression approach used by many states (including Florida) is not replicable by non-technical individuals. Further, it cannot be replicated within a district by individuals with the technical expertise required because in order to replicate the results within a single district they would need access to the entire State's individual student data records.

Limitations/Delimitations

It is important to understand the frame within which this project was operating. As with most research projects there are numerous limitations and weaknesses that are endemic to a single project rather than a complete body of research. For this project some of the areas that should be considered when placing it within the larger context of value-added research are, areas that would be worthwhile future projects. For example, the current evaluation has been restricted to a comparison of a relatively complex and robust model for deriving a teacher's value-added based on each teacher's students' achievement. There are other methods of deriving a teacher's added value to each student, however it is not feasible to include the other complex methods in the comparison due to the lack of access to the individual scores and teacher information for the entire state. Linked to the availability of other methodological approaches is that the results of this project will not be generalizable to the other complex models that are commonly used in education. Additionally, this project does not explore the existing artifacts of

the education system overall (i.e., hiring practices that lead to sorting), professional development requirements and procedures, or other social variables (e.g., poverty or ethnicity).

Definition of Terms

The following definition of terms are specific to this study.

Value-added systems in education. Value-added systems are methods of examining student achievement data to determine the extent to which students have demonstrated gains or losses over time.

Value-added. Value-added is the process of using process data to make a summative statement about an output. For this study, Value-added refers to student achievement data to make evaluative statements about teacher effects on student achievement. Value-added has a broader definition outside of this study.

Value-added modeling. Value-added modeling (VAM), is a method of examining academic progress over time regardless of level of proficiency (Rubin, Stuart, & Zanutto, 2004).

Value table. A value-added approach that utilizes pre-test and post-test data, which assigns a value to a change in achievement level from the pre-test to the post test. This provides a point value assigned to the change between a pre-test and a post-test (Dougherty, 2007, 2008). Value tables are sometimes referred to as transition tables.

Florida comprehensive assessment test. The Florida Comprehensive Achievement Test (FCAT) is a criterion reference test developed and used in Florida to satisfy the assessment requirements as per the Elementary and Secondary Education Act.

Mathematics developmental scale scores. The mathematics developmental scale score (MDSS) is a score on a vertical scale that reflect the achievement of a student on the associated test (retrieved from http://fcat.fldoe.org/mediapacket/2013/pdf/2013UFR.pdf).

Chapter 2

Literature Review

Systems in education that are used to evaluate teachers are often linked directly to or have an implied basal connection to the students and their learning. Systems that are used to tease out the impact of a teacher on a student are direct teacher observation, simple gains model using a pre-test and post-test analysis, and complex regression models. This literature review provides a broad overview of tacit concerns expressed for many years concerning teacher evaluations systems and the inclusion of student outcomes in the evaluation of teachers, and provide an overview of student achievement accountability approaches used in Florida. The review also examines sorting and poverty based on the existing research related to the non-randomness from sorting and the impact that poverty has on student outcomes, and then provide some of the examinations of value-added systems as they have been applied and tested in educational venues.

Teacher Evaluation Systems

Historically, research into improving teacher evaluation systems and improving teacher effectiveness has yielded mixed conclusions. The majority of researchers have expressed frustration over the perceived uselessness and lack of application of the existing research (Boyce, 1915; Darling-Hammond, Amrein-Beardsley, Haertel, & Rothstein 2012; Haefele, 1992, 1993; Lamke, 1955; Yamamoto, 1963). Researchers also indicate that there is a general lack of clarity surrounding what purpose they should fulfill and it is also often not clear what

actual process a teacher evaluation should follow (Brock, 1981; Lantham & Wexley, 1982; Lower, 1987; Scriven, 1980; Wise et al., 1984). Some have argued cogently that difficulties arise from the different uses, needs, and purposes of the evaluation systems (Darling-Hammond, Wise, & Pease, 1983; Smith & Fey, 2000). Others have argued that the overall needs to unify the teacher evaluation issues are rooted in a deeper need for a performance management system for education (Wiener & Jacobs, 2011). Weiner and Jacobs (2011) argue that a performance management system would not just implement improvements to an evaluation system; it would also result in increasing teacher effectiveness and student achievement. As detailed above, there has been a consistent outcry for improving teacher evaluation systems there has not been a high level of agreement on the most appropriate methods or techniques to use in an evaluation system

The changes to the design of teacher evaluations have many advocates, however, they do not all agree as to the appropriate approach. Some have argued that there should be no inclusion of a measure of student achievement (e.g., value-added measure) (Darling-Hammond, Amrein-Beardsley, Haertel, & Rothstein 2012; Medley, Coker, & Soar, 1984). Others have argued that the most logical and defensible measure of teachers' performance is their students' achievement (Goldhaber, 2002; Goldhaber & Brewer, 2000; Hanushek, Kain, O'Brien, & Rivkin, 2005; Wright, Horn, & Sanders, 1997). The passionate arguments and disparate positions of each group provide a partisan environment that makes it seem unlikely that they would be combined, however there have been attempts to blend the two, adopting the strengths from each approach (Aaronson, Barrow, & Sander, 2007; Doyle & Han, 2012; Haertel, 2009; Hanushek, Kain, O'Brien, & Rivkin, 2005; Rockoff, 2004; Scherrer, 2011; Weisberg, Sexton, Mulhern, & Keeling, 2009). There have been strenuous efforts at local levels to build an

observation method that is integrated with measures of student achievement for the overall evaluation. These local attempts have occurred in the District of Columbia; Charlotte-Mecklenburg, North Carolina; San Francisco, California; and Hillsborough County, Florida (Curtis 2012a, 2012b; District of Columbia Public Schools, 2010). There has also been a push for the same type of integration at the national level, specifically the RTTT initiative which provides funding to states with the requirement that both teacher observations and student achievement measures are included in the overall evaluation of the teacher.

The supporters of teacher observations founded their argument in the perception that observation of teachers in the classroom provides a glimpse into the pedagogical practices that occur inside of the classroom. The traditional observation method used to collect this information is for the administrator of a site to observe the teaching practices of their teachers (Haefele, 1980; Lower, 1987; Sweeney & Manatt, 1986). However, Jacob and Lefgren (2008) determined that principals' ability to identify the teachers with the highest and lowest achievement was relatively high, but they could not reliably differentiate between the teachers who had student achievement in the middle range. There is also evidence from extant evaluation data from multiple states that when teachers are scaled dichotomously or on an expanded range scale, 99% and 94%, respectively, were rated at the proficient level by principals (Weisberg, Sexton, Mulhern, & Keeling, 2009). Within the groups of researchers who argue for an observation of teacher practices, there is a tacit acknowledgement of this difficulty, with the argument that attention should be paid to providing methods for an "external" process for validating the principal observation portion of the evaluation (Darling-Hammond, Amrein-Beardsley, Haertel, & Rothstein, 2012; Medley, Coker, & Soar, 1984).

The advocates of inclusion of student achievement data in teacher evaluations rely on a foundational argument that the achievement of a teacher's students is the best method for demonstrating teacher outcomes (Goldhaber, 2002; Goldhaber & Brewer, 2000; Hanushek, Kain, O'Brien, & Rivkin, 2005; Wright, Horn, & Sanders, 1997; Harris & Sass, 2008). There is further evidence from research demonstrating that having a teacher identified at the highest level (e.g., 85th percentile) is associated with benefits similar to those seen from decreasing class sizes (Hanushek, Kain, O'Brien, & Rivkin, 2005; Rockoff, 2004). Others (Nye, Konstantopoulos, & Hedges, 2004; Sanders, 2000) have provided evidence that there is a measurable effect of good teaching as long as four years after students are in a high quality teacher's classroom. Sanders and Rivers (1996) have also demonstrated that a series of good teachers compared to a series of bad teachers have a large effect on students' long term outcomes. Opponents to value-added evaluation systems cite the findings of researchers who have documented a substantial variation in the findings of value-added measures of teacher effectiveness (Aaronson, Barrow, & Sander, 2007; Glazerman, Loeb, Goldhaber, Staiger, Raudenbush, & Whitehurst, 2010; Hanushek, Kain, O'Brien, & Rivkin, 2005; Rockoff, 2004). Dr. Edward Haertel, the Chair of the Board on Testing and Assessment for The National Academies, expressed the concerns of The National Academies to the federal government about the federally funded RTTT initiative placing such a high emphasis on student achievement without including other measures (Haertel, 2009).

Value Added Systems in Education

The overarching rationale for value-added analyses is quite simple. For example: Will the final value of a business be greater once an investment is applied to the business? Will the final score of an achievement test be greater once a teacher is applied to the student? Will the

overall yield of vegetable production be greater once a treatment is applied to the plant? Educational application of value-added approaches have come to prominence due in large part to a confluence of federal, state and local attempts to integrate student results into accountability systems. Value-added offers to disentangle the effects of teachers from student variables (Newton, Darling-Hammond, Haertel, & Thomas, 2010). Value-added systems vary in complexity from simple status models to complex regression models. Whereas much of the peer reviewed literature refers to value-added models in education as complex regression based models (Bock, Wolfe, & Fisher, 1996), the federal government accepts any type of examination of student gains as value-added (United States Department of Education [USDOE], 2006; 2008a, 2008b, 2009a, 2009b). In fact, federal funds have been awarded for various types of value-added systems which are currently being used to identify high performing teachers in school districts in Texas, Tennessee, North Carolina, Arkansas, Delaware, and Florida (USDOE, 2008b, 2009b).

Status and cohort gains models. The status model approach utilizes cross-sectional information consisting of achievement information for students who are in a school for a single year. The overall impact of a school on the students is assumed based on the estimate derived from a single year's data. In a status model approach, prior performance is not taken into account; instead status models look at the status of students who were enrolled in the prior year (Coleman, Campbell, & Kilgore, 1982). Gains models are change scores for groups of students in two adjacent years – for example, the third grade students in 2009 compared to the fourth grade students in 2010 (Lockwood, McCaffrey, Hamilton, Stecher, Le & Martinez, 2007). The state of Florida utilizes both status and gains in the state accountability system.

Value tables. Value tables or transition tables are similar to these simple gains models, in that they use pre-test and post-test data, however they are calculated at the individual student level (Doran & Izui, 2004; Hill, 2005, 2006a, 2006b; Hill, Marion, DePascale, Dunn, & Simpson, 2006). The value tables approach was utilized from the 2007-2008 school year through the 2009-2010 school year in Florida for the MAP program by multiple districts and by the state for the Charter School MAP program. The Florida value table system was similar to the value table and transition table systems utilized by the states of Delaware and Alaska (Taylor, 2008). The value tables in Florida were historically constructed using the five achievement levels from the state standardized assessment system (FCAT). For example, a point value is assigned to movement between pre-test and post-test achievement levels (Dougherty, 2007, 2008). These points are averaged for each of the subject area tests that teachers' students take while the teacher is responsible for the student.

Value tables are constructed by assigning a value to a change in achievement level from the Elementary and Secondary Education Act state test (FCAT in the case of Florida) through a process similar to that used in Delaware, Alaska and Florida (Taylor, 2008). A point value is assigned to the change between levels between a pre-test and a post-test (Dougherty, 2007, 2008). In order to mirror the methods used in Florida, and because documentation of the processes for assigning points and building a value table were not available in a published document, information was gathered from district personnel in a large Florida district that created and used value tables. This process is detailed below:

The value tables were created using the following guidelines: first, categories were set for each pre-test and post-test. This was accomplished in Reading & Mathematics courses using the five achievement levels of the FCAT which were derived based on input from a standard setting

panel using the modified Angoff method. The second step involved creating a cross tabulation of the pre-test levels and the post-test levels, with the post-test levels being listed in the columns and the pre-test levels being placed in the rows. The proportion of pre-test levels was then calculated for the number of students with changes in achievement in each cell. An example of cross tabulation of the pre-test levels and the post-test levels is below in Table 1 (see Appendix B for cross tabulation tables for all pre-test levels):

Table 1

Example of a Cross Tabulation of Pre-Test and Post-Test Levels

Pre-test		Post-test Level					
Level*	Students	1	2	3	4	5	Total
1	N	20	30	40	30	20	140
	Percentage	0.14	0.21	0.29	0.21	0.14	1.00
2	N	25	35	50	40	10	160
	Percentage	0.16	0.22	0.31	0.25	0.06	1.00

** Note: Cross tabulation of pre-test levels 3, 4, and 5 are not included in this example.*

The third and final step entailed assigning values to each cell based on the following rule-set (Michelle Watts personal communication, 2014):

1. The product of the proportion of cases in each cell and the value was summed to equal 100.
2. Students earned negative value points for going down in level unless they were at level 5.
3. Students earned no points for staying in level 1.
4. The points at each level should be approximately equal.
5. Students earned positive points if they move from level 5 to level 4.

When the district personnel were asked for an example of the final product resulting from this rule set, they provided the following, as shown in Table 2 (see Appendix B for detailed rule set tables).

Table 2

Example Rule Set Cross Tabulation of Pre-Test and Post-Test Levels

		Post-test Level					
Pre-test Level*		1	2	3	4	5	Total
1	N	20	30	40	30	20	140
	Proportion	0.14	0.21	0.29	0.21	0.14	1
	Raw Points	0	50	100	150	200	
	Value Points **	0	10.72	28.57	32.15	28.58	100***
2	N	25	35	50	40	10	160
	Proportion	0.16	0.22	0.31	0.25	0.06	1.0
	Raw Points	-55	70	120	170	220	
	Value Points **	-8.60	15.30	37.50	42.50	13.75	100***

** Note: Cross tabulation of pre-test levels 3, 4, and 5 are not included.*

*** Value Points is calculated by multiplying Proportion and Raw Points.*

*** Value Points Totals are rounded to the nearest whole number.

A teacher's value-added score is then calculated using the average points for each teacher's students' performance. For example (using Table 2), if a teacher has 30 students, 15 of which increased one level from pre-test 2 to post-test 3 (37.5 points each), eight of which increased two levels from pre-test 1 to post-test 3 (28.75 points each), and seven of which remained at level 1 (0 points each); the resulting score would be [(15*37.5) + (8*28.75) + (7*0)]/30 = 26.416.

Regression models. Regression models are more complex statistical procedures, such as covariate regression and multilevel modeling (Goldschmidt, Choi, Martinez - US Department of Education, 2004; Kingsbury, McCahon & McCall, 2004; Lyons, 2004; Doyle & Han, 2012).

The most recognized application of a regression based value-added model on student assessment results is the Tennessee Value-Added Assessment System (TVAAS). TVAAS was integrated into the educational reforms put in place in the Tennessee Educational Improvement Act of 1992. TVAAS was developed by a statistician who originally worked in agricultural statistics (Sanders, 1989) and applied complex regression techniques to assessment results to produce a measure of student and teacher effects using the extant test results. The expansion and integration of regression-type approaches has been greatly supported by the federally funded RTTT initiative. As the TVAAS system was integrated into accountability systems such as Houston Independent School District, the moniker of TVAAS has shifted to the Educational Value Added Assessment System (EVAAS). EVAAS is seen by many as the beginning of the integration of a value-added assessment method into state accountability systems across the nation and (Carey, 2004; Doran & Izumi, 2004; Hershberg, Simon, & Lea-Kruger, 2004; Kupermintz, 2003; Eckert & Dabrowski, 2010; Bianchi, 2003; McCall, Kingsbury, & Olson, 2004).

While EVAAS is arguably the most visible and widely used model in accountability systems, there have been multiple approaches from multiple vendors for integrating value-added approaches which use complex regression into accountability systems (McCaffrey, Lockwood, Koretz, Louis, & Hamilton, 2004). These vendors include the SAS EVAAS model, the American Institutes for Research (AIR), Mathematica, the National Center for the Improvement of Educational Assessment, and the Value Added Research Center (Goldhaber & Theobald, 2013; Ballou, Sanders, & Wright, 2004; Briggs & Weeks, 2009; Lockwood, Doran, & McCaffrey, 2003; Lockwood, McCaffrey, Hamilton, Stecher, Le, & Martinez, 2007; McCaffrey, Lockwood, Koretz, Louis, & Hamilton, 2004; Schmitz & Raymond, 2008; Wiley,

2006). While there are many models that could be examined this project specifically focused on the two models used in Florida. The value tables approach utilized from the 2007-2008 school year through the 2009-2010 school year in Florida for the MAP program by multiple districts and by the state for the Charter School MAP program. The American Institutes of Research (AIR) Covariate Adjustment model developed and adopted by committee in Florida for use in the 2010-2011 school year and forward.

The value-added scores used are derived by using the Florida value-added model - a covariate adjustment model (McCaffrey, Lockwood, Koretz, Louis, & Hamilton, 2004). The Florida model utilizes the current year's achievement as the outcome variable and includes multiple predictor variables, both dichotomous and continuous. The dichotomous indicators are ELL (in the program for less than two years), receipt of services for each possible disability classification (including gifted), number of Mathematics subject-relevant courses enrolled, and attendance. The continuous variables are two year prior achievement scores, mobility (number of transitions); difference from modal age in grade (months difference from the modal age for students enrolled in the same grade); class size (the number of students linked to a specific teacher); and homogeneity of entering test scores (interquartile range of student scores in the class).

The covariate regression model selected by the Florida Department of Education for the state of Florida (American Institutes for Research, 2011a, 2011b) to be employed in this study is

$$y_i = \mu + \sum_{g=1}^{M} \delta_g x_g + \sum_{j=1}^{K} \beta_j x_j + \theta_{(k)i} + \omega_{(mk)i} + \varepsilon_i;$$

where y_i represents the test score for student i, δ_g is the coefficient for the gth prior test score, β_j is the coefficient for covariate j, θ is the common school component of school k, ω is the effect

of teacher m in school k, and ε is the random error term. Since the teacher effect is the weighted mean of the student level residuals (ε_i), the individual student residuals are then aggregated for each teacher,

$$\tilde{\theta}_{\boldsymbol{t}} = \frac{N_j \sigma_t^{\mathbf{2}}}{N_j(\sigma_s^2 + \sigma_t^2) + \sigma_e^2} \frac{\sum_{i=1}^{N_j} \varepsilon_{(j)i}}{N_j}$$

where σ_t^2 is the teacher level variance, σ_s^2 is the school level variance, σ_e^2 is the residual variance, N_j is the number of students in class j, and ε (j)i is the residual for student i in class j. The final teacher effect is calculated based on the inclusion of a weighted mean for the school level,

$$\theta_t^* = \theta_t + .5\theta_{(s)t}$$

where θ_t^* is the estimate of the teacher effect, $\theta_{(s)t}$ is the estimate of the unique school component, and *s*(*t*) representing that teacher *t* in school *s*. The data elements utilized in the state of Florida's covariate regression model as provided by AIR are listed in Table 3.

Table 3
Predictor Variables Used in AIR Covariate Regression Model

English Language Learner Status (dichotomous)
Special Education Status (each dichotomous)
Language Impaired (D) Deaf or Hard of Hearing
Visually Impaired
Emotional/Behavioral Disability
Specific Learning Disability
Autism Spectrum Disorder
Traumatic Brain Injured
Other Health Impaired
Intellectual Disability
Gifted Student Indicator

Table 3 (Continued)
Predictor Variables Used in AIR Covariate Regression Model

Number of Mathematics courses enrolled in greater than one (dichotomous)
Enrolled in 2 or more Courses
Enrolled in 3 or more Courses
Enrolled in 4 or more Courses
Enrolled in 5 or more Courses
Homogeneity of class
Homogeneity of Class 1 Prior Year Test Scores
Homogeneity of Class 2 Prior Year Test Scores
Missing Homogeneity of Class 2 Prior Year Test Scores
Homogeneity of Class 3 Prior Year Test Scores
Missing Homogeneity of Class 3 Prior Year Test Scores
Homogeneity of Class 4 Prior Year Test Scores
Missing Homogeneity of Class 4 Prior Year Test Scores
Homogeneity of Class 5 Prior Year Test Scores
Missing Homogeneity of Class 5 Prior Year Test Scores
Homogeneity of Class 6 Prior Year Test Scores
Missing Homogeneity of Class 6 Prior Year Test Scores
Class size
Number of Students in Class 1
Number of Students in Class 2
Number of Students in Class 3
Number of Students in Class 4
Number of Students in Class 5
Number of Students in Class 6
Difference from Modal Age
Achievement: Two Years Prior
Achievement: Prior Year

Replication of the state's teacher value-added score would require the entire state's individual student achievement, course information, demographics, and attendance data. It is not practicable to reproduce the value. The state-derived and provided scores for individual teachers and the student level residuals was requested with teacher and student identifiers encoded in the same manner as the course file, which allowed the values to be appended to the data file.

While the decision to use value-added models has already been made in the state of Florida there are some real concerns about issues germane to education that researchers have attempted to examine, specifically sorting and poverty.

Sorting

Sorting in education refers to the distribution of students and teachers across schools and within schools. Teachers are not randomly assigned to schools, and students are not randomly assigned to teachers or to schools. There is a large body of literature that suggests that the sorting inherent to the public education system and teacher labor market is a biased process (Kalogrides, Loeb, & Beteille, 2013; Boyd, Lankford, Loeb, and Wyckoff, 2005a; Hanushek, Kain, & Rivkin, 2004; Lankford, Loeb, & Wyckoff, 2002).

This has led some researchers to examine the effect that sorting and non-random assignment has on the value-added modeling used in teacher evaluations. Ome (2013) demonstrated that in Columbia, South America, where teachers are restricted to what jobs they can apply based on proficiency tests, education, and experience, that teachers with higher scores and more seniority were in schools where students scored better on achievement tests. Betts, Rueben, and Danenberg (2000) examined California schools and found that schools with high poverty have more teachers with less experience, lower scores on the Praxis exams, and fewer advanced degrees. Bonesronning, Falch, and Strom (2005) examined data from Norwegian schools and found that teacher supply and demand was linked to the composition of the student body of a school. Lankford, Loeb, and Wyckoff (2002) examined New York Schools' urban schools and found that students who were low income and low achieving were often placed in classes with the least skilled teachers. Paufler and Amrein-Beardsley (2014) surveyed Arizona principals concerning student assignment to classes and found that many of the factors

identified for classroom assignment are not accounted for in typical value-added modeling, providing the conclusion that value-added is biased based on non-random assignment of students.

Paulfer and Amrein-Beardsley's position supports the work done by Rothstein (2010) who proposed and offered a model for testing scores derived from value-added methods. Rothstein used his model to examine the appropriateness of value-added scores for teachers and demonstrated some teachers seem to have a large effect on their students' previous year's achievement (during their initial year of interaction). His findings seem to demonstrate that the non-random assignment of students to teachers provided a situation in which a teacher's value-added score would be biased based on the sorting of students.

Goldhaber and Chaplin (2015), and Guarino, Reckase, and Wooldridge (2015), have both examined Rothstein's falsification test and the issue of sorting. Goldhaber and Chaplin's examinations found that Rothstein's test provides an accurate determination of whether there is sorting of students to a teacher; however upon further examination they also found that the falsification test provides improbable scores for randomly assigned students. Goldhaber has used simulations to demonstrate that Rothstein's approach will falsify VAMs that are not biased and also fails to falsify biased VAMS. Guarino and colleagues found that sorting could be demonstrated in large datasets at the building level, however when examined within a building it was much more difficult to demonstrate (2015). Kinsler (2012) also demonstrated that Rothstein's approach performed poorly with small samples. Sorting of teachers and assignment of students in a non-random manner is an issue in the analysis of educational data and is a real issue in relation to poverty and equity (Betts, Reuben, & Danenberg, 2000; Lankford, Loeb, & Wyckoff, 2002; Bonesronning, Falch, & Strom, 2005; Clotfelter, Ladd & Vigdor, 2006; Peske

& Haycock, 2006; Boyd, Lankford, & Wyckoff, 2007; Rothstein, 2010). This is an ongoing discussion in the literature and while there are clear indications that sorting exists it is not yet clear what the overall effect on the application of value-added modeling will be in education.

Poverty

A common concern from policy makers and stakeholders is that there may be a differential effect of school poverty on student achievement. This is a valid concern; in fact, there is a plethora of research that tells us that students in poverty have many barriers that they must overcome to be successful (Brooks-Gunn & Duncan, 1997; Janus, Walsh, Viverios, & Duku, 2003; Ferguson, Bovaird, & Mueller, 2007). Among early childhood indicators, higher poverty neighborhoods tend to have more students who are not ready for school (Janus, Walsh, Viverios, & Duku, 2003); children living in poverty have worse achievement outcomes, a higher incidence of learning disabilities, and are often developmentally delayed (Brooks-Gunn & Duncan, 1997). The home environments for children in poverty are much more likely to have chronic stressors and less likely to have the necessary social and emotional supports for success (Lacour & Tissington, 2011; Jensen, 2009). The research on sorting of educators (Ome, 2013; Betts, Reuben, & Danenberg, 2000; Lankford, Loeb, & Wyckoff, 2002; Bonesronning, Falch, & Strom, 2005; Clotfelter, Ladd, & Vigdor, 2006; Peske & Haycock, 2006; Boyd, Lankford, & Wyckoff, 2007), which has been conducted across institutions, states, and countries, often demonstrates that students in higher poverty schools tend to have teachers who are not as qualified as the teachers in lower poverty schools in terms of certification, experience, and Praxis performance (teacher qualification exams). To further compound this issue, existing research suggests the impact of a teacher influences a student's future achievement in a cumulative manner (e.g., Sanders & Rivers, 1996, Sanders, 2000; Nye, Konstantopoulos, &

Hedges, 2004; Aaronson, Barrow & Sanders, 2003; Rockoff, 2004; Rivkin, Hanushek, & Kain, 2005; Kane, Rockoff, & Staiger, 2006). All of these issues are very important in relation to the barriers and opportunities available to students, as well as the lasting impact of teacher quality on students. However, this study compared the differences between the State of Florida's adopted model and the value table approach to value-added modeling, so did not examine the possible influences of poverty.

Examining Value-Added Systems Applied to Education

The research community continues to examine the validity of value-added methodologies. No study has been able to definitively establish the causal relationship that policy makers assume. There has been an appropriate examination of the validity and reliability of value-added methodologies and there have been and continue to be multiple investigations into demonstrating and improving the precision of approaches, and searching for a means of implementing value-added in a fair and equitable manner that is both valid and reliable for all teachers. This section seeks to capture the zeitgeist of the research examining value-added systems in education; it is not exhaustive of all research on the comparisons of value-added systems in education.

There are many studies that examine different approaches to value-added modeling. These studies include examinations of reliability (stability, bias, or sensitivity) and validity. For the lay practitioner these studies often provide valuable insight into the application of value-added methodologies to extant data and contribute to the understanding of educational practitioners. Further, many of these same studies have embedded examinations of stability or reliability and provide information concerning the ability of value-added methodologies to provide reliable data. The examination of methodological approaches provides information

about convergence between methods and consistent demonstration of a real difference that is unique to the outcome level at which the data are being examined (which in most cases for teacher evaluation is at the classroom or teacher level). Examination of the available methodological approaches for both reliability and validity is extremely important given the high stakes that the results for value-added methodologies have taken in teacher evaluations across the nation. Further, because of the intertwined nature of reliability and validity, researchers have had the opportunity to contribute information that demonstrates both.

Application. The ability of value-added methods to provide unique teacher level information across groups of students has been examined in multiple ways and are often referenced by the lay-person in relation to the appropriateness of the methodology for education. For example, Sanders and Rivers (1996) demonstrated that teacher effects have some persistence and accumulate over time. Sanders and Rivers found that when students were taught by the least effective teachers for three years the students' scores were consistently below similar students taught by the most effective teachers. The research was duplicated by Mendro, Jordan, Gomez, Anderson, and Bembry (1998) and Kain (1998) using data from the Dallas Independent School District with consistent results. Another example of a study that examined the unique teacher level information was conducted by Rowan, Correnti, and Miller (2002) examined two cohorts of students from a nationwide sample of schools and demonstrated classroom level results in reading and mathematics scores that accounted for the variability in growth in student achievement scores.

Reliability. While Sanders and Rivers's findings of the persistence of teacher effectiveness as defined by the teacher residuals from their value-added methodology have been demonstrated to be consistent (Mendro, Jordan, Gomez, Anderson, & Bembry, 1998; Kain,

1998; Rowan, Correnti, & Miller, 2002), it is a direct example of the reliability of the methods. Consistency across methodologies has also been demonstrated by Tekwe, Carter, Ma, Algina, Lucas, Roth, Ariet, Fisher, and Resnick (2004), who examined the differences between three different types of value-added models and found that the simplest model (simple Fixed Effects Model) had similar results to the more complex Hierarchical Linear Models and Layered Mixed Effects Models, with some differences when they controlled for minority and socio-economic status. Conversely, Goldhaber, and Theobald (2012) found high correlations between models that account for student background and those which do not, provided that each include multiple measures for prior student achievement. When examining the consistency of the teacher effects as teachers move across contexts, Sanders, Wright, Springer, and Langevin (2008) observed stability across disparate student populations. Others examined "inter-temporal stability of teacher effects" for teachers across multiple years and found consistent results at the teacher level (Lockwood, McCaffrey, & Sass, 2008) and moderate relationships when aggregated at the school level (Sass, 2008). Koedel and Betts (2007) found teachers in the tails of the distribution demonstrate somewhat higher stability. Further, Lockwood and McCaffrey (2008) examined the impact of heterogeneity of students and found that the teacher effect varied only a small amount (3-4%), and the overall impact of heterogeneity of students does not have an appreciable impact.

Validity. Lockwood, McCaffrey, Hamilton, Stecher, Le, and Martinez (2007) used a large longitudinal dataset to examine the estimated teacher impact using three Stanford 9 scores (total mathematics, and Procedures and Problem Solving subscores) for four value-added methodological approaches. They also varied the types of student level controls that were used in each of the models. When they examined the results of each of the models for the specific test they found that the teacher effect results were highly correlated, yet demonstrated specificity of

the teacher results based on the focus of the test. They found that when examining the differences across types of test there was not a high correlation for the teacher impact for the different content of the achievement test, demonstrating that value-added methodology can tie an impact to the specific type of skill measured by the achievement test (Lockwood, McCaffrey, Hamilton, Stecher, Le, & Martinez, 2007). This is convergent with other researchers' identification that the content (Hamilton, 2004), and structure (Martineau, 2005 & 2006; Briggs, Weeks, & Wiley, 2008; Briggs & Weeks, 2009; Schmidt, Houang, & McKnight, 2005) of the assessment may have an impact on the resulting value-added estimates.

Experimental studies that randomized students' assignment across classes found similarities between the value-added results for randomly and non-randomly assigned students (Nye, Konstantopoulos, & Hedges, 2004; Kane & Staiger, 2008; Kane, McCaffrey, Miller, & Staiger, 2013). In an attempt to examine differential effects of teachers based on the consistency of the achievement of their students, Koedel and Betts (2005) found some evidence of an interaction with value-added results for teachers when examining groups of students with prior test scores above and below the median. Hanushek, Kain, O'Brien, and Rivkin (2005) also found that gains for students are related to their prior achievement. However, these findings are also convergent with the findings that students demonstrate positive impacts from high achievement in prior years. When examining fixed and random effect models, researchers have found that fixed effects are sensitive to sampling error with a small number of data points for individual teachers and random effect using shrinkage has an impact on the teachers at the extremes of the distribution (Sanders, Saxton, & Horn, 1997; Ballou, Sanders, & Wright, 2003; Rowan, Correnti, & Miller, 2002).

While there have been some arguments that external factors outside of the teachers' control, such as the poverty level of their students, have an impact on teachers' estimates, Chetty, Friedman, and Rockoff (2014a, 2014b, 2014c) combined IRS tax data with student achievement data for a large urban school district and found that the families socio-economic status as derived from the tax elements are not correlated with the teachers' estimates. Bacher-Hicks, Kane, and Staiger (2014) duplicated Chetty et al.'s approach and found consistent results using data from Los Angeles. Both researchers also examined the students as they moved through school in relation to staffing changes and found that while there was some bias associated with staffing changes, the amount of bias was relatively small (2.6%). There appears to be a level of robustness across models that implies many of the assumptions can be violated and similar results can be derived (Sass, Semykina, & Harris, 2014). Further, as more achievement tests are constructed and used across schools and districts, the increase in student data will improve the predictive value of value-added approaches (Goldhaber & Hansen, 2013) and decrease the standard error as sample sizes increase (McCaffrey, Sass, Lockwood, & Mihaly, 2009).

The findings of researchers generally are consistent with Lockwood, McCaffrey, and Sass (2008) assertion that changing model specifications minimally affects the stability of teacher effect. While this may seem reassuring in relation to the reliability and validity of value-added methodological approaches, it is not and should not be used as a means of dismissing the concerns surrounding the impact of implementation on the individual teachers and schools; while the results are stable there are individuals who can be impacted adversely and inappropriately. There is ample demonstration of the stability and consistency of the operation of value-added models and consistent direction from statisticians on the appropriate application of the models, as well as published opinions on the appropriate and inappropriate uses. Some may argue that,

when considering the application of value-added estimates as a performance metric, there is similar stability to those found for salespeople, securities analysts, sewing-machine operators and baseball players (Glazerman, Loeb, Goldhaber, Staiger, Raudenbush, & Whitehurst, 2010; McCaffrey, Sass, Lockwood, & Mihaly, 2009).

The policy requirements that have driven the examination of extant data and longitudinal databases have provided a great deal of information and important conversations linked to poverty, sorting, and equity in education. It is extremely important and useful at both the national and local levels to identify methods for partialing out the attributable impact that a teacher has on an individual student given the high stakes that the results for value-added methodologies have taken in teacher evaluation. The same can be said concerning the reliability and validity of value-added methodologies and providing information about convergence between methods. The more accurate the methodology, the better the information and application of the results to the betterment of understanding and improving education.

Even though we do see some differences between methodological approaches for measuring teacher impact on students, the implications of collapsing the results of the disparate methodological approaches into four categories may lend to the minimization of differences that may be seen in the results of the full models. The State of Florida requires that teachers be classified into four categories based on their value-added scores. The results of this study may assist districts with an alternative means to make decisions based on data that is readily available to them which teachers will most likely fall within the four evaluation categories as defined by the FLDOE. While it is very important to understand the differences in the models, the policy requirements have rendered much of the discussion and concerns to a hypothetical arena that

have no practical application in the use of student data from state standardized assessments and teacher impact on student academic performance.

An application of *lex parsimoniae* (law of parsimony) would tell us that if the methodological approaches provide the same classifications for teachers it is not appropriate to use the more complex solution. Methodologists who are working with educational institutions and policy makers focus on the most accurate and parsimonious means of partialing out a teacher's impact on students without considering the larger context or the reality of how the process will be applied.

Systems in education that are used to evaluate teachers are often linked directly to or have an implied basal connection to student academic achievement. The application and codification of a causal relationship in the policy treatment for teacher evaluation rather than a carefully constructed equitable approach to value-added methodologies creates an environment in which the practicable use of a simple methodological approach (value tables) has the added advantages of being replicable by educational stakeholders and more understandable to parents and the community. The primary goal of this research is to investigate the consistency of teacher value-added ratings and evaluation classifications using two value-added procedures. An examination of the published literature did not reveal an instance where there has been a published example of comparing value tables and covariate regression methods nor a comparison of the application of the methods to the requirements of the State of Florida. The following chapter discusses the methodologies used to evaluate the concordance and discordance of teacher ranking and category assignment derived from the value table approach versus the covariate regression approach currently used by the State of Florida.

Chapter 3

Methods

Purpose

The primary goal of the research was to investigate the consistency of teacher value-added ratings and evaluation classifications using two value-added procedures: value tables and the covariate adjustment model currently used by the state of Florida.

First, the degree to which there are differences between the rankings of mathematics teachers using the two methods was examined. Second, the degree to which there are differences between the assignments of mathematics teachers into the four state-mandated teacher evaluation categories using the two methods was investigated.

Research Questions

1. What is the degree of concordance and discordance between the mathematics teachers' ranking using value-added scores derived from the value table approach versus the covariate regression approach?
2. What is the degree of concordance and discordance of the categories to which mathematics teachers are assigned when the state's recommendations for the classification of teachers into the four evaluation categories are applied to their value-added scores by the value table approach versus the covariate regression approach?

Sample

The sample was from a Florida school district's data files. The data files did not include any direct personal identifiable information for students or teachers. Each student and teacher was assigned an encoded number that was consistent across files. Teachers included in the analyses had at least one rostered class of students in mathematics. The students included in the analyses were 5th through 9th graders with at least three years' worth of FCAT mathematics scores from 2009-2010 through 2011-2012. For the value table analysis, a student must have two years of FCAT data from the 2011-2012 and 2010-2011 school years in order to derive a student value-added score. For the covariate regression, students who have an FCAT mathematics score from 2011-2012 and the prior two years of achievement data (2009-2010 and 2010-2011) provided a residual score for each student. In order to maintain parity between the two methods, only students with data for the most recent two school years (2010-2011 and 2011-2012) and a residual score (2011-2012 residual score) were used to derive the teacher level results. This approach ensured the students included in each teacher's associated value-added score were the same, with the intention of the comparisons being based on the same sample.

Procedures

The study was implemented in five stages:

1. data acquisition and preparation,
2. verification of consistent students across models,
3. value table generation,
4. covariate regression aggregation, and
5. data analysis.

Stage 1 data acquisition and preparation. Data acquisition and preparation had two steps: 1.obtain data, 2. prepare data files.

Step 1: obtain data. The data requested was student and teacher data that were anonymized and linked through the use of an encoded student number and an encoded teacher number. Student, teacher, and course level data were requested for the 2011-2012 school year. Testing data were requested for the 2010-2011, and 2011-2012 school years. The student level data requested included demographics (grade of enrollment), mathematics achievement (FCAT MDSS for 2010-2011, and 2011-2012), student level residuals (for 2011-2012), mathematics course enrollment (state course code, school site (encoded), encoded teacher number), and attendance (total days enrolled and number of schools enrolled).

Data files and elements. Table 4 identifies the data file elements requested:

Table 4
Data File Elements

Data File Name	Element
Demo File (2011-12)	Encoded student ID Student grade Student residual
Course File (2011-12)	State Math course number (7 digit) & class period School of instruction Encoded teacher, student ID & school number
FCAT Mathematics (2010-11, 2011-12)	MDSS
Attendance file	Encoded student ID Encoded School number Days enrolled

Step 2: prepare data files. Prior to conducting the analyses, the following data management tasks were planned. The files would be merged based by individual encoded student id. Specifically, the demographics file would be merged with the Mathematics achievement file and the resulting file would then be merged with the course files. The final file

would then include all of the necessary information to derive a student level residual for the covariate adjustment model and to derive the value table score for each student.

In order to maintain consistency with the accountability systems of the state, Based on the information for each student aggregated at the school level from the attendance file, students were to be selected only if they were enrolled for at least 80% of the 112 days (90 days) of the school year between the first state reporting period (October) and the administration of the FCAT in April. However, Data were provided for each grade level in a single already matched Statistical Package for Social Science (SPSS 22) data file. The data provided met the needs of the analysis and selection process. Therefore the planned preparation of the data files was not necessary. The Variable names and a short description of the variable from each of the data files are provided in Table 5.

Table 5
Variable Name and Description of Each Grade Level File

Variable Name	Variable description
teacher_encode	Encoded teacher variable
student_Encode	Encoded student variable
@_1011_TestedGrade	Student's Grade for 1011 test
@_1011_ScaleScore	Student's Scale score for 1011 test
@_1112_ScaleScore	Student's Scale score for 1112 test
@_1112_StudentEnrollment	Student's Enrollment for 1112
@_1112_Predictedscore	Student's Predicted score for students 1112 scale score
Resid	Student's Residual (@_1112_Predictedscore - @_1112_ScaleScore)

Stage 2 data verification. Verification of consistent students across models had three steps:

1. Verification of two years of achievement data and existence of a residual for each student linked to a teacher.

2. Flag any student from the sample who does not have both a residual for 2011-2012 school year and achievement data for school years 2010-2011 and 2011-2012).
3. Flag any student who did not have 90 days or greater at an individual school site.

Stage 3 value table generation. Value table generation had three steps:

1. Construct value table,
2. Generate value-added scores for teachers based on students from their classes, and,
3. Sort the teacher scores into state categories using the procedure provided by the FLDOE for identifying cut scores based on the distribution of teacher value-added scores.

Value table generation. The value tables were constructed by assigning a value to a change in achievement level from the FCAT. A point value was assigned to the change between levels for a pre-test and a post-test (Dougherty, 2007, 2008). In order to emulate the methods used in Florida, the value table model mirrored the process used by HCPS, as provided by district personnel. The Value Tables were generated using the following steps:

1. Categories were set for each pre-test and post-test based on published conversions (FLDOE , 2014) with level 1 separated into a low and high scoring category consistent with extant information (Lassila, 2006) and confirmation from feedback from retired district personnel in a large Florida district (Michelle Watts personal communication, 2014).
2. Created a cross tabulation of the pre-test levels and the post-test levels, with the post-test levels being listed in the columns and the pre-test levels being placed in the rows;

3. Calculated the proportion of pre-test levels for the number of students with changes in achievement in each cell;
4. Assign a value to each cell based on the rules provided during the interview with the retired employee of a large Florida district and the example table provided in a presentation by Lassila (2006) at the November 2006 Florida Association of School Personnel Administrators conference. Values were assigned in consistent point values across each cells for each pretest level using an excel spreadsheet to concurrently evaluate number of points assigned in each cell.

The fourth step in the creation of the value tables was the assignment of values to each cell based on the following rule-set (Michelle Watts personal communication, 2014):

1. The product of the proportion of cases in each cell and the value was summed to equal 100.
2. Students earned negative value points for going down in level unless they were at level 5.
3. Students earned no points for staying in level 1.
4. The points at each level should be approximately equal.
5. Students earned positive points if they move from level 5 to level 4.

Based on the resulting value table each student was assigned a value table score based on their pretest and posttest scores (value table computations for each grade level are detailed in Appendix B.). The teacher's value table score was then calculated using the average points for all of their assigned students who were not eliminated from the analysis for missing data or attendance. Teachers where then assigned a state category using the procedure provided by the

FLDOE for identifying cut scores based on the distribution of teacher value-added scores (Copa, 2012).

Stage 4 covariate regression aggregation. Covariate regression aggregation had two steps: 1. aggregate the student's, who met the inclusion criteria, residuals for their teachers, and 2. classify the teacher scores into state categories using the procedure provided by the FLDOE for identifying cut scores based on the distribution of teacher value-added scores. The study used the four classification categories mandated in the state evaluation system (Florida Statute 1012.34) combined with the state's guidance on the dispersion of the teacher level aggregated scores for classification of teachers into those categories (Copa, 2012).

Stage 5 data analysis. Finally, the fourth stage of the process was the analysis stage and had two steps: 1.examine for concordance/discordance the resulting rankings from the two procedures and, 2. examine for concordance/discordance the classification of teachers into categories

The aggregated data analysis file contained a teacher score for both the value table and regression models and did not include any encoded student information. This file contained the information necessary to conduct the comparisons for the project. The data analysis stage followed, with each step in the analysis stage directly linked to the research questions:

Table 6
Data Analysis Steps

	Analysis Step	Research Questions
1	Examine for concordance/ discordance the resulting rankings from the two procedures	What is the degree of concordance and discordance between the mathematics teachers' ranking using value-added scores derived from the value table approach versus the covariate regression approach?
2	Examine for concordance/ discordance the classification of teachers into categories	What is the degree of concordance and discordance of the categories to which mathematics teachers are assigned when the state's recommendations for the classification of teachers into the four evaluation categories are applied to their value-added scores by the value table approach versus the covariate regression approach?

The first step of the analysis was to examine for concordance/discordance the resulting rankings from the two procedures. This consisted of a simple comparison of the ranking of individual teachers based on their value-added scores derived from each approach. This comparison was accomplished by conducting three analyses: the distribution characteristics of each method were calculated and examined (frequency distribution, skew and kurtosis), a Pearson Product Moment Correlation was generated for the two sets of value-added scores to examine the consistency between scores, and a raw rank difference was calculated with quintiles examined for teachers who were classified differently between methodologies.

The second step of the data analysis was to examine for concordance/discordance the sorting of teachers into categories. This was accomplished by examining of the differences in categorical assignment between approaches. This examination was conducted using a matrix which identified the consistent classification and the degree of difference between classifications of the two methods. The classifications were also used in calculating a Kendall tau and a Goodman and Kruskal gamma for each grade level. These measures where used based on the

simulation for Doubly Ordered Square Contingency Tables which was conducted by Göktaş and İşçi (2011).

Data Management

The project required student level data. In order to ensure compliance with student data privacy laws, data was requested in an anonymized form with student and teacher numbers encoded to allow for matching without providing any personally identifiable information for either the teacher or the student. For security purposes, all electronic files were encrypted and secured on an external hard drive. When the data were not being used, the hard drive was stored in a locked cabinet to which only the researcher had access. The computer used to conduct the analysis was not connected to the internet during analysis, was password protected, and locked after 15 minutes of inactivity. Only the researcher had the password for the encrypted files and to the computer. Upon completion of the project all individual student-level data received was destroyed and the external drive was reformatted by using a disk utility to write over the entire drive.

Chapter 4

Results

Study Sample

The data gathering and analysis plan (Appendix A) was followed for the treatment and conduct of the analyses. The initial study sample consisted of students in grades 4 through 8 who had any value added results from the FLDOE files. In order to ensure comparable groups and consistency between the types of analysis, students were removed from the data set if two consecutive grade levels of test data for school years 2010-11 and 2011-12 were not available. For each grade level a crosstab was generated based on the total grade level population who had taken both the post-test in the current grade (2011-12) and pre-test the prior year (2010-11) (Tables 7- 11). These crosstabs were used in the generation of the value table scores.

Table 7
Grade 4 Cross Tabulation of Pre-Test (2010-11) by Post-Test (2011-12)

		Post-Test (2011-12 Mathematics FCAT)					
		Low Level 1	High Level 1	Level 2	Level 3	Level 4	Level 5
Pre-test 2010-2011 Mathematics FCAT	Low Level 1	228	140	20	2	0	0
	High Level 1	218	961	418	90	3	0
	Level 2	30	698	1415	812	131	12
	Level 3	5	126	902	1761	901	135
	Level 4	0	1	95	684	912	451
	Level 5	0	1	5	131	512	687

Table 8

Grade 5 Cross Tabulation of Pre-Test (2010-11) by Post-Test (2011-12)

		Post-Test (2011-12 Mathematics FCAT)					
		Low Level 1	High Level 1	Level 2	Level 3	Level 4	Level 5
Pre-test 2010-2011 Mathematics FCAT	Low Level 1	291	200	16	1	0	0
	High Level 1	260	1034	643	100	2	0
	Level 2	33	525	1434	816	102	4
	Level 3	3	83	834	1760	782	142
	Level 4	0	2	85	601	1084	565
	Level 5	0	0	3	51	348	786

Table 9

Grade 6 Cross Tabulation of Pre-Test (2010-11) by Post-Test (2011-12)

		Post-Test (2011-12 Mathematics FCAT)					
		Low Level 1	High Level 1	Level 2	Level 3	Level 4	Level 5
Pre-test 2010-2011 Mathematics FCAT	Low Level 1	270	200	27	1	0	0
	High Level 1	349	987	493	44	2	0
	Level 2	99	801	1568	583	40	2
	Level 3	11	109	801	1646	576	31
	Level 4	1	3	71	593	1207	348
	Level 5	0	0	4	49	424	764

Table 10

Grade 7 Cross Tabulation of Pre-Test (2010-11) by Post-Test (2011-12)

		Post-Test (2011-12 Mathematics FCAT)					
		Low Level 1	High Level 1	Level 2	Level 3	Level 4	Level 5
Pre-test 2010-2011 Mathematics FCAT	Low Level 1	262	286	48	8	2	0
	High Level 1	225	859	602	100	6	0
	Level 2	59	522	1291	686	60	2
	Level 3	4	56	591	1612	599	45
	Level 4	0	6	29	579	1327	521
	Level 5	0	0	4	29	399	952

Table 11
Grade 8 Cross Tabulation of Pre-Test (2010-11) by Post-Test (2011-12)

		Post-Test (2011-12 Mathematics FCAT)					
		Low Level 1	High Level 1	Level 2	Level 3	Level 4	Level 5
Pre-test 2010-2011 Mathematics FCAT	Low Level 1	219	214	26	6	0	0
	High Level 1	243	900	340	65	3	0
	Level 2	52	648	1097	500	24	1
	Level 3	5	141	828	1703	321	26
	Level 4	0	7	71	829	1062	377
	Level 5	1	0	0	74	447	985

The cross tabulations seen in tables 7 through table 11 reflect the movement of students from the pre-test achievement level to the post-test achievement level. The movement of students across every grade from their pre-test score was centered on the corresponding post-test score with most movement occurring into the next highest or lowest level. Very few students moved more than a single level from pre-test to post-test. This may be interpreted that the test levels are relatively consistent across time for students.

Value tables were generated using the proportions from these tables. After the students with an attendance rate of less than 0.8 were removed from the data set, the sample data set was further restricted to teachers who had more than ten students in the remaining student dataset assigned to them (Lassila, 2006). The remaining students made up the data set for this study and were assigned a value from the value tables generated earlier. Teacher level value added scores were aggregated by averaging the residuals for the students assigned to them and the value table score associated with their students, into separate variables. The distribution of students and final number of teachers is provided in Table 12.

Table 12
Distribution of Students and teachers across teachers by grade to identify study sample

Grade	Original *N*	2 years test data*	Attendance rate of >= 0.8	>10 students per teacher**	Number of teachers**
4th grade	13045	12487	12013	11103	526
5th grade	13665	12590	12127	11066	509
6th grade	13209	12104	11438	11063	213
7th grade	13039	11771	11105	10620	210
8th grade	12485	11215	10528	10094	177

*Utilized to create crosstabs of pretest posttest for value table creation
** Final sample size utilized in study reflected only teachers and their assigned students where there are more than 10 students assigned to an individual teacher.

The teachers in 4th and 5th grade have on average 21 students associated with them. The 6th through 8th grade teachers have on average 53 students associated with them.

Value Tables

In order to replicate the methods recommended by the FLDOE, and because specific documentation of the processes for assigning values to each change in achievement level and building a value table were not found in published documents, information was gathered from two sources: extant data from the FLDOE website (Lassila, 2006) and several discussions/interviews with a retired district employee in a large Florida district that created and used value tables extensively in the MAP program (Michelle Watts personal communication, 2014). State achievement levels were consistent with the published conversions used by FLDOE (2014). Achievement level 1 was separated into low and high sub-levels that were also consistent with extant information (Lassila, 2006) and confirmation from interviews with retired district personnel in a large Florida district. Table 13 shows the FCAT mathematics developmental scale score for each grade level.

Table 13
*FCAT mathematics developmental scale score Ranges for Grade Levels**

FCAT 2.0 Mathematics Developmental Scale Scores (140 to 298)						
Grade	Low Level 1	High Level 1	Level 2	Level 3	Level 4	Level 5
3	140-161	162-182	183-197	198-213	214-228	229-260
4	155-175	176-196	197-209	210-223	224-239	240-271
5	163-183	184-204	205-219	220-233	234-246	247-279
6	170-191	192-212	213-226	227-239	240-252	253-284
7	179-199	200-219	220-233	234-247	248-260	261-292
8	187-207	208-228	229-240	241-255	256-267	268-298

* these ranges are applicable to the 2010-11 and 2011-12 achievement information

The process for developing value tables in enumerated in chapter 3 and the tables demonstrating the calculations for arriving as the values to populate each grade level value table is found in Appendix B. The final values for the value tables were placed into a table for each grade (Tables 14 through 18).

Table 14
Fourth Grade Value Tables for Pre-test (2010-11) and Post-test (2011-12) Student Achievement Levels

Pre-Test Level	Post-test Low Level 1	Post-test High Level 1	Post-test Level 2	Post-test Level 3	Post-test Level 4	Post-test Level 5
Low Level 1	0	210	420	629	839	1049
High Level 1	-50	72	194	316	438	560
Level 2	-100	-5	89	184	279	373
Level 3	-150	-67	17	100	183	267
Level 4	-200	-121	-43	36	114	200
Level 5	-250	-171	-91	-12	68	147

Table 15
Fifth Grade Value Tables for Pre-test (2010-11) and Post-test (2011-12) Student Achievement Levels

Pre-Test Level	Post-test Low Level 1	Post-test High Level 1	Post-test Level 2	Post-test Level 3	Post-test Level 4	Post-test Level 5
Low Level 1	0	217	434	651	868	1085
High Level 1	-50	66	183	299	416	532
Level 2	-100	-7	86	179	272	365
Level 3	-150	-67	16	99	182	265
Level 4	-200	-125	-49	26	101	200
Level 5	-250	-174	-98	-22	53	129

Table 16

Sixth Grade Value Tables for Pre-test (2010-11) and Post-test (2011-12) Student Achievement Levels

Pre-Test Level	Post-test Low Level 1	Post-test High Level 1	Post-test Level 2	Post-test Level 3	Post-test Level 4	Post-test Level 5
Low Level 1	0	194	388	581	775	969
High Level 1	-50	83	216	349	482	616
Level 2	-100	6	111	217	323	428
Level 3	-150	-63	24	111	198	286
Level 4	-200	-122	-44	35	113	200
Level 5	-250	-173	-97	-20	56	133

Table 17

Seventh Grade Value Tables for Pre-test (2010-11) and Post-test (2011-12) Student Achievement Levels

Pre-Test Level	Post-test Low Level 1	Post-test High Level 1	Post-test Level 2	Post-test Level 3	Post-test Level 4	Post-test Level 5
Low Level 1	0	146	293	439	586	732
High Level 1	-50	63	175	288	400	513
Level 2	-100	-3	94	190	287	384
Level 3	-150	-66	17	101	184	268
Level 4	-200	-125	-51	24	98	200
Level 5	-250	-175	-100	-25	50	125

Table 18

Eighth Grade Value Tables for Pre-test (2010-11) and Post-test (2011-12) Student Achievement Levels

Pre-Test Level	Post-test Low Level 1	Post-test High Level 1	Post-test Level 2	Post-test Level 3	Post-test Level 4	Post-test Level 5
Low Level 1	0	164	328	491	655	819
High Level 1	-50	80	210	341	471	601
Level 2	-100	5	109	214	318	423
Level 3	-150	-59	32	123	214	304
Level 4	-200	-120	-39	41	121	200
Level 5	-250	-174	-98	-22	54	130

The value tables are generally neutral with a beginning value of zero or negative and a consistent value being assigned across each post-test level with the least number of points being awarded to students who decrease in level and the greatest number of points being awarded to students moving up levels. Each student received a value table score based on their pretest to

posttest performance in their associated grade level. The individual student value table scores were then averaged for each teacher for all students in their class that met the inclusion criteria, providing a teacher level value table value added score.

Covariate Residuals

Teachers covariate value added scores were generated based on the teachers individual student's difference score based on their predicted scale score on the 2011-12 Mathematics FCAT and their actual Mathematics FCAT scale score. The average residual was calculated for each teacher providing a teacher level covariate value added score.

Analysis

Research question one. What is the degree of concordance and discordance between the mathematics teachers' ranking using value-added scores derived from the value table approach versus the covariate regression approach?

This research question was evaluated in steps. First the distribution characteristics of the teacher covariate value added scores and value table value added scores was examined. Second a scatterplot and Pearson Product Moment Correlation between the two value-added scores was generated. Third the relationship between the relative locations each position had in the distribution using a Tukey Mean Difference Plot of the value added scores. Finally, each value added score was divided into quintiles and the accuracy, agreement and disagreement between the quintiles was examined

Distributions. The distribution characteristics of each method were then examined. A 95% confidence interval was generated for the skew and kurtosis and it was observed that for each of the grades the distribution measures approached normal.

Table 19
Distribution for Teacher covariate Value Added Scores for Research Question 1

Grade	*N*	Mean	*SD*	Skew	95% Confidence Interval	Kurtosis	95% Confidence Interval
4th	526	-0.64	5.81	0.14	-0.068 to 0.348	0.61	0.197 to 1.031
5th	509	0.00	4.84	-0.03	-0.238 to 0.186	-0.36	-0.785 to 0.061
6th	213	-2.59	3.82	-0.03	-0.360 to 0.294	0.37	-0.282 to 1.020
7th	210	-0.48	3.50	0.11	-0.218 to 0.440	0.14	-0.516 to 0.794
8th	177	-0.99	3.41	0.64	0.281 to 0.999	1.10	0.391 to 1.813

Table 20
Distribution for Teacher Value Table Value Added Scores for Research Question 1

Grade	*N*	Mean	*SD*	Skew	95% Confidence Interval	Kurtosis	95% Confidence Interval
4th	526	101.72	35.12	0.020	-0.193 to 0.225	-0.07	-0.483 to 0.350
5th	509	101.08	31.72	-0.160	-0.375 to 0.049	-0.21	-0.637 to 0.210
6th	213	94.41	30.53	-0.012	-0.339 to 0.315	1.61	0.961 to 2.263
7th	210	97.77	26.02	-0.088	-0.417 to 0.241	0.30	-0.355 to 0.955
8th	177	99.51	26.19	-0.216	-0.575 to 0.143	-0.44	-1.155 to 0.267

Scatter Plots and Correlations. The scatter plots for each grade were generated and examined (figure 1-5).

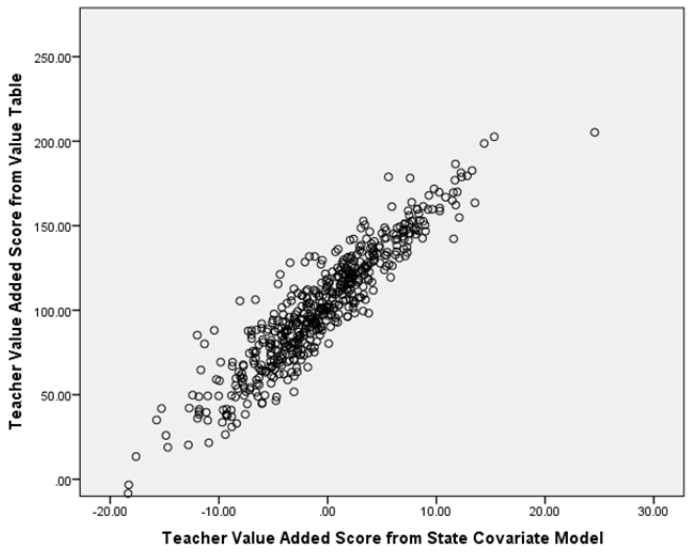

Figure 1 *Grade 4 Scatterplot of Teachers Value Table Score by Covariate Model Score*

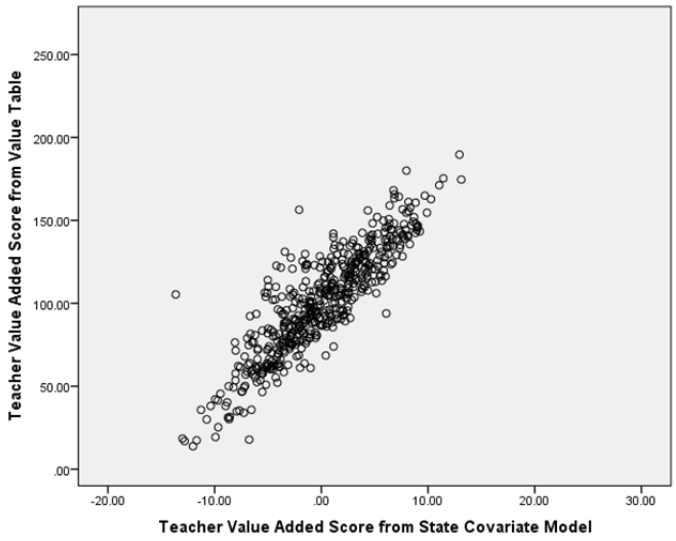

Figure 2 *Grade 5 Scatterplot of Teachers Value Table Score by Covariate Model Score*

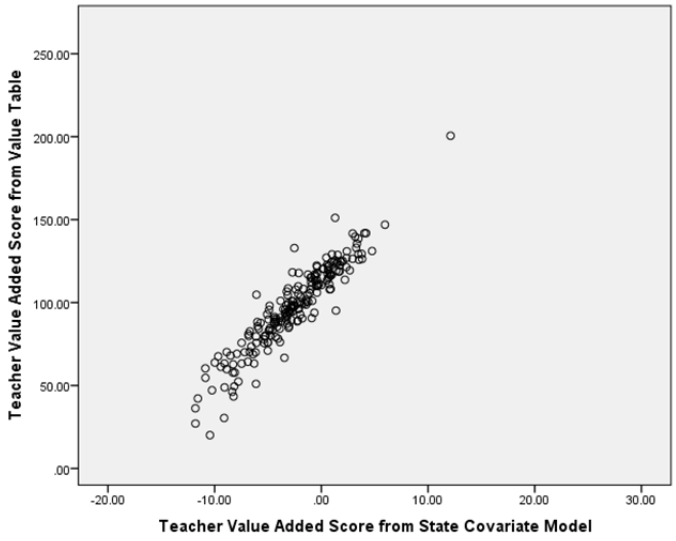

Figure 3 *Grade 6 Scatterplot of Teachers Value Table Score by Covariate Model Score*

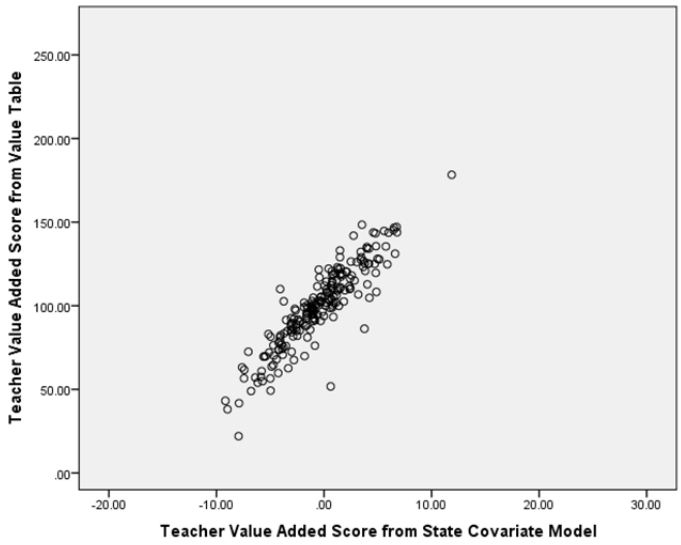

Figure 4 *Grade 7 Scatterplot of Teachers Value Table Score by Covariate Model Score*

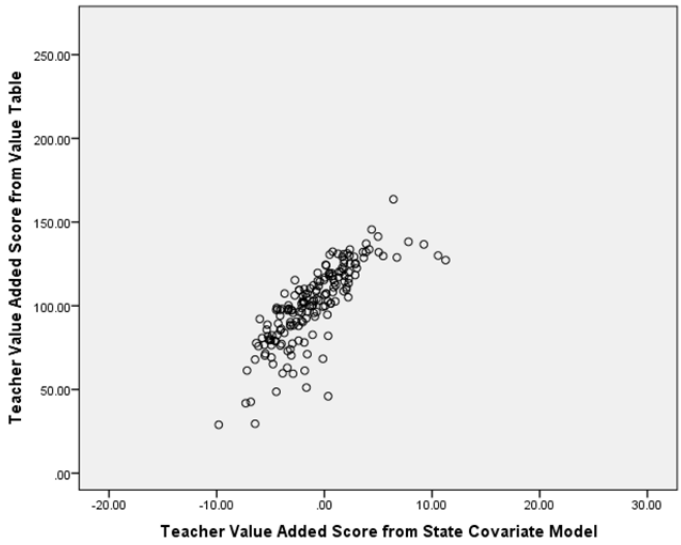

Figure 5 *Grade 8 Scatterplot of Teachers Value Table Score by Covariate Model Score*

The two sets of value added scores for the teachers were then examined using a Pearson Product Moment Correlation as shown in Table 19. The correlation coefficients were all significant, positive, and ranged from .981 to .772.

Table 21

Pearson Product Moment Correlation for Research Question 1

Grade Level	r	p	n
Grade 4	.926	<.0001*	526
Grade 5	.880	<.0001*	509
Grade 6	.981	<.0001*	213
Grade 7	.907	<.0001*	210
Grade 8	.772	<.0001*	177

**denotes significant results at the .05 level*

Tukey mean-difference (Bland-Altman diagram). In order to examine the agreement between the distribution of the scores for each teacher's value table score was converted to a z-score and then transformed into the same scale as the covariate regression score using the mean and standard deviation of the covariate regression score. The scores where then plotted using a Tukey mean-difference plot or a Bland-Altman diagram (1986). The resulting graphic provides a visual representation of the scores in a manner that demonstrates a relatively high level of agreement between the Florida covariate residual approach and the value table approach Figure 1 through Figure 5. The plot includes a horizontal band which visually demonstrates the boundaries for 95% limit of agreement. In the figures, each point represents a teacher based on the difference between the relative locations once the value table score was transformed to a similar scale to the covariate regression value in the same grade. The teachers who fall between the lines representing a 95% limit of agreement are within an acceptable range calculated by using 1.96 standard deviations from the mean for each grade level (Bland & Altman, 1986).

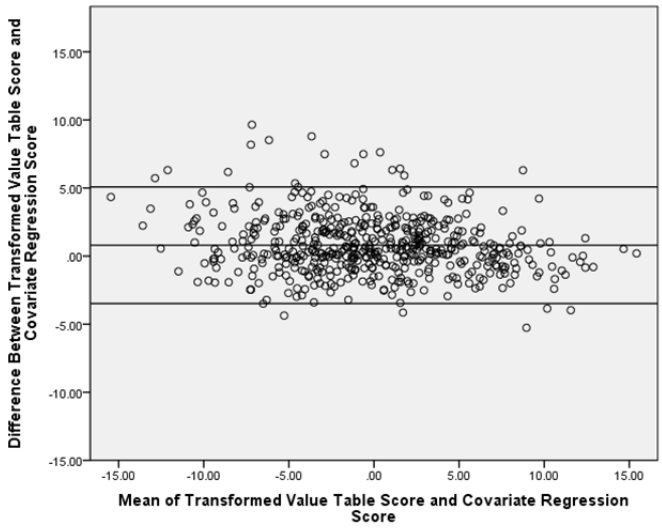

Figure 6 *Comparison of 4th Grade Covariate Regression and Transformed Value Table Scores for Research Question 1*

For fourth grade there are 4% (n = 22) of the 526 teachers who fall outside of the 95% limit of agreement. The upper and lower bound of the 95% limit of agreement were 5.08 and -3.49 respectively with a range of 8.56 points

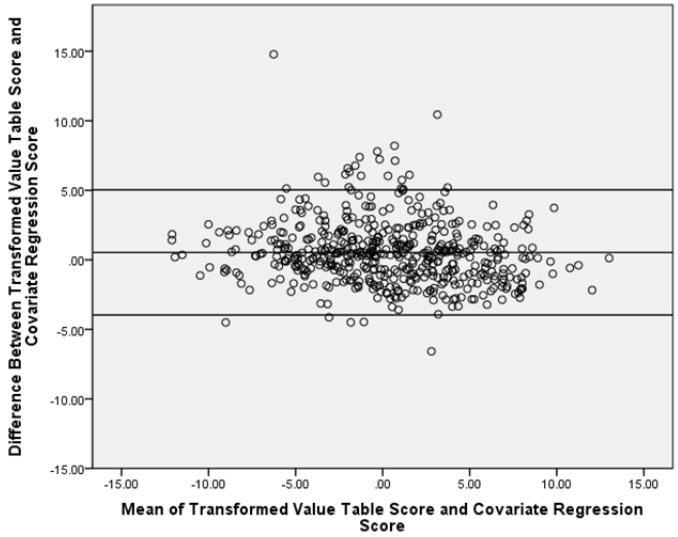

Figure 7 *Comparison of 5th Grade Covariate Regression and Transformed Value Table Scores for Research Question 1*

For fifth grade there are 5% (n = 27) of the 509 teachers who fall outside of the 95% limit of agreement. The upper and lower bound of the 95% limit of agreement were 5.03 and -3.97 respectively with a range of 9.00 points.

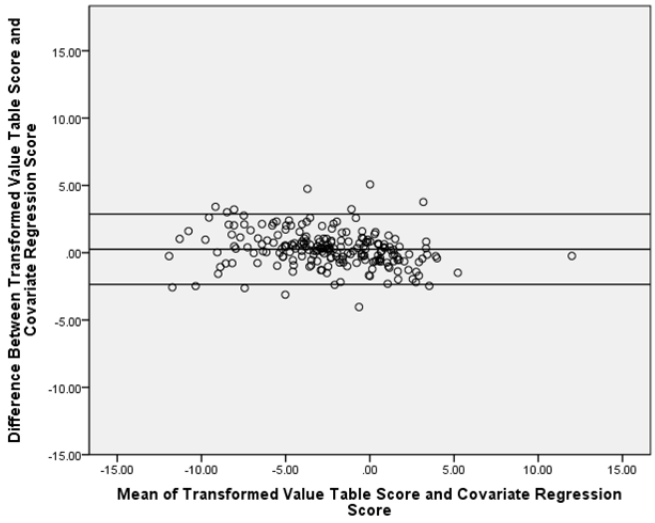

Figure 8 *Comparison of 6th Grade Covariate Regression and Transformed Value Table Scores for Research Question 1*

For Sixth grade there are 5% (n = 10) of the 213 teachers who fall outside of the 95% limit of agreement. The upper and lower bound of the 95% limit of agreement were 2.87 and -2.35 respectively with a range of 5.22 points

For Seventh grade (Figure 9 page 56) there are 6% (n = 13) of the 210 teachers who fall outside of the 95% limit of agreement. The upper and lower bound of the 95% limit of agreement were 3.75 and -2.27 respectively with a range of 6.02 points

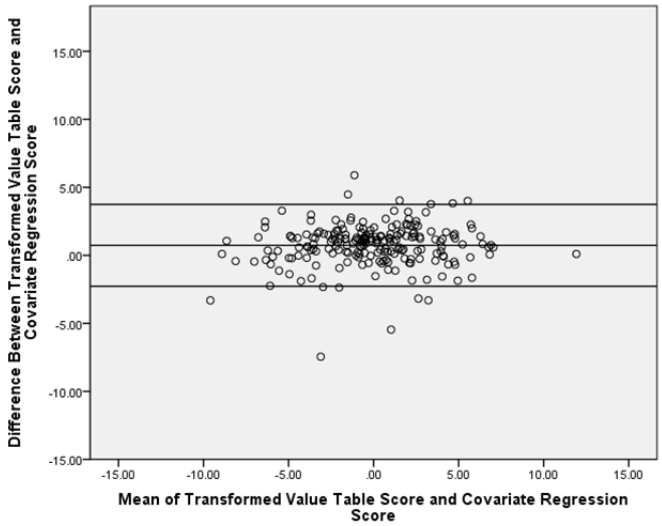

Figure 9 *Comparison of 7th Grade Covariate Regression and Transformed Value Table Scores for Research Question 1*

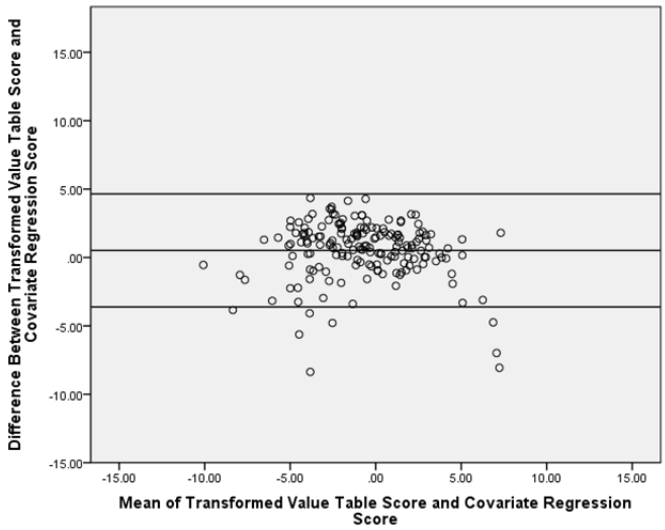

Figure 10 *Comparison of 8th Grade Covariate Regression and Transformed Value Table Scores for Research Question 1*

For eighth grade there are 5% (n = 8) of the 177 teachers who fall outside of the 95% limit of agreement. The upper and lower bound of the 95% limit of agreement were 4.65 and -3.61 respectively with a range of 8.26 points

Examination of quintiles. Quintiles for the Florida covariate residual scores and the value table scores were established and examined, to further examine the accuracy, agreement and disagreement. The first quintile is the lowest quintile, the third quintile is the middle quintile, and the fifth quintile is the highest quintile. Higher quintiles are to the right of the diagonal, lower quintile are those scores to the left of the diagonal together they represent the disagreement between the compared approaches.

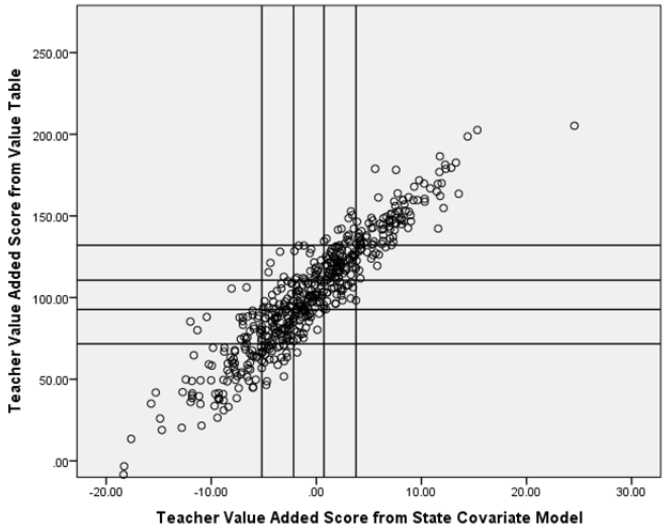

Figure 11 *Grade 4 Scatterplot of Teachers Value Table Score by Covariate Model Score with Quintile Bands Superimposed*

Table 22
Fourth Grade Quintiles for Research Question 1

Total Number of Teachers by Residual Score and Value Table Score Quintiles					
Teacher Residual	Teacher Value Table Score Quintiles				
Score Quintiles	First	Second	Third	Fourth	Fifth
First	79	25	2	0	0
Second	26	50	24	4	0
Third	1	30	54	20	0
Fourth	0	0	24	68	13
Fifth	0	0	1	14	91

*Quintiles that agree between methods are highlighted

Comparing the quintiles of the state residuals to the value table quintiles there are 16% (n = 88) in a higher quintile and 18% (n = 96) were in a lower quintile by the value table scores than by the state residuals. Thus there is a disagreement of 35% for the fourth grade approaches. The accuracy between the two methods is 65% (n = 342). When the immediate neighboring quintiles are considered (for example the second and fourth quintile when considering the third quintile) the level of agreement increases to 98% (n = 518).

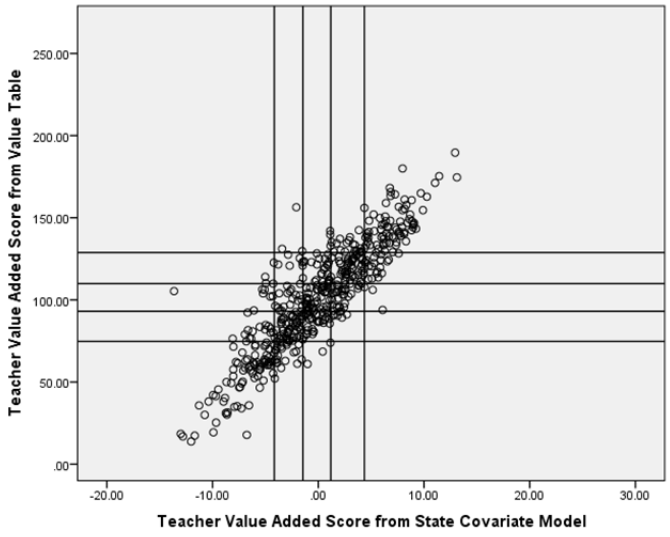

Figure 12 *Grade 5 Scatterplot of Teachers Value Table Score by Covariate Model Score with Quintile Bands Superimposed*

Table 23
Fifth Grade Quintiles for Research Question 1

Total Number of Teachers by Residual Score and Value Table Score Quintiles					
Teacher Residual	Teacher Value Table Score Quintiles				
Score Quintiles	First	Second	Third	Fourth	Fifth
First	75	15	8	3	0
Second	23	52	18	6	3
Third	3	30	42	25	3
Fourth	0	6	31	46	19
Fifth	0	0	3	21	77

*Quintiles that agree between methods are highlighted

Comparing the quintiles of the state residuals to the value table quintiles there are 20% (n = 100) in a higher quintile and 23% (n = 117) were in a lower quintile by the value table scores than by the state residuals. Thus there is a disagreement of 43% for the fourth grade approaches. The accuracy between the two methods is 57% (n = 292). When the immediate neighboring quintiles are considered (for example the second and fourth quintile when considering the third quintile) the level of agreement increases to 93% (n = 474).

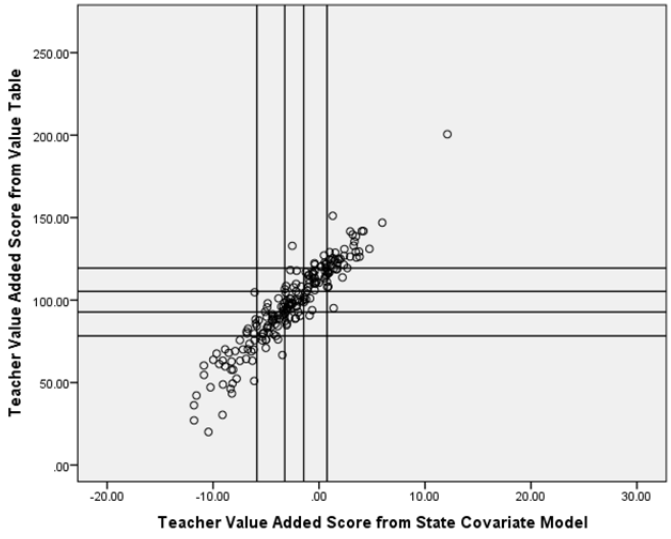

Figure 13 *Grade 6 Scatterplot of Teachers Value Table Score by Covariate Model Score with Quintile Bands Superimposed*

Table 24
Sixth Grade Quintiles for Research Question 1

Total Number of Teachers by Residual Score and Value Table Score Quintiles					
Teacher Residual	Teacher Value Table Score Quintiles				
Score Quintiles	First	Second	Third	Fourth	Fifth
First	35	6	1	0	0
Second	8	26	9	1	0
Third	0	10	23	7	1
Fourth	0	1	7	24	11
Fifth	0	0	1	11	31

*Quintiles that agree between methods are highlighted

Comparing the quintiles of the state residuals to the value table quintiles there are 17% (*n* = 36) in a higher quintile and 18% (*n* = 38) were in a lower quintile by the value table scores than by the state residuals. Thus there is a disagreement of 35% for the fourth grade approaches. The accuracy between the two methods is 65% (*n* = 139). When the immediate neighboring quintiles are considered (for example the second and fourth quintile when considering the third quintile) the level of agreement increases to 98% (*n* = 208).

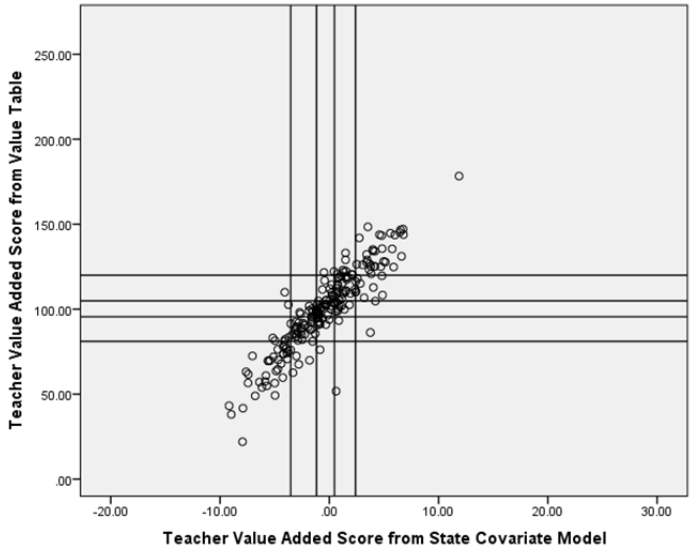

Figure 14 *Grade 7 Scatterplot of Teachers Value Table Score by Covariate Model Score with Quintile Bands Superimposed*

Table 25
Seventh Grade Quintiles for Research Question 1

Total Number of Teachers by Residual Score and Value Table Score Quintiles					
Teacher Residual	Teacher Value Table Score Quintiles				
Score Quintiles	First	Second	Third	Fourth	Fifth
First	35	5	1	1	0
Second	5	26	11	0	0
Third	1	9	20	10	2
Fourth	1	1	9	23	8
Fifth	0	1	1	8	32

*Quintiles that agree between methods are highlighted

Comparing the quintiles of the state residuals to the value table quintiles there are 18% (*n* = 38) in a higher quintile and 17% (*n* = 36) were in a lower quintile by the value table scores than by the state residuals. Thus there is a disagreement of 35% for the fourth grade approaches. The accuracy between the two methods is 65% (*n* = 136). When the immediate neighboring quintiles are considered (for example the second and fourth quintile when considering the third quintile) the level of agreement increases to 96% (*n* = 201).

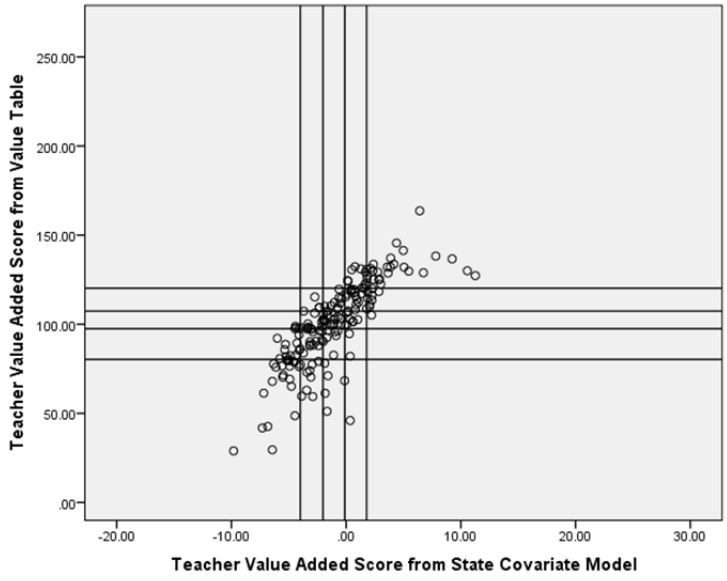

Figure 14 *Grade 8 Scatterplot of Teachers Value Table Score by Covariate Model Score with Quintile Bands Superimposed*

Table 26
Eighth Grade Quintiles for Research Question 1

Total Number of Teachers by Residual Score and Value Table Score Quintiles					
Teacher Residual	Teacher Value Table Score Quintiles				
Score Quintiles	First	Second	Third	Fourth	Fifth
First	20	12	3	0	0
Second	9	14	10	3	0
Third	5	7	14	9	0
Fourth	1	2	8	15	10
Fifth	0	0	1	8	26

*Quintiles that agree between methods are highlighted

Comparing the quintiles of the state residuals to the value table quintiles there are 27% (n = 47) in a higher quintile and 23% (n = 41) were in a lower quintile by the value table scores than by the state residuals. Thus there is a disagreement of 50% for the fourth grade approaches. The accuracy between the two methods is 50% (n = 89). When the immediate neighboring quintiles are considered (for example the second and fourth quintile when considering the third quintile) the level of agreement increases to 92% (n = 162).

Research question two. What is the degree of concordance and discordance of the categories to which mathematics teachers are assigned when the state's recommendations for the classification of teachers into the four evaluation categories are applied to their value-added scores by the value table approach versus the covariate regression approach?

The teacher covariate and the value table value added scores were classified into the four State categories using the guidance provided by the state. Specifically: two standard deviations (SD) above the mean – Highly effective; Less than two SD above the mean and more than one SD below the mean – Effective; one SD below the mean – Needs Improvement; and two SD below the mean – Unsatisfactory (Copa, 2012).

Examination of classification. The accuracy, agreement and disagreement were examined for the classifications (Tables 27 through 31).

Table 27
Fourth Grade Classification for Accuracy and Agreement for Research Question 2

Total Number of Teachers by Residual and Value Table Categories				
	Classification Based on Value Table Category			
Classification Based on Residual Category	Unsatisfactory	Needs Improvement	Effective	Highly Effective
Unsatisfactory	0	3	0	0
Needs Improvement	0	21	11	0
Effective	0	4	484	1
Highly Effective	0	0	0	2

Comparing the classification of the state residuals to the value table classification (Table 27) there are 3% (n = 15) in a higher classification and 1% (n = 4) were in a lower classification by the value table scores than by the state residuals. Thus there is a disagreement of 4% for the fourth grade classification. The accuracy between the two methods is 96% (n = 507).

Table 28
Fifth Grade Classification for Accuracy and Agreement for Research Question 2

Total Number of Teachers by Residual and Value Table Categories				
	Classification Based on Value Table Category			
Classification Based on Residual Category	Unsatisfactory	Needs Improvement	Effective	Highly Effective
Unsatisfactory	0	0	0	0
Needs Improvement	0	22	13	0
Effective	0	3	471	0
Highly Effective	0	0	0	0

Comparing the classification of the state residuals to the value table classification (Table 28) there are 2.6% (n = 13) in a higher quintile and .6% (n = 3) were in a lower classification by

the value table scores than by the state residuals. Thus there is a disagreement of 3% for the fifth grade classification. The accuracy between the two methods is 97% (n = 493).

Table 29
Sixth Grade Classification for Accuracy and Agreement for Research Question 2

Total Number of Teachers by Residual and Value Table Categories				
	Classification Based on Value Table Category			
Classification Based on Residual Category	Unsatisfactory	Needs Improvement	Effective	Highly Effective
Unsatisfactory	0	0	0	0
Needs Improvement	0	7	8	0
Effective	0	4	193	0
Highly Effective	0	0	0	1

Comparing the classification of the state residuals to the value table classification (Table 29) there are 4% (n = 8) in a higher classification and 2% (n = 4) were in a lower classification by the value table scores than by the state residuals. Thus there is a disagreement of 6% for the sixth grade classification. The accuracy between the two methods is 94% (n = 201).

Table 30
Seventh Grade Classification for Accuracy and Agreement for Research Question 2

Total Number of Teachers by Residual and Value Table Categories				
	Classification based on Value Table Category			
Classification Based on Residual Category	Unsatisfactory	Needs Improvement	Effective	Highly Effective
Unsatisfactory	0	0	0	0
Needs Improvement	0	4	4	0
Effective	0	2	200	0
Highly Effective	0	0	0	0

Comparing the classification of the state residuals to the value table classification (Table 30) there are 2% (n = 4) in a higher classification and 1% (n = 2) were in a lower classification by the value table scores than by the state residuals. Thus there is a disagreement of 3% for the seventh grade classifications. The accuracy between the two methods is 97% (n = 204).

Table 31
Eighth Grade Classification for Accuracy and Agreement for Research Question 2

Total Number of Teachers by Residual and Value Table Categories				
	Classification based on Value Table Categories			
Classification Based on Residual Category	Unsatisfactory	Needs Improvement	Effective	Highly Effective
Unsatisfactory	0	0	0	0
Needs Improvement	0	1	0	0
Effective	0	1	175	0
Highly Effective	0	0	0	0

Comparing the classification of the state residuals to the value table classification (Table 31) there are 0% ($n = 0$) in a higher classification and 1% ($n = 1$) were in a lower classification by the value table scores than by the state residuals. Thus there is a disagreement of 1% for the eighth grade classification. The accuracy between the two methods is 99% ($n = 176$).

The relationship between the resulting 4X4 classification table was examined using Kendall's tau-b and Goodman and Kruskal gamma (Table 32). Finally, the accuracy, agreement and disagreement between the categories were examined.

Table 32
Kendall's tau-b and γ

Grade Level	T_b	*p*	γ	*p*	*n*
Grade 4	. 756	<.0001*	.993	<.0001*	526
Grade 5	. 729	<.0001*	.993	<.0001*	509
Grade 6	.553	<.0001*	.960	.004*	213
Grade 7	.563	<.0001*	.980	.042*	210
Grade 8	.705	<.0001*	1.000	.313	177

* *denotes significant results at .05 level*

When examining the T_b and γ the results are consistent with a 4X4 which has an ordinal measure of association for $\rho = .9$ (Göktaş & İşçi, 2011). T_b and γ are both evaluated on a scale of 1.00 to -1.00 with the high positive score representing a strong relationship in a positive direction. T_b and γ are both significant for every grade but the eighth grade. It is interesting to

note that the γ for eighth grade is 1.00 and not significant while the T_b for the same data is significant. The lack of significance for γ is not surprising given that γ examines the concordant and discordant pairs and ignores tied pairs while T_b uses a correction for ties. The distribution of teachers within state categories in the eighth grade has only one teacher who has a discordant classification and all other teachers in the concordant classification.

Chapter 5

Discussion

Summary of Findings

Comparison of Initial Classification Based on Residuals and Value Tables

Research question 1: What is the degree of concordance and discordance between the mathematics teachers' ranking using value-added scores derived from the value table approach versus the covariate regression approach?

The relationship between the teachers' ranking derived from the value table approach versus the covariate regression approach was examined using several approaches. Initially the relationships between the scores were examined using Pearson product moment correlations. Correlations of the teachers' value added scores are above .90 for grades 4, 6, and 7 with the correlation for grade 5 teachers value added scores at .88 and the lowest correlation for the value added scores being .77 for the eighth grade teachers. These correlations demonstrate a strong relationship between the two approaches. The next step was to examine the distribution of scores for each approach. When examining the confidence intervals the skew and kurtosis for the distributions indicate that all are near normal. The scores were compared using a Tukey Mean difference plot or Bland-Altman diagram. The Bland-Altman diagram indicates that the two measures are providing measures that are consistent. The scores were then separated into quintiles and examined for the level of agreement between the quintiles. The 4^{th}, 6^{th} and 7^{th} grade teacher scores had higher agreement than the 5^{th} and 8^{th} grade scores.

These initial examinations of the value added scores for the teachers based on each of the value added methodologies provide support for a high level of agreement and similarity between the other methodological approaches. Similar results have been found in other research examining the relationship between other types of value added approaches. For example, Tekwe, Carter, Ma, Algina, Lucas, Roth, Ariet, Fisher, and Resnick (2004) found that there were little differences between three different types of value-added models and found that the simplest model (simple Fixed Effects Model) had similar results to the more complex Hierarchical Linear Models and Layered Mixed Effects Models. Goldhaber and Theobald (2012) observed high correlations between models that included differing amounts of information for student background. Similar high correlations were observed by Sanders, Wright, Springer, and Langevin (2008), who observed similar stability in scores across disparate student populations.

The actual placement of individual teachers within the ranking, as examined using quintiles, demonstrates that while there is a strong relationship between the methods (65% to 50%) the disagreement (35% to 50%) between the quintiles indicates that it may not be appropriate to utilize one method in lieu of the other depending on the process used to quantify the teachers' position in the ranking based on their value added score. Because this project used actual data from a large southeastern school district it is not possible to identify which method does a more accurate job of correctly classifying the teachers. Further, while the strong correlations combined with the observed differences between the methods at this stage of the analysis are illustrative of the differences between the methods, these differences and similarities accentuate the concerns voiced by researchers related to the understanding of how teachers are evaluated and the relationship of the value added methods (Darling-Hammond, Amrein-Beardsley, Haertel, & Rothstein, 2012; Haertel, 2009).

Relationship of the Two Methods When State Classification is Applied

Research question 2. What is the degree of concordance and discordance of the categories to which mathematics teachers are assigned when the state's recommendations for the classification of teachers into the four evaluation categories are applied to their value-added scores by the value table approach versus the covariate regression approach?

The teacher covariate and the value table value added scores were classified into the four State categories using the guidance provided by the state (Copa, 2012). The relationship between the resulting 4 x 4 classification table was examined using Kendall's tau-b and Goodman and Kruskal's gamma. The Kendall tau-b correlation for each grade level was significant and ranged from .55 to .76. The Goodman-Kruskal's gamma statistic was significant for every grade but 8th where there was only 1 teacher who has a discordant classification and no teacher in the concordant classification. The correlational findings are consistent with the existing research demonstrating agreement between methodologies (Lockwood, McCaffrey, & Sass, 2008; Goldhaber & Theobald, 2012; Lockwood, McCaffrey, Hamilton, Stecher, Le, & Martinez, 2007; Tekwe, Carter, Ma, Algina, Lucas, Roth, Ariet, Fisher, & Resnick, 2004; Goldhaber & Theobald, 2012; Sanders, Wright, Springer, & Langevin, 2008). Finally, the accuracy, agreement and disagreement between the categories were examined. The disagreement for each of the grade levels was very small in every grade and the accuracy of the two methods was very high (94% to 99%).

Considering these findings, in combination with the examination of the quintiles from the first examination, reinforces the argument that a careful consideration should be made when deciding on a means for classifying teachers' value added scores into categories. While there is

a close relationship between the methodologies there are differences at the teacher level that could adversely affect individual teachers' salaries and evaluations. However, given that it is the state classification that is the recommended procedure the extremely high level of agreement between the two methods provides evidence that the value table methods may be a viable proxy that can be used by district administrators for planning.

Limitations of the Study

There are multiple limitations that should be considered when using the results of this project. Data were obtained from a large Florida school district, and given the sample size for each grade ranged from 12,590 to 11,215 the data should mirror the general population of the state, however, the sample is not a random sample of the state and therefore even with the large sample size there is no means to examine representativeness or generalizability. While the overall results support the consistency between the two methods once the classification into categories is used, the results may be different with the inclusion of more districts; without the data from other districts in the state to replicate the results it is not appropriate to generalize these findings to the entire state.

The data used for this study were provided from the information systems of a large district. While there are data correction processes that are employed by the state of Florida to ensure that accountability information is as accurate as possible, there is no way to tell the extent of accuracy of that data. It may be possible that there are errors in the data files provided by the district. Further, in relation to errors in the data and uncertainty, this study did not examine the measures of uncertainty that are available from the calculations of the covariate regression method.

This study was not meant as an examination of the "best method" for identifying the value added by a teacher to their students. This study does not purport to have identified a better method, rather it provides evidence that once the state guidance is applied to the results of the two approaches the differences between the classifications that they provide are nearly negligible for the data that were examined. If one were to want to examine the accuracy of the methods for identifying the value added of a teacher on their individual students it would be more appropriate to conduct a simulation study, which was built on simulated data which reflects known impacts of the individual teacher units and thus allows an examination of the accuracy of the methods. Value Tables were constructed for each grade level from the provided data files. Lack of historical state value tables for mathematics required the reconstruction of the value tables based on feedback from individuals who had constructed value tables for districts and extant information from nontechnical references. It is possible that the resulting tables deviate from the tables used in the past. However the information included in this study should provide adequate guidance in the construction of tables to replicate this project.

Directions for Future Research

There are several directions for future research that this project opens up. Given the findings it would be prudent to conduct a parallel examination of English Language Arts achievement results for the same year using the same techniques. Also, to ensure that the findings were not a spurious result from an anomaly in the years examined, it would be appropriate to replicate the study over several years to ensure that the findings are consistent. Additionally, replicating the project using either the entire state of Florida's data or including a larger sample of districts would provide validation of the findings. An evaluation of the return on

investment for conducting the complex covariate analysis versus the value table approach in conjunction with future replications could provide some useful fiscal feedback for policymakers.

Additionally, it would be useful to explore specifics of the limitations and accuracy of the two methods using a simulation study. There are many possible approaches for related simulation studies. A simulation study would provide the opportunity to control the data thereby having an actual record of the overall impact for each teacher. Such a study could examine the accuracy of each method for classifying teachers based on their impact on students. A simulation study could also provide an opportunity to examine the impacts of errors in the data in relation to the actual accuracy of each method for identifying the impacts of teachers on students and the classification of those teachers into state categories. Further, simulation studies could provide a means of examining the uncertainty and associated measures in each of the approaches in relation to the accuracy of the classification of individual teachers. Simulation studies would provide useful information to both validate the existing findings and for exploring the impact of the limitations for the present project.

Conclusions

Policy Implications. In the late 1990's and early 2000's the state utilized value tables to assess teacher impacts on their assigned students. In an attempt to make teacher evaluations more fair and equitable policymakers pursued complex approaches to parsing the value added by an individual teacher to their students. The decision to use covariate regression makes sense when the statistical controls afforded by the more complex approach are considered. Covariate regression parses information at a more complex level and allows for the inclusion of controls for many exogenous variables that value tables cannot control for. However, in order to maintain consistency across the state, the guidance for classification of results essentially renders the

additional information and controls provided by the covariate regression impractical. The application of the ruleset to classify the value added results has provided a system that takes the results of a complex approach to identification of the value added by a teacher to individual students and converted it to a level that has little to no practicable difference from a value table approach when the same classification ruleset is applied. Thus, it could be argued that it is impractical to use a more complex model when the overall classification from a simpler and more parsimonious system has such high levels of agreement. Further, the use of a complex model that is not easily understood and virtually impossible for a non-technical individual to replicate could serve to alienate teachers and districts from the required accountability and evaluation procedures (Darling-Hammond, Amrein-Beardsley, Haertel, & Rothstein 2012; Haertel, 2009).

Practical Applications. The consistency between approaches once the state guidance is applied provides some useful practical applications at multiple levels within school districts. Since the covariate approach is the adopted state process (Florida statute 1012.986, State board of Education Rule 6A-5.081), it would be inappropriate to replace the resulting information with the value table approach without the approval of the state. However, it takes the state two to three months from initial FCAT results released to districts until the state can provide results from the covariate regression analysis to districts (for example the 2011-12 FCAT results were released to districts at the beginning of June 2012 and the preliminary value added scores were released to the districts in August of 2012). The results of teacher evaluations are used for planning at the school and district level to include assignment of teachers, needed professional development and other program improvement activities. The necessary planning at schools and the district for the upcoming year's activities, especially professional development and teacher

assignment, takes time and is extremely problematic when the evaluation results that are instrumental in the planning are not released until immediately before the beginning of the next school year. A simpler method that districts can calculate and make use of for planning purposes is very attractive. At the individual teacher level value tables could be used to generate a value-added score for individual teachers based on their students' achievement results. This combined with other teacher observation evaluative information can be used to inform their overall professional development plans over the summer while they are waiting on the results of the covariate regression from the state. At the school level principals use all the teacher evaluation information to evaluate professional development needs, general student impacts at the teacher level and as a point for discussion in teacher evaluation processes. At the school type level (elementary, middle, high) and at the district level, administrators can use the general information for districts and school boards to conduct needs assessments and planning when the assessment results are released instead of waiting for an extensive period of time for the covariate regression results to be calculated and returned to the district. The resulting information would only be for the state tested grades, districts would still need to utilize other points of information for planning to include untested grades, the value tables could provide an extremely useful means of developing plans which, based on the results of this project, would be very close to the covariate regression results. While the applicability and usefulness of this approach seems promising as a proxy for the existing required approach it is important to replicate this project at a larger level with more data in a manner that is representative of the entire state to ensure that the results are generalizable. Until such a study is completed and other projects replicate and confirm the findings of this project it is very important for districts to

examine the results for their own teachers carefully for prior school year's data to ensure that the findings for their data are consistent with this project.

References

Aaronson, D., Barrow, L., & Sander, W. (2007). Teachers and student achievement in Chicago public high schools. *Journal of Labor Economics, 25*(2), 95-135.

Alternative Teacher Certification, House Bill 5596, Michigan Statute, 2011

American Institutes for Research. (2011a). *Value Added Model Recommendation to the Commissioner of Education from the Student Growth Implementation Committee.* Retrieved from http://www.fldoe.org/core/fileparse.php/7503/urlt/0072159-summaryfinalrecommendation.pdf.

American Institutes for Research. (2011b). *Value Added Model white Paper*. Retrieved from : http://www.fldoe.org/core/fileparse.php/7503/urlt/0072161-value-added-model-white-paper.doc.

Ballou, D., Sanders, W., & Wright, P. (2004). Controlling for student background in value-added assessment of teachers. *Journal of Educational and Behavioral Statistics, 29*(1), 37-65.

Briggs, D., & Weeks, J. (2009). The impact of vertical scaling decisions on growth interpretations. *Educational Measurement: Issues & Practice, 28*(4). (Special Issue).

Betts, J., Rueben, K., & Danenberg, A. (2000). *Equal Resources, Equal Outcomes? The Distribution of School Resources and Student Achievement in California,* Public Policy Institute of California, San Francisco, California, 2000.

Boyd, D., Lankford H., & Wyckoff, J. (2007). Closing the student achievement gap by increasing the effectiveness of teachers in low-performing schools, Ladd and Fiske (eds). *Handbook of Research in Education Finance and Policy.*

Bacher-Hicks, A., Kane, T. J., & Staiger, D. O. (2014). *Validating teacher effect estimates using changes in teacher assignments in Los Angeles*, National Bureau of Economic Research Working Paper Series No. 20657 Issued in November 2014.

Bock, R. D., Wolfe, R. G., & Fisher, T. H. (1996). *A review and analysis of the Tennessee value-added assessment system,* Nashville, Tennessee Comptroller of the Treasury, Office of Educational Accountability.

Boyd, D, Lankford H., Loeb S, & Wyckoff J. (2005). The draw of home: how teachers' preferences for proximity disadvantage urban schools. *Journal of Policy Analysis and Management* 24:113-132.

Boyce, A. (1915). Methods of measuring teachers' efficiency. *Fourteenth yearbook of the national society for the study of education, Part II.* Bloomington, IL: Public School Publishing Co.

Bonesronning, H., Falch, T., & Strom, B. (2005). Teacher sorting, teacher quality, and student composition. *European Economic Review, Elsevier, 49*(2), 457-483.

Brock, C. (1981). Education in Latin America: aspects and issues in the mid-twentieth century. *International Journal of Educational Development, 1*, 50-64.

Ballou, D., Sanders, W., & Wright, P. (2004). Controlling for student background in value-added assessment of teachers. J*ournal of Educational and Behavioral Statistics, 29*(1), 37-65.

Bianchi, A.B. (2003). A new look at accountability: 'value-added' assessment. *Forecast: Emerging Issues in Public Education, 1*(1), 1-4.

Bland, J.M., Altman, D.G. (1986). Statistical methods for assessing agreement between two methods of clinical measurement *Lancet 327 (8476): 307–10.* doi:10.1016/S0140-6736(86)90837-8.

Brooks-Gunn, J., & Duncan, G.J. (1997). The future of children. *Children and Poverty, 7*(2), 55-71.

Carey, K. (2004). The real value of teachers: Using new information about teacher effectiveness to close the achievement gap. *Thinking K-16: A Publication of the Education Trust, 8*(1), 3-42.

Chetty, R., Friedman, J.N., & Rockoff, J.E. (2014a). Measuring the impacts of teachers in evaluating bias in teacher value-added estimates. *American Economic Review; 104*(9), 2593-2632.

Chetty, R., Friedman, J.N., & Rockoff, J.E. (2014b). Prior test scores do not provide valid placebo tests of teacher switching research designs. Retrieved from http://obs.rc.fas.harvard.edu/chetty/va_prior_score.pdf.

Chetty, R., Friedman, J.N., & Rockoff, J.E. (2014c). Response to Rothstein (2014) 'revisiting the impacts of teachers.' Retrieved from http://obs.rc.fas.harvard.edu/chetty/Rothstein_response.pdf.

Copa, J. (2012). Florida's value-added model measuring student learning growth. *Presentation at the 57th annual meeting of the Florida Educational Research Association*, Gainesville, FL.

Coleman, J. S., Hoffer, T., & Kilgore, S. (1982). *High School achievement: Public, Catholic, and private schools compared*. New York: Basic Books.

Clotfelter, C.T., Ladd, H.F., & Vigdor, J.L. (2007). How and why do teacher credentials matter for student achievement? *National Center for Analysis of Longitudinal Data in Education Research*. Retrieved from http://www.caldercenter.org/sites/default/files/1001058_Teacher_Credentials.pdf.

Curtis, R. (2012a). *Building it together: the design and implementation of Hillsborough county public schools' teacher evaluation system*. The Aspen Institute: Education & Society Program. Retrieved from http://www.aspendrl.org/portal/browse/DocumentDetail?documentId=1068&download.

Curtis, R. (2012b). *Putting the pieces in place: Charlotte-Mecklenburg public schools' teacher evaluation system.* The Aspen Institute: Education & Society Program. Retrieved from http://www.aspeninstitute.org/sites/default/files/content/docs/pubs/ed-CharlotteREP4.pdf.

Darling-Hammond, L., Wise, A.E., & Pease, S.R. (1983). Teacher evaluation in the organizational context: A review of the literature. *Review of Educational Research, 53*(3), 285-328.

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

Dougherty, C. (2008). The power of longitudinal data: Measuring student academic growth. *National Center for Educational Accountability, Data Quality Campaign.*

Dougherty, C. (2007). Measures of adequate growth. *National Center for Educational Accountability, Data Quality Campaign, 2*(4), 2.

Doyle, D., & Han, J.G. (2012). *Measuring teacher effectiveness: A look "under the hood" of teacher evaluation in 10 sites*. New York: 50CAN; New Haven, CT: ConnCAN; and Chapel Hill, NC: Public Impact. Retrieved from http://www.conncan.org/learn/research/teachers/measuring-teacher-effectiveness.

Doran, H.C., & Izumi, L.T. (2004). *Putting education to the test: A value-added model for California*. Pacific Research Institute. Retrieved from https://www.heartland.org/policy-documents/putting-education-test-value-added-model-california.

District of Columbia Public Schools. (2010). *IMPACT: The District of Columbia Public Schools effectiveness assessment system for school-based personnel 2010-2011. Group 1 general education teachers with individual value-added student achievement data.* Washington, DC: DCPS.

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

Education Law Repeals, House Bill 7087, Florida Statute § 1012.34, 2011

Failing School Reform, House Bill 4787, Michigan, 2011

FLDOE, (2014) FCAT 2.0 and Florida EOC Assessments Achievement Levels Retrieved from http://www.fldoe.org/core/fileparse.php/5662/urlt/0095809-achlevel.pdf

Ferguson, H., Bovaird, S., & Mueller, M. (2007). The impact of poverty on educational outcomes for children. *Pediatrics & Child Health*, *12*(8), 701–706.

Goldschmidt, P., Choi, K., & Martinez, F. (2004). *Using Hierarchical Growth Models To Monitor School Performance Over Time: Comparing NCE to Scale Score Results.* Los Angeles: University of California, Center for the Study of Evaluation (CSE).

Goldhaber D., & Chaplin D. (2015). Assessing the "Rothstein falsification test." Does it really show teacher value-added models are biased*? J*ournal *of Research on Educational Effectiveness,* 8(1), 8–34.

Göktaş, A., & İşçi, Ö., (2011) A Comparison of the Most Commonly Used Measures of Association for Doubly Ordered Square Contingency Tables via Simulation Metodološki zvezki, Vol. 8, No. 1, 2011, 17-37

Guarino, C., Reckase, M., & Wooldridge, J. (2015). Can value-added measures of teacher performance be trusted? *Education Finance and Policy*, *10*(1).

Goldhaber, D. (2002). The mystery of good teaching. *Education Next*, 50–55.

Goldhaber, D. and Theobald R. (2012). *Do Different Value-Added Models Tell Us the Same Things?* Carnegie Knowledge Network. Retrieved from http://carnegieknowledgenetwork.org/briefs/value-added/different-growth-models.

Goldhaber D., & Brewer D. (1997). *Evaluating the effect of teacher degree level on educational performance*. In Fowler (Ed.), Developments in school finance, 1996 (pp. 197–210). Washington, DC: National Center for Education Statistics, U.S. Department of Education.

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

Glazerman, S., Goldhaber, D, Loeb, S., Staiger, D., Raudenbush, S., & Whitehurst, G.J. (2010). *Evaluating teachers: The important role of value-added*. Washington, DC: The Brookings Institution.

Goldhaber, D., & Hansen, M. (2013). Is it just a bad class? Assessing the long-term stability of estimated teacher performance. *Economica*, *80*(319), 589-612.

Hanushek, E.A., Kain, J.F,, & Rivkin, S.G. (2004). Why public schools lose teachers. *Journal of Human Resources, 39*(2), 326-354.

Hanushek, E.A., Kain, J.F,, O'Brien, D., & Rivkin, S.G. (2005). The market for teacher quality. *National Bureau of Economic Research (NBER) Working Papers* 11154, National Bureau of Economic Research, Inc.

Haefele, D. (1980). How to evaluate thee, teacher — Let me count the ways. *Phi Delta Kappan*, *61*(5), 349-352.

Haefele D. (1993) Evaluating teachers: A call for change. *Journal of Personnel Evaluation in Education, 7*(1)*,* 21-31.

Haefele, D. (1992). Evaluating teachers: An alternative model. *Journal of Personnel Evaluation in Education*, *5*(4), 335-345.

Haertel, E.H. (2009). *Letter Report to the U.S. Department of Education on the Race to the Top Fund*. Washington, DC: The National Academies Press. Retrieved from http://www.nap.edu/catalog.php?record_id=12780.

Harris, D.N., & Sass, T.R. (2008). Teacher training, teacher quality and student achievement. *Journal of Public Economics, 95*(8)*,* 798-912.

Hamilton, L. S. (2004). Improving inferences about student achievement: A multi-dimensional perspective. In S. J. Paik (Ed.), *Advancing educational productivity: Policy implications from national databases* (pp. 175–201). Greenwich, CT: Information Age Publishing.

Henry, G.T., Thompson, C.L., Bastian, K.C., Kershaw, D.C., Purtell, K.M., & Zulli, R.A. (2011). Does teacher preparation affect student achievement? *Chapel Hill, NC: Carolina Institute for Public Policy Working Paper, Version dated February*, *7*, 2011. Retrieved from http://www.maxwell.syr.edu/uploadedFiles/cpr/events/cpr_seminar_series/Henry_paper.pdf.

Hershberg, T., Simon, V. A., & Lea-Kruger, B. (2004). Measuring what matters. *National School Boards Association*. Retrieved from http://www.asbj.com/2004/02/0204asbjhershberg.pdf.

Hill, R. (2005). *Measuring student growth through value tables*. Paper presented at the Council of Chief State School Officers. LSA conference, San Antonio, TX.

Hill, R. (2006a). *Developing a value table for Alaska's public school performance incentive program*. Juneau, AK: Alaska Department of Education and Early Development. Retrieved from http://www.eed.state.ak.us/spip/DevelopingValueTableforAlaska.pdf.

Hill, R. (2006b). *Using value table for a school-level accountability system*. Paper presented at the 2006 annual meeting of the National Council on Measurement, San Francisco, CA.

Hill, R., Marion, S., DePascale, C., Dunn, J., & Simpson, M. (2006). *Using value tables to explicitly value student growth*. In R.W. Lissitz (Ed.), Longitudinal and value-added models of student performance (pp. 255-282). Maple Grove, MN: JAM Press.

Janus, M., Walsh, C., Viveiros, H., Duku, E., & Offord, D. (2003). *School readiness to learn and neighborhood characteristics.* Poster presented at the Biennial meeting of the Society for Research on Child Development. Tampa. FL.

Jensen, E. (2009). *Teaching with Poverty in Mind: What Being Poor Does to Kids' Brains and What Schools Can Do About It*. Alexandria, VA: Association for Supervision and Curriculum Development.

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

Kalogrides, D., Loeb, S., & Beteille, T. (2013). Systematic sorting: Teacher characteristics and class assignments. *Sociology of Education*, *86*(2), 103-123.

Kain, J. F. (1998, October). The impact of individual teachers and peers on individual student achievement. In *annual meeting of the Association for Public Policy Analysis and Management, New York, NY*.

Kane, T., Rockoff, J, & Staiger, D. (2006). *What does certification tell us about teacher effectiveness? Evidence from New York City.* Working paper no. 12155, National Bureau of Economic Research, Cambridge, MA.

Kane, T., & Staiger, D. (2008). *Estimating teacher impacts on student achievement: An experimental evaluation.* NBER Working Paper No. 14607.

Kane, T., McCaffrey, D.F., Miller, T., & Staiger, D. (2013). *Have we identified effective teachers? Validating measures of effective teaching using random assignment*. Seattle, WA: Bill and Melinda Gates Foundation.

Kingsbury, G.G., Olson, A., McCahon, D., & McCall, M.S. (2004, July). *Adequate yearly progress using the hybrid success model: A suggested improvement to No Child Left Behind*. Paper presented at a forum on NCLB sponsored by the Center for Education Policy, Washington, D.C. Retrieved from http://www.nwea.org/research/grd.asp.

Kinsler, J. (2012), Assessing Rothstein's critique of teacher value-added models. *Quantitative Economics, 3*(2), 333-362.

Koedel, C., & Betts, J. (2005). *Re-examining the role of teacher quality in the educational production function (Technical report).* Department of Economics, UCSD.

Koedel, C. and Betts, J. (2007). *Re-Examining the Role of Teacher Quality in the Educational Production Function.* Working Paper 07-08, University of Missouri, Columbia.

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

Lockwood, J., Doran, H., & McCaffrey, D. F. (2003). Using r for estimating longitudinal student achievement models. *The Newsletter of the R Project, 3*(3), 17–23.

Lockwood, J., McCaffrey, D., Hamilton, L., Stecher, B., Le, V., & Martinez, J. (2007). The sensitivity of value-added teacher effect estimates to 31 different mathematics achievement measures. *Journal of Educational Measurement, 44*(1), 47–67.

Lyons, R. (2004, August 5). The influence of socioeconomic factors on Kentucky's public school accountability system: Does poverty impact school effectiveness? *Education Policy Analysis Archives, 12*(37). Retrieved from http://epaa.asu.edu/epaa/v12n37/.

Lankford, H., Loeb, S., & Wyckoff, J. (2002). Teacher sorting and the plight of urban schools: a descriptive analysis. *Educational Evaluation and Policy Analysis, 24*(1), 37-62.

Lamke, T.A., & Chairman, (1955 June). Introduction. Review of Educational Research, *25*(3), 192.

Latham, W. (1982). *Increasing productivity through performance appraisals*, Addison Wesley, Readin MA

Lacour, M., & Tissington, L.D. (2011). The effects of poverty on academic achievement *Educational Research and Reviews*, *7*(9), 522-527.

Lassila, C (2006). Special Teachers Are Rewarded (STAR), Paper presented at FASPA Conference November 2, 2006 Retrieved from http://www.faspa.net/STAR-FASPA-Conference.pdf.

Lower, M. (1987). *A study of principals' and teachers' perceptions and attitudes toward the evaluation of teachers* (Doctoral dissertation). The Ohio State University, Columbus, OH.

Lockwood J.R., McCaffrey D.F., Hamilton L.S., Stecher B.M., Le V., & Martinez, F. (2007). The sensitivity of value-added teacher effect estimates to different mathematics achievement measures. *Journal of Educational Measurement*, *44*(1), 45-65.

Lockwood, J., & McCaffrey, D. (2008). Exploring student-teacher interactions in longitudinal achievement data – Florida longitudinal dbase - Duval, Hillsborough, Orange and Palm Beach.

Lockwood, J., McCaffrey, D., & Sass, T. (2008). The intertemporal stability of teacher effect estimates. Paper presented at the Wisconsin Center for Educational Research National Conference on Value-Added Modeling, Madison, WI, April 2008.

Merit Award Program for Instructional Personnel and School-Based Administrators , FL § 1012.225

McCall, M.S., Kingsbury, G.G., & Olson, A. (2004). *Individual Growth and School Success.* Lake Oswego, OR: Northwest Evaluation Association

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

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

Martineau, J. A. (2006). Distorting value-added: The use of longitudinal, vertically scaled student achievement data for value-added accountability. *Journal of Educational and Behavioral Statistics, 31*(1)*,* 35–62.

Martineau, J.A. (2005). Un-Distorting measures of growth: alternatives to traditional vertical scales. Paper presented at the 35th Annual National Conference on Large-Scale Assessment of the Council of Chief State School Officers, June 19, 2005.

Medley, D., Coker, H., & Soare R. (1984). Measurement-based evaluation of teacher performance: An empirical approach. New York: Longman.

Mendro, R., Jordan, H., Gomez, E., Anderson, M., & Bembry, K. (1998). An application of multiple linear regression in determining longitudinal teacher effectiveness. Paper presented at the 1998 Annual Meeting of the AERA, San Diego, CA.

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

Newton, X., Darling-Hammond, L., Haertel, E., & Thomas, E. (2010). Value-added modeling of teacher effectiveness: an exploration of stability across models and contexts. *Education Policy Analysis Archives*, *18*(23).

Nye, B., Konstantopoulos, S., & Hedges, L. (2004). How large are teacher effects? *Educational Evaluation and Policy Analysis, 26*(3), 237-257.

Office of Program Policy Analysis and Government Accountability. (2007). *Restrictive District Requirements Limited Participation in Performance Pay Systems*, OPPAGA report 07-01, Joint Legislative Auditing Committee, Florida.

Ome, A. (2013). Teacher Sorting: Should We Be Concerned? The Case of Colombia Poster Paper: at APPAM Annual Fall research conference, November 2013.

Paufler, N., & Amrein-Beardsley, A. (2014). The random assignment of students into elementary classrooms implications for value-added analyses and interpretations. *American Educational Research Journal*, *51*(2), 328-362.

Personnel evaluation procedures and criteria , FL § 1012.34

Peske, H.G., & Haycock, K. (2006). *Teaching inequality: How poor and minority students are shortchanged on teacher quality*. The Education Trust.

Public Employment Relations Act (PERA), House Bill 4788, Michigan Statute, 2011

Rothstein, J. (2010). Teacher quality in educational production: tracking, decay, and student achievement. *The Quarterly Journal of Economics*, *125*(1), 175-214.

Rockoff, J.E. (2004). The impact of individual teachers on student achievement: Evidence from panel data. *American Economic Review Papers and Proceedings, 94*(2), 247-252.

Rice, J. (2003). *Teacher quality: Understanding the effectiveness of teacher attributes*. Economic Policy Institute, 1660 L Street, NW, Suite 1200, Washington, DC 20035.

Rivkin, S. G., Hanushek, E. A., & Kain, J. F. (2005). Teachers, schools, and academic achievement. *Econometrica*, *73*(2), 417-458.

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

Sanders, W.L. (1989). *A multivariate mixed model* In: *Applications of mixed models in agriculture and related disciplines*. Southern Cooperative Series Bulletin No. 343. Louisiana Agricultural Experiment Station, pp. 138–144

Sanders, W.L. (2000). Value-added assessment from student achievement data: opportunities and hurdles. *Journal of Personnel Evaluation in Education*, *14*(4), 329–339.

Sanders, W.L. & Rivers, J. (1996). *Cumulative and residual effects of teachers on future student academic achievement.* Knoxville, TN: University of Tennessee Value-Added Research Center

Sanders, W., Saxton, A., & Horn, B. (1997). The Tennessee Value-Added Assessment System: A quantitative outcomes-based approach to educational assessment. In J. Millman (Ed.), *Grading Teachers, Grading Schools: Is Student Achievement a Valid Evaluational Measure?* Thousand Oaks, CA: Corwin Press, Inc., 137–162.

Sanders, W., Wright, S., Springer, M., & Langevin, W. (2008). D*o teacher effects persist when teachers move to schools with different socioeconomic environments?* Presented at National Center for Performance Incentives conference Performance Incentives: Their Growing Impact on American K-12 Education, Vanderbilt University, Nashville, TN, February 2008.

Sass, T. R. (2008). The stability of value-added measures of teacher quality and implications for teacher compensation policy. Brief 4. *National Center for Analysis of Longitudinal Data in Education Research*.

Sass, T.R., Semykina, A., & Harris, D.N. (2014). Value-added models and the measurement of teacher productivity. *Economics of Education Review*, *38*(C), 9-23.

Schmitz, D.D., & Raymond, K.J. (2008). *The utility of the cumulative effects model in a statewide analysis of student achievement.* Paper presented at the American Educational Research Association Annual Meeting, New York, NY on March 27, 2008.

Schmidt, W. H., Houang, R. T., & McKnight, C. C. (2005). Value-added research: Right idea but wrong solution? In R. Lissitz (Ed.), *Value-added models in education: Theory and practice* (pp. 272–297). Maple Grove, MN: JAM Press.

Scherrer, J. (2011). Measuring teaching using value-added modeling: The imperfect panacea. *NASSP Bulletin, 95*(2), 122-140.

Schools of Excellence, Senate Bill 981, Michigan Statute, 2011

Scriven, M. (1980). The evaluation of college teaching. National Council of States on Inservice Education. (ERIC Document Reproduction Service No. 203 729).

Smith, M.L., & Fey, P. (2000). Validity and accountability in high-stakes testing. *Journal of Teacher Education*, *51*(5), 334-344.

Student Success Act of 2011, Senate Bill 736, Florida Statute 2011 (1012.34, F.S.)

Sweeney, J., & Manatt, R. (1986). Teacher evaluation. In R. Berk (Ed.), Performance assessment: Methods and applications (pp. 446–468). Baltimore: John Hopkins University Press.

Taylor, R. (2008). *Delaware's Growth Model and Results from Year One.* Retrieved from www.ecs.org/html/meetingsEvents/NF2008/resources/GrowthModelpresentation-ECS-070208.ppt

Teacher Data System/Basic Instruction Supplies, Senate Bill 926, Michigan Statute, 2011

Tekwe, C., Carter, R., Ma, C., Algina, J, Lucas, M. E., Roth, J., Ariet, M., Fisher, T., & Resnick, M. B. (2004). An empirical comparison of statistical models for value-added assessment of school performance. *Journal of Educational and Behavioral Statistics, 29*(1), 11-36.

U.S. Department of Education. (2006). *Guidance: Cross Cutting Document.* Retrieved from http://www.ed.gov/admins/lead/account/growthmodel/index.html.

U.S. Department of Education. (2008a). *Cross Cutting Guidance: Growth Model Proposal Peer Recommendation for the NCLB Growth Model Pilot Applications.* Retrieved from http://www.ed.gov/admins/lead/account/growthmodel/index.html.

U.S. Department of Education. (2008b). *Press release: U. S. Secretary of Education Margaret Spellings Approves Additional Growth Model Pilots for 2007-2008 School Year* Retrieved from http://www.ed.gov/admins/lead/account/growthmodel/index.html.

U.S. Department of Education. (2009a). *Guidance: Guidance for States to Include Growth Models in AYP Determinations* Retrieved from http://www.ed.gov/admins/lead/account/growthmodel/index.html.

U.S. Department of Education. (2009b). *Press release: Secretary Spellings Approves Additional Growth Model Pilots for 2008-2009 School Year* Retrieved from http://www.ed.gov/admins/lead/account/growthmodel/index.html.

Wiley, E. W. (2006). A practitioner's guide to value-added assessment: Educational policy studies laboratory research monograph. Retrieved from http://nepc.colorado.edu/publication/a-practitioners-guide-value-addedassessment-educational-policy-studies-laboratory-resea

Wiener, R., & Jacobs, A. (2011). *Designing and Implementing Teacher Performance Management Systems: Pitfalls and Possibilities.* Retrieved from http://www.aspendrl.org/portal/browse/DocumentDetail?documentId=580&download.

Wayne, A., & Young, P. (2003). Teacher characteristics and student achievement gains: A review. *Review of Educational Research*, *73*(1), 89-122.

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

Wilson, S., & Floden, R. (2003). *Creating Effective Teachers: Concise Answers for Hard Questions. An Addendum to the Report "Teacher Preparation Research: Current Knowledge, Gaps, and Recommendations."*. AACTE Publications, 1307 New York Avenue, NW, Suite 300, Washington, DC 20005-4701.

Wilson, S., Floden, R., & Ferrini-Mundy, J. (2001). Teachers preparation research: Current knowledge, gaps, and recommendations. Seattle, WA: Center for the Study of Teaching and Policy.

Wright, S.,Horn, S, and Sanders W. (1997). Teacher and classroom context effects on student achievement: implications for teacher evaluation. *Journal of Personnel Evaluation in Education* 11, 57–67.

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

Yamamoto, K. (1963). Creative writing and school achievement. *School and Society, 91*, 307-308.

Appendices

Appendix A:
Data gathering and analysis plan

1. Procure the data
    a. Request file
2. Clean and prepare data
    a. Identify students with two years of test data
        i. Identify individual students achievement level from Developmental Scale Scores based on FLDOE conversions and extant data sources.
        ii. Produce cross tabulation of pretest and posttest results for use in the construction of value tables.
    b. Identify students who meet inclusion criteria
        i. Residual score and;
        ii. Achievement data for school years 2010-11 and 2011-12 and;
        iii. Attendance rate of .8 (90 days)
3. construct value tables
    a. Create a value table for each grade.
    b. Calculate individual student value table score based on pretest posttest information.
4. Prepare data for aggregation of teacher information

a. Remove students who did not have: a residual score, achievement data for the school years 2010-2011 and 2011-2012 and an attendance rate of 90 days or more.
b. Remove teachers who had 10 or fewer students assigned to them.

5. Aggregate teacher value added scores
    a. Value table scores: Aggregate average teacher value table scores based on assigned students who meet the inclusion criteria
    b. Regression residual score: Aggregate average teacher residual based on assigned students who meet the inclusion criteria
6. Calculate quintiles
    a. Identify and execute cut points for value table scores into quintiles based on visual binning procedure in SPSS 22.0
    b. Identify and execute cut points for residuals (state regression) scores into quintiles based on visual binning procedure in SPSS 22.0
7. Calculate state categories
    a. Identify and execute cut points for value table scores into state categories based on state guidance.
    b. Identify and execute cut points for residuals (state regression) scores into state categories based on state guidance
8. Compare value table scores with aggregated (regression) residual scores
    a. Examine distributions of value table scores and residual scores
    b. Pearson Product Moment Correlation for each grade level

c. Construct Tukey mean-difference plots (Bland-Altman Diagrams) to examine relationship of relative classification for each method

d. Examine disagreement and agreement of quintile classification for each value added method.

9. Compare state classifications for value table and residuals (state regression)

   a. Examine accuracy, agreement, and disagreement between categories.

   b. Examine the relationship of the resulting 4X4 classification table using Kendall's tau-b (T_b) and Goodman and Kruskal gamma (γ).

Appendix B:
Value Tables

The values in the tables are explained in the following manner:
The N row reflects the number of students from the level of achievement on the pretest to the associated level of achievement on the post test. The proportion row is the proportion of those students at the pretest level who were found in the associated posttest level. The value row is the associated value for that level which is distributed as closely as possible across each posttest level to provide a check value of 100.

Table B.1 4th grade Value Tables calculations

Achievement Level on Post-test for students with Low Achievement Level 1 on pretest

	Low level 1	High Level 1	Level 2	Level 3	Level 4	Level 5	Total N
N	228	140	20	2	0	0	**390**
Proportion	0.58	0.36	0.05	0.01	0.00	0.00	1.00
Value	0	210	420	629	839	1049	
Check	0	75	22	3	0	0	100

Achievement Level on Post-test for students with High Achievement Level 1 on pretest

	Low level 1	High Level 1	Level 2	Level 3	Level 4	Level 5	Total N
N	218	961	418	90	3	0	**1690**
Proportion	0.13	0.57	0.25	0.05	0.00	0.00	1.00
Value	-50	72	194	316	438	560	
Check	-6	41	48	17	1	0	100

Achievement Level on Post-test for students with Achievement Level 2 on pretest

	Low level 1	High Level 1	Level 2	Level 3	Level 4	Level 5	Total N
N	30	698	1415	812	131	12	**3098**
Proportion	0.01	0.23	0.46	0.26	0.04	0.00	1.00
Value	-100	-5	89	184	279	373	
Check	-1	-1	41	48	12	1	100

Achievement Level on Post-test for students with Achievement Level 3 on pretest

	Low level 1	High Level 1	Level 2	Level 3	Level 4	Level 5	Total N
N	5	126	902	1761	901	135	**3830**
Proportion	0.00	0.03	0.24	0.46	0.24	0.04	1.00
Value	-150	-67	17	100	183	267	
Check	0	-2	4	46	43	9	100

Achievement Level on Post-test for students with Achievement Level 4 on pretest

	Low level 1	High Level 1	Level 2	Level 3	Level 4	Level 5	Total N
N	0	1	95	684	912	451	**2143**
Proportion	0.00	0.00	0.04	0.32	0.43	0.21	1.00
Value	-200	-121	-43	36	114	200	
Check	0	0	-2	11	49	42	100

Achievement Level on Post-test for students with Achievement Level 5 on pretest

	Low level 1	High Level 1	Level 2	Level 3	Level 4	Level 5	Total N
N	0	1	5	131	512	687	**1336**
Proportion	0.00	0.00	0.00	0.10	0.38	0.51	1.00
Value	-250	-171	-91	-12	68	147	
Check	0	0	0	-1	26	76	100

Table B.2 5th grade Value Tables calculations

Achievement Level on Post-test for students with Low Achievement Level 1 on pretest

	Low level 1	High Level 1	Level 2	Level 3	Level 4	Level 5	Total N
N	291	200	16	1	0	0	**508**
Proportion	0.57	0.39	0.03	0.00	0.00	0.00	1.00
Value	0	217	434	651	868	1085	
Check	0	85	14	1	0	0	100

Achievement Level on Post-test for students with High Achievement Level 1 on pretest

	Low level 1	High Level 1	Level 2	Level 3	Level 4	Level 5	Total N
N	260	1034	643	100	2	0	**2039**
Proportion	0.13	0.51	0.32	0.05	0.00	0.00	1.00
Value	-50	66	183	299	416	532	
Check	-6	34	58	15	0	0	100

Achievement Level on Post-test for students with Achievement Level 2 on pretest

	Low level 1	High Level 1	Level 2	Level 3	Level 4	Level 5	Total N
N	33	525	1434	816	102	4	**2914**
Proportion	0.01	0.18	0.49	0.28	0.04	0.00	1.00
Value	-100	-7	86	179	272	365	
Check	-1	-1	42	50	10	1	100

Achievement Level on Post-test for students with Achievement Level 3 on pretest

	Low level 1	High Level 1	Level 2	Level 3	Level 4	Level 5	Total N
N	3	83	834	1760	782	142	**3604**
Proportion	0.00	0.02	0.23	0.49	0.22	0.04	1.00
Value	-150	-67	16	99	182	265	
Check	0	-2	4	48	39	10	100

Achievement Level on Post-test for students with Achievement Level 4 on pretest

	Low level 1	High Level 1	Level 2	Level 3	Level 4	Level 5	Total N
N	0	2	85	601	1084	565	**2337**
Proportion	0.00	0.00	0.04	0.26	0.46	0.24	1.00
Value	-200	-125	-49	26	101	200	
Check	0	0	-2	7	47	48	100

Achievement Level on Post-test for students with Achievement Level 5 on pretest

	Low level 1	High Level 1	Level 2	Level 3	Level 4	Level 5	Total N
N	0	0	3	51	348	786	**1188**
Proportion	0.00	0.00	0.00	0.04	0.29	0.66	1.00
Value	-250	-174	-98	-22	53	129	
Check	0	0	0	-1	16	86	100

Table B.3 6th grade Value Tables calculations

Achievement Level on Post-test for students with Low Achievement Level 1 on pretest

	Low level 1	High Level 1	Level 2	Level 3	Level 4	Level 5	Total N
N	270	200	27	1	0	0	**498**
Proportion	0.54	0.40	0.05	0.00	0.00	0.00	1.00
Value	0	194	388	581	775	969	
Check	0	78	21	1	0	0	100

Achievement Level on Post-test for students with High Achievement Level 1 on pretest

	Low level 1	High Level 1	Level 2	Level 3	Level 4	Level 5	Total N
N	349	987	493	44	2	0	**1875**
Proportion	0.19	0.53	0.26	0.02	0.00	0.00	1.00
Value	-50	83	216	349	482	616	
Check	-9	44	57	8	1	0	100

Achievement Level on Post-test for students with Achievement Level 2 on pretest

	Low level 1	High Level 1	Level 2	Level 3	Level 4	Level 5	Total N
N	99	801	1568	583	40	2	**3093**
Proportion	0.03	0.26	0.51	0.19	0.01	0.00	1.00
Value	-100	6	111	217	323	428	
Check	-3	1	56	41	4	0	100

Achievement Level on Post-test for students with Achievement Level 3 on pretest

	Low level 1	High Level 1	Level 2	Level 3	Level 4	Level 5	Total N
N	11	109	801	1646	576	31	**3174**
Proportion	0.00	0.03	0.25	0.52	0.18	0.01	1.00
Value	-150	-63	24	111	198	286	
Check	-1	-2	6	58	36	3	100

Achievement Level on Post-test for students with Achievement Level 4 on pretest

	Low level 1	High Level 1	Level 2	Level 3	Level 4	Level 5	Total N
N	1	3	71	593	1207	348	**2223**
Proportion	0.00	0.00	0.03	0.27	0.54	0.16	1.00
Value	-200	-122	-44	35	113	200	
Check	0	0	-1	9	61	31	100

Achievement Level on Post-test for students with Achievement Level 5 on pretest

	Low level 1	High Level 1	Level 2	Level 3	Level 4	Level 5	Total N
N	0	0	4	49	424	764	**1241**
Proportion	0.00	0.00	0.00	0.04	0.34	0.62	1.00
Value	-250	-173	-97	-20	56	133	
Check	0	0	0	-1	19	82	100

Table B.4 7th grade Value Tables calculations

Achievement Level on Post-test for students with Low Achievement Level 1 on pretest

	Low level 1	High Level 1	Level 2	Level 3	Level 4	Level 5	Total N
N	262	286	48	8	2	0	**606**
Proportion	0.43	0.47	0.08	0.01	0.00	0.00	1.00
Value	0	146	293	439	586	732	
Check	0	69	23	6	2	0	100

Achievement Level on Post-test for students with High Achievement Level 1 on pretest

	Low level 1	High Level 1	Level 2	Level 3	Level 4	Level 5	Total N
N	225	859	602	100	6	0	**1792**
Proportion	0.13	0.48	0.34	0.06	0.00	0.00	1.00
Value	-50	63	175	288	400	513	
Check	-6	30	59	16	1	0	100

Achievement Level on Post-test for students with Achievement Level 2 on pretest

	Low level 1	High Level 1	Level 2	Level 3	Level 4	Level 5	Total N
N	59	522	1291	686	60	2	**2620**
Proportion	0.02	0.20	0.49	0.26	0.02	0.00	1.00
Value	-100	-3	94	190	287	384	
Check	-2	-1	46	50	7	0	100

Achievement Level on Post-test for students with Achievement Level 3 on pretest

	Low level 1	High Level 1	Level 2	Level 3	Level 4	Level 5	Total N
N	4	56	591	1612	599	45	**2907**
Proportion	0.00	0.02	0.20	0.55	0.21	0.02	1.00
Value	-150	-66	17	101	184	268	
Check	0	-1	3	56	38	4	100

Achievement Level on Post-test for students with Achievement Level 4 on pretest

	Low level 1	High Level 1	Level 2	Level 3	Level 4	Level 5	Total N
N	0	6	29	579	1327	521	**2462**
Proportion	0.00	0.00	0.01	0.24	0.54	0.21	1.00
Value	-200	-125	-51	24	98	200	
Check	0	0	-1	6	53	42	100

Achievement Level on Post-test for students with Achievement Level 5 on pretest

	Low level 1	High Level 1	Level 2	Level 3	Level 4	Level 5	Total N
N	0	0	4	29	399	952	**1384**
Proportion	0.00	0.00	0.00	0.02	0.29	0.69	1.00
Value	-250	-175	-100	-25	50	125	
Check	0	0	0	-1	15	86	100

Table B.5 8th grade Value Tables calculations

Achievement Level on Post-test for students with Low Achievement Level 1 on pretest

	Low level 1	High Level 1	Level 2	Level 3	Level 4	Level 5	Total N
N	219	214	26	6	0	0	**465**
Proportion	0.47	0.46	0.06	0.01	0.00	0.00	1.00
Value	0	164	328	491	655	819	
Check	0	75	18	6	0	0	100

Achievement Level on Post-test for students with High Achievement Level 1 on pretest

	Low level 1	High Level 1	Level 2	Level 3	Level 4	Level 5	Total N
N	243	900	340	65	3	0	**1551**
Proportion	0.16	0.58	0.22	0.04	0.00	0.00	1.00
Value	-50	80	210	341	471	601	
Check	-8	47	46	14	1	0	100

Achievement Level on Post-test for students with Achievement Level 2 on pretest

	Low level 1	High Level 1	Level 2	Level 3	Level 4	Level 5	Total N
N	52	648	1097	500	24	1	**2322**
Proportion	0.02	0.28	0.47	0.22	0.01	0.00	1.00
Value	-100	5	109	214	318	423	
Check	-2	1	52	46	3	0	100

Achievement Level on Post-test for students with Achievement Level 3 on pretest

	Low level 1	High Level 1	Level 2	Level 3	Level 4	Level 5	Total N
N	5	141	828	1703	321	26	**3024**
Proportion	0.00	0.05	0.27	0.56	0.11	0.01	1.00
Value	-150	-59	32	123	214	304	
Check	0	-3	9	69	23	3	100

Achievement Level on Post-test for students with Achievement Level 4 on pretest

	Low level 1	High Level 1	Level 2	Level 3	Level 4	Level 5	Total N
N	0	7	71	829	1062	377	**2346**
Proportion	0.00	0.00	0.03	0.35	0.45	0.16	1.00
Value	-200	-120	-39	41	121	200	
Check	0	0	-1	14	55	32	100

Achievement Level on Post-test for students with Achievement Level 5 on pretest

	Low level 1	High Level 1	Level 2	Level 3	Level 4	Level 5	Total N
N	1	0	0	74	447	985	**1507**
Proportion	0.00	0.00	0.00	0.05	0.30	0.65	1.00
Value	-250	-174	-98	-22	54	130	
Check	0	0	0	-1	16	85	100