Teacher evaluation based on an aspect of classroom practice and on student achievement: A relational analysis between student learning objectives and value-added modeling

Author: Jiefang Hu

BOSTON COLLEGE

Lynch School of Education

Department of
Educational Research, Measurement, and Evaluation

TEACHER EVALUATION BASED ON AN ASPECT OF CLASSROOM PRACTICE AND ON STUDENT ACHIEVEMENT: A RELATIONAL ANALYSIS BETWEEN STUDENT LEARNING OBJECTIVES AND VALUE-ADDED MODELING

Dissertation
by

JIEFANG HU

submitted in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy

May 2015

TEACHER EVALUATION BASED ON AN ASPECT OF CLASSROOM PRACTICE AND ON STUDENT ACHIEVEMENT: A RELATIONAL ANALYSIS BETWEEN STUDENT LEARNING OBJECTIVES AND VALUE-ADDED MODELING

by

Jiefang Hu

Dr. Henry I. Braun, Dissertation Chair

ABSTRACT

With teachers being largely held accountable for student learning outcomes, it is of critical importance to identify effective and ineffective teachers through the development and implementation of a successful teacher evaluation system. Addressing the call to explore indicators of teacher effectiveness and enhance the traditional methods and practices of teacher evaluation, this study extends current efforts investigating different approaches to measuring teacher effectiveness through exploration of the relationship between two indicators of teacher effectiveness: the value-added estimates based on student test performance and growth, and the quality of student learning objectives (SLO) developed by teachers. It uses data from a large school district in North Carolina, comprising student achievement outcomes in mathematics and reading across five grades and three years. Different hierarchical linear models are employed to obtain teachers' VAM estimates with regression adjustments for prior years of achievement, student background characteristics, and teacher level covariate adjusted for each set of

models. Weighted Least Squares (WLS) analysis, logistic regression, and point-biserial analysis are used to examine the variations in the relationships among teachers' VAM estimates, SLO quality and SLO attainment status across years and grades. The HLM results revealed fluctuations in teachers' VAM rankings obtained at different stages of the model sequence that caused the correlations with SLO quality to vary as well. The WLS results indicated that the correlations between VAM and SLO quality also varied across years and grades. Further data analysis revealed generally weak associations between SLO quality and attainment status, as well as those between teachers' VAM estimates and whether their SLOs were achieved.

ACKNOWLEDGEMENTS

I have gained tremendous support from many people during this dissertation journey, without whom I would never have made it through. I would like to express my deepest gratitude to each and every one of them.

First of all, I am extremely thankful for my Committee Chair, Dr. Henry Braun, who devoted a lot of his time in guiding me through the entire process, teaching me how to think and write in a more effective and efficient way. This is the single most meaningful and valuable learning experience in my whole life.

I would also like to thank the other two Committee members, Dr. Larry Ludlow and Dr. Dennis Shirley who have provided great support in this process. Dr. Ludlow has been an incredible mentor in all these years, giving me kind help at every step of my path towards a Ph.D. degree. Dr. Shirley's knowledge and insight was essential to deepen my understanding and complete the dissertation.

There are other ERME faculty members to whom I would like to say thank you. Drs. Ina Mullis and Michael Martin granted me great support during the years I worked at TIMSS & PIRLS International Study Center. I am also grateful for Dr. Camelia Rosca for her generous support and practical advice on the dissertation data. My gratitude also goes to my colleagues at CTAC, Bill Slotnik, Gerry Harge, Barbara Helms, and Joe Frey who supported my access and understanding of the dataset.

Finally, the love of my family has been the spiritual support through my highs and lows during these years. My mother and father gave me endless trust while I was faced with difficulties. I owe many thanks to my lovely son and daughter, Zayne and Julaine – you don't know how many times I got more motivated by watching your sleeping faces at midnight. Of course, the most special thank you should go to my husband – I would not have accomplished this without your love and inspiration.

To all and those I didn't name – thank you for the help on my dissertation.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1. INTRODUCTION

1.1 Description of the Problem

Student learning is fundamental to the mission of education and, therefore, has been deemed to be the focus of most educational institutions. A variety of factors contribute to student learning, and generally those associated with schools and classrooms are believed to be essential. A number of recent studies argue that access to an effective teacher is the single most important school-related factor responsible for incremental student learning (Thum, 2003; Rivkin, Hanushek, & Kain, 2005; Haycock, 1998; Jordan, Mendro, & Weerasinghe, 1997; Sanders & Horn, 1998, 1995). During the last decade, holding schools and teachers accountable for student learning outcomes has gained unprecedented prevalence in K-12 education[1].

Teachers demonstrate differential effectiveness in fostering student progress. Teaching quality matters with respect to student learning outcomes as measured by test performance. Jordan, Mendro, and Weerasinghe (1997) investigated teacher ranking estimates from the Dallas Classroom Effectiveness Indices model for three years to identify effective and ineffective teachers, and pointed out that a few years with effective teachers can put even the most disadvantaged students on the path to college. On the other hand, a few years with ineffective teachers can deal students an academic blow

[1] As of January 2012, most of the 46 states including the Districts of Columbia that have adopted the Common Core State Standards (CCSS) are changing their teacher preparation and evaluation systems (Center of Education Policy, 2012).

from which they may never recover. Clearly, such substantial differences in teacher effectiveness have meaningful consequences for student performance and growth. Some investigators have asserted that a student who is taught by an ineffective teacher for 2 years in a row can never recover the learning lost during those years (Sanders, 2000; Webster, Mendro, Orsak, and Weerasinghe, 1996, 1998).

Since teaching quality plays such an important role in student learning and academic progress, identifying effective and ineffective teachers is of critical importance. A successful teacher evaluation system, therefore, is called for to perform a key duty in advancing student learning: Teacher evaluations, appropriately designed and implemented, should identify and evaluate the instructional strategies, professional behaviors, and delivery of content knowledge that affect student learning (Danielson & McGreal, 2000; Shinkfield & Stufflebeam, 1995).

However, the reality of teacher evaluation in public school districts nationwide has been disappointing. Traditional classroom observations, as nearly the only source of evidence regarding measuring teacher effectiveness, have failed to distinguish great teaching from good, good from fair, and fair from poor. A teacher's effectiveness—the most important factor for schools in improving student achievement—is not measured, recorded, or used to inform decision-making in any meaningful way (Weisberg et al, 2009).

Moreover, the extensive research on teacher quality in recent years has concluded that there are large and significant differences among teachers with respect to their capacity to promote student achievement. However, these differences are not well captured by common measures of teacher qualifications (Schacter & Thum, 2003;

Hanushek, 2003). The typical methods and practices of teacher evaluation currently employed are based on simplistic criteria with marginal relevance to what teachers need to perform to enhance student learning (Danielson & McGreal, 2000), and therefore are characterized as inaccurate, unsupportive (Peterson, 1995), superficial (Stiggens & Duke, 1988), and of low reliability and validity (Darling-Hammond, Wise, and Pease, 1983).

Discussions regarding the characteristics and indicators of teacher quality that can be utilized to measure teacher effectiveness are ongoing. Only recently have some states and districts begun to develop more useful systems for evaluating teacher performance and providing teachers with the feedback they need to improve their practice (Bill and Melinda Gates Foundation, 2012). With the increasing amount of attention directed toward teacher evaluation, there have been a number of studies investigating a range of teacher-relevant factors that may influence student learning. For instance, Darling-Hammond (2000) indicated that the variables presumed to be indicative of teachers' competence which have been examined for their relationship to student learning include measures of academic ability, years of education, years of teaching experience, measures of subject matter and teaching knowledge, certification status, and teaching behaviors in the classroom. This assertion was re-emphasized in her latest book (Darling-Hammond, 2010).

In recent years, with the increasing availability of longitudinal student achievement data, researchers and policy makers have started to explore more scientific ways to quantify the heterogeneity in students' test score trajectories and how to use this rich data to measure a key aspect of teacher effectiveness. Value-added models, the complex statistical models that attempt to attribute some fraction of student progress to

their teachers based on those trajectories (National Research Council, 2011), are intensively studied and widely employed to examine the effectiveness of teachers in facilitating students' progress in their academic achievement. Currently, many states and districts have begun to adopt the VAM approach for their teacher evaluation, for example, nearly one fourth of the 65 member districts of the Council of Great City Schools have implemented some form of value-added based school or teacher rewards program (Hill, Kapitula & Umland, 2011)

Value-added models (VAMs) hold out the promise of isolating the effects of teachers or schools from that of other factors such as family background, poverty or school leadership. Employing a collection of complex statistical techniques to analyze multiple years of students' test score data, VAMs attempt to quantify the extent to which changes in student performance can be attributed to the effect of students attending the class of a particular teacher or school rather than another. The estimates of relative effectiveness derived from a value-added analysis can be compared to one another or to that of the typical teacher or school. The VAM approach for estimating teacher effectiveness is an increasingly popular, as well as a controversial education reform policy, and has garnered a great deal of attention among both policymakers and researchers (Kane, Rockoff & Staiger, 2008; Jacob & Lefgren, 2008; Goldhaber & Hansen, 2010). When compared to performance assessment in other fields or to evaluations of teachers based on other sources of information, VAM looks respectable and can still provide the best signal for measuring the effectiveness of teachers in improving student learning outcomes (Glazerman et al, 2010).

However, the inferences one can make from the VAM estimates still raise concerns as researchers hope to link student learning outcomes with teacher effectiveness in this era of accountability. As Braun (2005) noted, causal attributions cannot be confidently made about the quality of teaching due to the lack of randomization – no matter how complex the statistical model is and how sophisticated the method of analysis is. There could be many other unmeasured attributes associated with the results from VAM models, which, when used in high-stakes situations, can bring unintended negative consequences. Therefore, the results from student learning outcomes should be properly used so as "to inform decision making and improve teaching and learning", rather than only with high-stakes accountability purposes (Kuh & Ikenberry, 2009, p.4).

As teacher evaluation attracts more and more attention, other approaches to measuring teacher effectiveness are emerging. Another popular approach to measuring teacher performance is called Student Learning Objectives (SLOs), which started to be tied to high-stakes decisions in Denver in 1999, when a performance-pay system was piloted in the Denver Public Schools that required teachers of both tested and non-tested grades and subjects to set "student growth objectives" (CTAC, 2001). Currently SLOs have been broadly known for their use in measuring teacher effectiveness for the non-tested subjects and grades, and many states and school districts take SLOs as the solution of choice to the challenge of integrating teachers of non-tested grades and subjects into the overall evaluation and compensation systems that require measuring student growth (Reform Support Network, 2012). The Race to the Top (RTTT) Technical Assistance (TA) Network defines SLOs as: "A participatory method of setting measurable goals, or objectives, based on the specific assignment or class, such as the students taught, the

subject matter taught, the baseline performance of the students, and the measurable gain in student performance during the course of instruction (RTT Technical Assistance, 2010)".

With SLOs, teachers establish learning objectives for individual students, the class as a whole or particular target student groups based on their knowledge of the students and their instructional plans. Once the learning objectives are created for students, the extent to which these objectives are achieved during a particular learning period can be evaluated. SLOs can be implemented with a variety of assessment formats, such as nationwide standardized tests, state or district assessments, and even teacher-developed measures[2]. A key advantage of the SLO approach over traditional test-centered approaches to accountability is the active involvement of a teacher. SLOs are designed to reflect and incentivize good teaching practices such as setting clear learning targets, differentiating instruction for students, monitoring students' progress toward these targets, and evaluating the extent to which students have met the targets (Marion et al, 2012). Therefore, SLOs should have instructional value as well as assessment value.

Professionals have different strengths in their own disciplines. In the field of teaching, some teachers have extraordinary success in fostering student success on assessments, while others are more skillful at daily classroom teaching practice, creating more vigorous and dynamic academic environment for students, and cultivating their interests. Hence, the extent to which different approaches to measuring teacher

[2] For example, New York State Department of Education requires K-2 teachers must use one of the assessment options for SLOs: State approved 3rd party assessment; District, regional, or BOCES-developed assessment; School-or BOCES-wide, group, or team results based on state assessments; Teachers at 3rd grade must use 3rd grade state assessment (ELA and math), and teachers of 4-8 grade must use State provided growth SGP/VA.

effectiveness correspond to one other is an important question. Answering this question can provide insights to better understand teacher effectiveness and help improve the design and implementation of teacher evaluation in the future.

A number of studies have investigated a range of teacher-linked factors that may influence student learning. Variables presumed to be indicative of teachers' competence are generally regarded as indicators of the effectiveness of teachers, which may comprise assorted aspects of teachers' characteristics. Two types of indicators of teacher effectiveness are investigated in this study: (i) VAM estimates, an indicator of teacher effectiveness based on student achievement; and (ii) SLO quality, an indicator of teacher effectiveness based on one aspect of classroom practice.

A common approach to SLOs is to treat them as the goals that students are expected to attain within a certain learning period of time, and the extent to which they are achieved can be used as an indicator of teacher effectiveness. Denver, Colorado and Charlotte-Mecklenburg, North Carolina are at the forefront of this approach (Buckley & Marion, 2011). This study focuses on one aspect of SLOs and employ a different approach to understanding and analyzing SLOs; that is, teachers have been required to develop the SLOs for their students, either individually or focusing on a group of students or the class as a whole. The quality of these objectives for student learning provided by the teachers have been evaluated and accordingly is considered as an indicator of the effectiveness of these teachers, reflecting an aspect of the attribute of their classroom practice. In other words, SLOs are treated as the written objectives created by teachers, and the quality of these written SLOs is used as a proxy for teacher quality in this study.

Employing a dataset from a large school district in North Carolina, this dissertation examines the extent to which teachers' SLOs quality scores are related to the estimated teacher success in contributing to student achievement, as indicated by value-added model estimates. Analyses in this study involve generating a series of multi-level value-added models, structured by subject and grade level, to obtain the value-added estimates of teacher effectiveness based on their students' test score trajectories with contextual characteristics accounted for. As a practice-based indicator of teacher effectiveness, the SLOs quality scores are correlated with the VAM estimates in further analyses. In addition, an indicator of whether the SLOs have been achieved is also examined in conjunction with SLOs quality scores. The relationship between test-based value-added estimates of teacher effectiveness and practice-based estimates of teacher effectiveness are subsequently compared and contrasted across different models and settings. Results are also be aggregated by elementary and middle school levels.

1.2 Purpose and Research Questions

To help bring clarity, this study explores the problems and uncertainties in measuring teacher effectiveness by focusing on value-added methodology issues and on the quality of teacher developed SLOs. Here, the purpose of value-added modeling is to estimate relative teacher effectiveness with respect to their students' progress on test-based outcomes, while taking into account students' prior achievement and other associated factors (e.g. background characteristics) – both at the student and the teacher level. On the other hand, the quality of teacher-developed SLOs is employed to provide the estimates of teachers' effectiveness based on their classroom practices.

The framework for the analyses proposed for this study comprises six research questions concerning the relationships between different approaches to measuring teacher effectiveness. Although the answers to these questions are of great interest in their own right, more broadly they offer insights into improved design of accountability systems. The research questions are:

1. How do rankings based on value-added estimates of teacher effectiveness compare (by grade and subject) to the rankings derived from the practice-based estimates of teacher effectiveness based on SLOs quality scores?
2. To what extent do the student's contextual characteristics impact these relationships?
3. To what extent are the relationships between value-added estimates of teacher effectiveness and practice-based estimates of teacher effectiveness affected by teacher-level characteristics?
4. To what extent do the relationships between value-added estimates of teacher effectiveness and practice-based estimates of teacher effectiveness vary by grade, by year, by type of school?
5. To what extent do the SLOs quality scores correspond to whether the SLOs have been achieved?
6. To what extent do teachers' VAM estimates agree with the achievement status of the SLOs?

The first three research questions aim to identify the student and teacher characteristics that may have an impact on the relationship between the different indicators of teacher effectiveness. The findings help to ensure the credibility of

employing teacher value-added estimates as an effectiveness indicator in the response to the fourth research question, where further comparisons with the indicator of teacher effectiveness based on an aspect of classroom practice are carried out. The last two research questions concern the credibility of using SLOs as measures of teacher effectiveness. Moreover, issues associated with using student learning objectives as an indicator of teacher effectiveness in high-stakes settings, as well as the implications of using assessment-as-accountability measures for educators and policy makers are discussed.

1.3 Significance of the Study

Because of the key role teachers play in improving student academic performance and the diverse problems with most teacher evaluation systems nationwide and internationally, there are many important issues that are yet to be addressed. For one, the impact and validity of various approaches to measuring teacher effectiveness need be further investigated. In fact, as emphasized by the NRC report (2011) and the Measures of Effective Teaching (MET) study (Bill & Melinda Gates Foundation, 2012), validation of value-added estimates of teacher effectiveness remains an important area of research. The proposed study seeks to examine the relationship between different approaches to estimating teacher effectiveness, which could produce one type of validity evidence[3] for using the value-added approach, as well as SLOs approach, to measuring teacher effectiveness. Ideally, this can help provide criterion-related validity evidence for the use of the performance evaluation scores as the basis for a performance-based pay system or

[3] Strictly speaking this is not a validity study of VAM or SLO quality scores as we couldn't suggest whether VAM or SLO should be used.

other decisions with consequences for teachers (Milanowski, 2004). Alternately, the findings could indicate significant problems with VAM and SLOs that policy makers should address before continuing with these assessment approaches.

Secondly, this study also helps examine other approaches to teacher evaluation, such as the validity of using SLOs for measuring teacher effectiveness. Through investigating the extent to which the SLOs quality scores may correspond to the status of whether the SLOs can be achieved by students, the analyses reveals the association between teachers' ability to develop the written SLOs and their efficacy in helping students achieve those SLOs. As such, this dissertation provides evidence for employing the SLOs quality as indicators of teacher effectiveness.

Furthermore, results in this study also provides evidence as to the extent to which different approaches to value-added models in estimating teacher effectiveness correspond to one other, as well as how they are correlated to teacher classroom practice and performance. These findings also are likely to be useful to policy makers and can contribute to the growing research and debates on the reliability and validity of different kinds of VAM.

There is an urgent call for evidence regarding the validity of the different approaches to measuring teacher effectiveness. This research is highly relevant and topical in many nations that are working on improving teacher performance and appraisal. This research enables researchers and policy makers to better understand the relationship between teachers' proficiency in one aspect of classroom practice and their effectiveness in fostering student academic achievement. These findings can, in turn,

inform the design and implementation of sound teacher evaluation systems as well as correct errors that may exist in current performance appraisals.

1.4 Outline of the Dissertation

After an introduction to the purpose and key research questions of the dissertation in this first chapter, chapter 2 reviews literature in the field of teacher evaluation employing three perspectives: (1) the importance and impact of teacher quality; (2) the traditional approaches and current efforts in the field of teacher evaluation; and (3) indicators of teacher effectiveness.

Chapter 3 introduces the proposed methods of analysis for this study. Data used in this study are described and the methods that are employed to construct the variables are presented. Specifically, this chapter describes some widely used approaches to the value-added modeling. The statistical models for each research question are presented in detail. Finally, this chapter discusses the integrity of the research design and the limitations of the results based on the research design and data collected.

In chapter 4, results from the empirical data analyses are provided. Beginning with a detailed description of the analysis sample, the chapter then delineates the results of the multi-level value-added statistical models. The preliminary results include those from the stages of variable construction, descriptive analyses of distributions and patterns on the SLOs quality variable, as well as other variables of interest. Results from each value-added model along with the corresponding correlational analyses with SLOs quality, are then presented. Further results from the subsequent Weighted Least Squares analyses, logistic regression analyses, as well as the point-biserial analyses are discussed.

Finally, chapter 5 reviews findings and their implications. The dissertation concludes with a discussion of limitations of the study and possible new research directions in the field.

CHAPTER 2. LITERATURE REVIEW

This chapter provides an overview of current research in measuring teacher effectiveness and demonstrates how this study relates to the broader issues in this field. The goal of this review of the literature is to demonstrate the significance of this dissertation by examining key aspects of the study. In this regard, this chapter is organized into three major sections.

First, a number of studies are discussed to document the importance of teacher quality. The differential impact of high- and low-quality teaching, as well as other factors that influence teaching and learning, are presented. The second section emphasizes the content of teacher evaluations, with a discussion of the traditional and current approaches to measuring teacher effectiveness. The third section focuses on a set of widely available indicators of teacher effectiveness based on student learning outcomes and teachers' classroom practices, respectively. In particular, value-added analysis is highlighted as a method of measuring teacher effectiveness using student achievement growth, and Student Learning Objectives (SLOs) are introduced as a practice-based measure of teacher effectiveness.

2.1 Importance and Impact of Teacher Quality

2.1.1 Importance of Teacher Quality

Teacher quality has been seen as the crucial driving force for improving student achievement and thus promoting a nation's economic competitiveness in the global

society. The National Academies (2007) addressed the importance of teacher quality in the study of Teacher Preparation Programs: “Teacher quality is widely recognized by policymakers, practitioners, and researchers alike to be the most powerful school-related influence on a child’s academic performance.” As student learning is a collective responsibility and is influenced by a variety of factors such as school resources and environment, peer interaction and classroom climate, teachers’ contributions to student learning outcomes, when being evaluated, need to be isolated from those of other factors.

Although the importance and impact of teacher quality on student achievement is nearly universally acknowledged, the construct itself has been defined and measured in many ways since the criteria for defining teacher quality vary from person to person, from one community to another, and from one era to the next (National Research Council, 2001). Teacher quality is a broad construct that involves various aspects of teachers’ characteristics, such as knowledge, skills, abilities, and dispositions. In addition to the contributions to student academic achievement and socio-emotional development, teacher quality may also be signaled by its ability to create a positive classroom environment that fosters and stimulates student learning. Besides, teacher quality has largely been measured by more distal characteristics such as teaching experience and formal qualifications, professional attitudes, skills in mentoring new teachers, and constructive cooperation with other staff. Further, due to the complexity of classroom dynamics and the role of exogenous factors, researchers have developed several different approaches to document and evaluate teaching practices, which include analysis of teacher assignments and student work, student evaluation of teachers, classroom

observations and videotapes of classroom observations, as well as combinations of these methods (Paek, Braun, Trapani, Ponte & Powers, 2010).

Although each of these characteristics could indicate a particular aspect of teacher quality, it might be difficult to achieve a uniform consensus on how to define teacher quality in a more comprehensive and unified way. For different purposes, teacher quality will likely need to be defined differently. For example, when used for making an initial hiring decision, in granting tenure, rewarding excellent performance or identifying and supporting struggling teachers, key aspects of teacher quality may well differ (Goe, 2007).

When teacher effects are compared to other factors, a number of important assumptions have been made and need to be articulated explicitly. First, the class effect is assumed to be causal; that is, the differences in student learning outcomes among students from different classes, if any, are completely attributed to the students being in one class rather than another, although there might be other factors involved, such as the differential levels of parental involvement and other types of external encouragement and support that may influence motivation and engagement. Second, the teacher effect is assumed to constitute the class effect; that is, teachers are entirely responsible for the differences in results among classrooms. These assumptions about teacher effect are also implicit in the present study.

Researchers have established the fact that teachers have a measurable effect on student learning and that teachers matter more than any other school-related factors with regard to student learning outcomes (National Academy of Education, 2005; Hanushek, 2011; National Research Council, 2010; Bill & Melinda Gates Foundation, 2012; RAND

Education, 2012). However, it is much more difficult to identify the specific characteristics of teachers or aspects of their pedagogy that can be linked to higher student achievement (Olson, 2003). Various approaches have been used to measure teacher quality and different conclusions regarding the variations of teacher quality are made. On the one hand, distributions of teacher quality scores based on teacher attributes and demographic characteristics tend to show very little variability. For example, studies typically found that less than 10 percent of the variation in teacher effectiveness can be attributed to readily observable credentials like degree and experience levels (e.g. Aaronson et al. 2007; Goldhaber et al. 2000). On the other hand, large variations in teacher quality are likely to be demonstrated while using student learning outcomes to measure teacher effectiveness. Studies have indicated that a variety of teacher characteristics are associated with student learning outcomes, such as the selectivity of the teacher preparation program (Rice, 2003), the general teacher aptitude including the cognitive ability demonstrated by intelligence test scores (Rockoff, Jacob, Kane, & Staiger, 2011), the verbal skills from vocabulary or word tests (Wayne & Youngs, 2003), and the scores on certification and licensure tests (Goldhaber & Anthony, 2007; Clotfelter, Ladd, and Vigdor, 2007).

While student achievement has been broadly used to measure teacher effectiveness, teacher quality can also be assessed through rubrics that do not involve student test scores but are based on norms of professional practice, such as the Danielson's framework or the National Board Professional Teaching Standards (NBPTS) framework (Danielson, 2013; NBPTS, 2014). The extent to which the rankings of teachers based on these standards correlate with those based on student achievement

growth remains an open question. Goldhaber and Anthony (2005) studied the relationship between the certification of teachers by NBPTS and elementary-level student achievement, and concluded with consistent evidence that NBPTS can identify more effective teachers and that teachers with National Board Certifications are generally more effective than those who never applied to the program. A similar study from Goldhaber and Brewer (2000) also found that mathematics teachers with a regular subject certification have a statistically significantly positive impact on student mathematics test scores than those who are not certified in their subject area.

In general, the extensive research on teacher quality in recent years has concluded that there are large and significant differences among teachers with respect to their capacity to improve student achievement that are not well captured by more commonly used measures of teacher qualifications (Schacter & Thum, 2003; Hanushek, 2003).

2.1.2 Impact of High- and Low- Quality Teaching

The concept of effectiveness is usually defined as the capability of producing a desired result, which suggests that measuring teacher effectiveness requires examining the quality of the main results of teaching – student learning. Therefore, educators and researchers need to agree on the desired outcomes of student learning and how to measure them well. One important aspect of student learning that can be relatively easily measured is academic achievement as evaluated by test performance, though the limitations regarding using tests to measure teacher effectiveness need be noted; that is, whether the valued learning objectives can be measured well by tests and how the

contributions of teachers should be isolated from those of other factors that influence student learning.

Apparently, differences in teacher effectiveness can have meaningful consequences for student progress in academic achievement. However, certain ways of defining teacher effectiveness have raised serious issues and concerns. In many studies, teacher quality and effectiveness is defined and measured by the magnitude of improvement in student test scores. In other words, differences in student learning outcomes determine, by definition, teacher effectiveness (Kuppermintz, 2003). These studies divided teachers into different "effectiveness" groups based on the rankings of their average student gains and obtained a variety of "findings" claiming that teacher effectiveness is the cause of student achievement gains. This type of reasoning is tautological, while the real interest regarding teacher effectiveness and student learning lies in the extent to which the difference in student performance can be expected from teachers at different percentiles in the effectiveness distribution. The effectiveness of teachers can be defined independent of student learning outcomes as well. For example, Weimer (2013) studied ways to define teacher effectiveness by collecting people's opinions, and found the most important abilities agreed by teacher, students, and administrators are – "cultivate thinking skills, stimulate interest in the subject, and motivate students to learn."

A variety of studies have demonstrated the magnitude of estimated differences in teacher effectiveness to be quite impressive. Using teacher ranking estimates computed from the student residuals of the Dallas Classroom Effectiveness Indices model, Jordan, Mendro, and Weerasinghe (1997) examined the effective and ineffective teachers in a 3-

year period and pointed out that a few years with effective teachers can put even the most disadvantaged students back on the pathway to college, while, conversely, a few years with ineffective teachers can deal students a major academic blow from which they may never recover. More recent studies have also replicated the findings regarding the impact of teacher effectiveness on student learning outcomes, for example, Clotfelter, Ladd & Vigdor (2007) explored the relationship between teacher quality and student achievement extensively using a dataset with 10 years of records, and concluded that a teacher's characteristics as well as credentials, including experience, test scores and regular licensure, all exhibit positive and large effects on student achievement. The effects are larger for mathematics than for reading, and can be comparable to those of changes in class size and to the effects of socio-economic characteristics of students, such as those measured by the educational levels of their parents. It is worth noting that part of the results of this study was disputed (Goldhaber & Anthony, 2007), which found the effects of the National Board for Professional Teacher Standards (NBPTS) certification process was not related to teacher effectiveness.

Rockoff (2004) used panel data to estimate teacher fixed effects from linear regressions of test scores while controlling for fixed student characteristics and classroom specific variables, and found consistently large and statistically significant differences among teachers. A one standard deviation increase in teacher quality, comprised of observable and unobservable characteristics such as teacher experience and highest education levels, is associated with the increase of student test scores by approximately .20 standard deviations in reading and .24 standard deviations in mathematics on

nationally standardized distributions of achievement. In particular, teaching experience is found to significantly raise student test scores in reading subject area.

Convincing evidence that teacher quality is strongly associated with student achievement has been documented by several other studies as well. Nye, Konstantopoulos, and Hedges (2004) applied a hierarchical linear model to sort out the between-class effects on student achievement gains as well as on achievement status and concluded that the teacher effects[4], consistent with findings of other studies, are substantial and are larger for mathematics than for reading. The estimated between-teacher variance components for reading is about half the size for mathematics.

Hanushek et al (2005) studied the matched data of students in Grades 3–8 and their classroom teachers in a single Texas district, and produced the lower bound estimates of the variance in teacher quality since the study was entirely based on within-school heterogeneity, and there was no control for school level factors such as the effectiveness of school principals or the composition of the students. The authors found a one standard deviation in teacher quality, comprised of background characteristics such as experience and degree, is associated with a 0.22 to 0.32 standard deviation difference in achievement gains. Results from the study also suggested that the effects of a costly ten student reduction in class size are smaller than the benefit of moving one standard deviation up the teacher quality distribution.

[4] The study (Konstantopoulos, and Hedges. 2004) mainly documented the between-classroom variations, and attributed the differences to teacher effects.

2.1.3 Compared to Other Factors Teacher Quality Matters the Most

A great deal of research has explored nearly all factors that may have a statistical association with student learning, including individual student factors, classroom effects, and school level characteristics. Many researchers have concurred that teacher effectiveness is still the most critical school-related factor in terms of the influence on student learning and can be regarded a powerful predictor of student performance. For example, Rockoff (2004), Rivkin et al (2005), Aaronson et al (2007) all contended that the single most crucial factor affecting student achievement is from teachers, and the effects of teachers on student achievement are both additive and cumulative. Further, they believe that lower achieving students are the most likely to benefit from the increases in teacher effectiveness. Likewise, Hanushek (1992) analyzed the relationship between teacher effects and student achievement in his study and claimed that the influence of teacher quality on their students' annual achievement can be more than one grade-level equivalent in test performance.

Rivkin et al (2001) used fixed effects and a value-added framework for variance decomposition in their analysis of the extraordinarily rich data set for student achievement in Texas, with the very large samples of over 3000 schools and a half million students. They believed that school quality matters for student achievement, and variations among teachers within schools dominate school quality differences. In contrast, class size, teacher education, and teacher experience appear to play only a small role. This portion of the findings conflicts with the study of Clotfelter, Ladd & Vigdor (2007), which concluded that a teacher's characteristics as well as credentials, including

experience, test scores and regular licensure, all exhibit positive and large effects on student achievement.

Recent research has shown that teacher and classroom effects on student learning carry the largest weight in the education system (Goldhaber, 2007). The comparative analysis of teacher effects with other factors on student learning is implemented in the study by Sanders (2000), where teacher effects are shown to have greater statistical association with student learning than class size, spending differences and several other factors. In the analysis of teacher preparation and student achievement across states, Darling-Hammond (2000) argues that teacher quality is more strongly related to student achievement than other factors such as class size, overall spending on education, and teacher salaries.

In her later book, Darling-Hammond (2010) re-emphasized this assertion. After reviewing a number of studies, she concluded that the differences in student achievement associated with teacher qualifications (characterized by certification, preparation, license test scores, degree, and teaching experience) are larger than the average differences attributed to race or socioeconomic status (i.e. difference between a White student with college-educated parents and a Black student with high school-educated parents). She specifically stated that improving teacher quality can reduce the achievement gap between the schools serving the poorest and most affluent student bodies by 25%. In fact, students' achievement was hurt most by having an inexperienced teacher on a temporary license.

In comparing teacher and school effects, a few studies find that the variations in student achievement among classrooms within the same schools are actually larger than the variation among schools (Hanushek, Kain, & Rivkin, 1998; Meyer, 2001; Webster et al., 1996). Webster, Mendro, Orsak, and Weerasinghe (1996), in another study of school and teacher effects in the Dallas Public Schools, concluded that a school's effect could essentially represent an aggregation of the individual effects of its teachers.

To isolate the contributions of teachers to student learning outcomes from those of other factors, various statistical techniques have been employed to evaluate the variances from different levels of a model. Some studies found that teacher level variance far exceeds grade-level, school-level and district-level variance (Marzano, 2003; Rivkin et al., 2001; Thum, 2003; Sanders & Horn, 1995). Meyer (2001) also asserted from a study of the Denver Public Schools that the teacher-related variables account for more than twice the total variation in student test score changes than do the school-related variables.

In summary, recent research has concluded that teachers are a critical determinant of student achievement and have substantial impact on student learning. In a study of the Cincinnati school district, Milanowski (2004) pointed out that the teacher evaluation scores from a rigorous teacher evaluation system can be positively related to student achievement gains and provide criterion-related validity evidence for the use of the performance evaluation scores as the basis for a performance-based pay system or other decisions with consequences for teachers. Darling-Hammond (2007), while discussing various reform efforts to improve schools and the outcomes of education, indicated that an important lesson is that teachers are the fulcrum that determines whether any school initiative will result in success or failure. She contended that nearly every aspect of

school reform that is aimed to improve student learning, after all, depends on the efforts of highly-skilled teachers to implement new strategies related to improved curricula or assessment. These accumulated findings indicate the appropriateness and indeed, the imperative, of continued research into the best ways of appraising and improving teacher quality.

2.2 Teacher Evaluation: Traditional Approaches and Current Efforts

2.2.1 Traditional Approaches to Teacher Evaluation

It has been broadly agreed that a teacher evaluation system, when used appropriately, should identify and measure the instructional strategies, professional behaviors, and delivery of content knowledge that affect student learning (Danielson & McGreal, 2000). In addition, well-designed teacher evaluation programs could have a direct and lasting effect on individual teacher performance given that teachers, through more conversations with colleagues and administrators about effective practices, could gain information and feedback from the evaluation program and thereof become generally more self-reflective. Taylor & Tyler (2012) studied a sample of midcareer elementary and middle school teachers in the Cincinnati Public Schools who were evaluated based on a yearlong classroom observation program. The authors found that teachers are more effective at raising student achievement during the school year when they are being evaluated than they were previously, and even more effective in the years after evaluation.

Other studies that investigated schools using the Teacher Advancement Program

(TAP) based on NBPTS and INTASC, as well as standards based assessment rubrics developed in Connecticut (Bill & Melinda Gates Foundation, 2010; Rothstein, 2011), found that the indicators of good teaching are practices associated with desired student outcomes. TAP teachers said this system, along with the intensive professional development offered, is substantially responsible for improving their practice and for student achievement gains in many TAP schools (Solmon, White, Cohen, & Woo, 2007).

Darling-Hammond et al (2012) reviewed different approaches to evaluating teachers and concluded that standards-based evaluation processes, like the National Board Certification and Performance Assessments for beginning teacher licensing[5] as well as district and school-level instruments based on professional teaching standards, have been found to be predictive of student learning gains and productive for teacher learning. Ideally, teacher evaluation can support accurate information about teachers, helpful feedback, well-grounded personnel decisions and be a useful part of a constantly improving teaching and learning system.

However, the reality of teacher evaluation in public school districts nationwide is generally disappointing. Traditional classroom observations, as nearly the only source of evidence regarding measuring teacher effectiveness, have failed to distinguish the teacher performance such as great teaching from good, good from fair, and fair from poor. A teacher's effectiveness—the most important factor for schools in improving student achievement—has not been measured, recorded, or used to inform decision-making in any meaningful way (Weisberg et al, 2009).

[5] American Association of Colleges for Teacher Education (AACTE) presented a new preservice teacher performance assessment, edTPA (2013); however, no study of the relationship between this type of assessment and others is not yet available.

In most public school districts, individual teachers receive little feedback on the work they do and teacher evaluation becomes an obligatory but perfunctory exercise. In too many schools principals go through the motions of visiting classrooms with a checklist in hand. In the end, virtually all teachers receive the same "satisfactory" rating (Bill and Melinda Gates Foundation, 2010). According to the recent extensively quoted "The widget effect" report from the New Teacher Project which surveyed over 15,000 teachers in 12 large school districts and 4 states, teacher evaluation systems are unsuccessful in differentiating performance among teachers. Most teacher evaluations are based on only two or fewer classroom observations, each 60 minutes or less, and they are even conducted by administrators without any extensive training. Evidently, such an evaluation system cannot reflect much variation among the teachers. Therefore, not surprisingly, a majority of teachers were highly evaluated. For example, in Denver schools that did not make adequate yearly progress (AYP), more than 98 percent of tenured teachers received the highest rating—"satisfactory." Peterson (2000) concluded from his review of the literature that the present teacher evaluation practices neither improve teachers nor accurately represent what happens in the classroom.

2.2.1.1 Problems and Consequences of Traditional Approaches

Every classroom should have a well-educated, knowledgeable, skilled and compassionate teacher. For that to happen, school systems should conduct teacher evaluation in a fair and systematic way so that effective teachers can be retained, those with remediable shortcomings be further guided and trained, and ineffective teachers who do not improve should be removed. However, in practice, traditional evaluation programs are often seen as perfunctory, unreliable, and insufficient to provide incentives to

improve teacher performance, while adding administrative burdens (Halverson, Kelley & Kimball, 2004). A recent study found that under current evaluation systems, American public schools generally fall short in efforts to improve the performance of less effective teachers, and failing that, of removing them (Baker et al, 2010). Chait (2010) explored the barriers to remove chronically ineffective teachers, and concluded the reasons why teacher dismissal is rarely pursued: the weak teacher evaluation practices or systems, the time and cost of dismissal cases, the difficulty of winning cases, a school culture that is uncomfortable differentiating among teachers, and the difficulty of hiring replacements in some districts.

In traditional teacher evaluation, educators relied on the observations and judgment of teacher performance in classroom; these methods are generally deemed to be of low reliability and validity (Darling-Hammond, 2008; Danielson & McGreal, 2000). Medley and Coker (1987) reviewed studies from the 1950s to 1970s and reckoned the relationship between a principal's ratings of teacher performance and student achievement as being generally weak. Their own study presumes the correlation between principal performance ratings and teacher effectiveness, which is estimated from students' pretest and posttest scores obtained at the beginning and the end of the same school year, to be quite low, in the range of 0.10 to 0.23.

In particular, many researchers believe that the traditional methods and practices of teacher evaluation are based on simplistic criteria with minimal relevance to the pedagogical practices that enhance student learning (Danielson & McGreal, 2000). They have been characterized, therefore, as inaccurate, unsupportive (Peterson, 1995), and superficial (Stiggens & Duke, 1988). Among other criticisms, teacher evaluation systems

have been discredited for lack of teacher buy-in and minimal district or school level commitment. Scholars contended that they have been based on criteria predicated upon narrow conceptions of teaching, inadequate feedback, and perceived subjectivity (Glazerman et al., 2011; Strong & Tucker, 1999; Johnson, 1997).

Teacher evaluation has frequently been used to weed out the poorest performing teachers rather than to hold all teachers accountable or to improve implementing teacher evaluation systems performance of all teachers (Darling-Hammond et al., 2009). Because of these constraints, teacher evaluation has had a limited impact on teacher performance and development (Peterson, 1995; Darling-Hammond, Wise & Pease, 1983).

One other common criticism of the current teacher accountability systems is that teacher seniority and credentials are often considered to be independent as well as important factors for teacher evaluation and compensation. Recent research suggests that over the first several years of practice teacher effectiveness does improve; however, it tends to flatten out after seven to ten years. Further, with the exception of degrees in mathematics or the sciences, teachers' additional educational credentials appear to be only weakly related to their students' test performance (Goldhaber, 2008).

For a long time, states and school districts have attempted to structure teacher evaluation practices to promote teacher accountability and improvement in practice or both (Peterson, 1982). One purpose of educational accountability could be not only to hold teachers responsible for student learning outcomes but also to contribute to the improvement of practice; however, there are often tensions and even direct conflicts between the two purposes of improvement and accountability. These problems could

probably be mitigated by collaborative involvement in data collection and analysis, collective responsibility for improvement, and a consensus on the accurate, meaningful, fair, broad and balanced indicators and metrics (Hargreaves and Braun, 2013).

However, traditional evaluation systems and repeated reforms appear to have done little to enhance either accountability or practice (Glazerman et al., 2011; Peterson, 1995; Joint Committee on Standards for Educational Evaluation, 1988). Principals typically have too little time and training to get prepared for adequately completing the job of assessing and supporting teachers. In many school districts nearly all teachers are judged to perform satisfactorily. However, a number of statistical analyses of large datasets confirm the long-held intuition of most teachers, students, and parents: teachers do vary substantially in their ability to promote student achievement growth. The ubiquity of "satisfactory"[6] ratings stands in contrast to a rapidly growing body of research that reveals differences in teachers' effectiveness at raising student achievement (Kane et al, 2011).

One of the other difficulties in establishing a good teacher evaluation system is related to the difficulty in maintaining objectivity and impartiality. Traditionally, evaluations of teachers' performance have been conducted by principals or supervisors, peers, students, and at times, self-ratings performed by the teachers themselves. The concern with many of these practices is that they are clearly subjective and vulnerable to the quirks and frailties of the raters, or what Glass and Martinez call the "politics of teacher evaluation," not to mention the professional incapacities of the raters (Alicias, 2005).

[6] In some districts this is a consequence of the contract achieved by collective bargaining.

2.2.1.2 Difficulty of Establishing a Good Teacher Evaluation System

It is generally acknowledged that establishing and implementing a good teacher evaluation system is a difficult task. Danielson (2000) pointed out that a good system of teacher evaluation must answer four questions: How good is good enough? Good enough at what? How do we know? Who should decide? If these questions were asked in a typical manufacturing enterprise, answers might be much easier to provide, as there would be clear standards and criteria at hand to measure the process and products. However, in the field of education, such standards and criteria that are commonly accepted for the evaluation of teacher performance have often not been available.

Sykes (1985) described teaching, like parenting, as a natural, spontaneous, organic human activity. As such, one's teaching style depends largely on one's personality, as well as on tacit, idiosyncratic approaches to human relations. In addition, a number of studies have suggested that the cultural context of both students and teachers should be observed in classroom teaching so as to avoid possible cultural conflicts and in order to promote a pleasant class environment. This type of pedagogy is referred as culturally-sensitive pedagogy (Thomas, 1997), or culturally responsive, culturally respective, culturally-rooted, culturally relevant, and culturally appropriate (Nguyen et al., 2006). In any case, the primary ingredients for success usually are defined as knowing one's subject matter and caring about children, around which technical embellishments can marginally matter. However, teaching is such an enormously complicated and philosophically multifaceted act that its full import has eluded the increasingly sophisticated methodological and conceptual tools of the social sciences (Sykes, 1985).

Defining what is a good teaching is by no means an easy assignment. The No Child Left Behind (NCLB) Act (2009) defines "highly qualified teachers" as those who must be fully licensed or certified by the state and must not have had any certification or licensure requirements waived on an emergency, temporary, or provisional basis." In addition, teachers also must demonstrate subject matter competence (Title IX, Part A, Sec. 9101). However, the certification standards for highly qualified teachers have been lowered by statute and the final regulations allow teachers who have enrolled in alternative-certification programs, not necessarily completing them, to be designated as highly qualified as well. Moreover, some states such as Texas, Florida and California have proposed standards that allow candidates who have not attended teacher preparation programs to be certified so long as they have a bachelor's degree and pass a state test (Darling-Hammond & Sykes, 2003). Obviously, teacher qualifications are being interpreted in a variety of ways throughout the country.

It is worth noting that the evaluation of teacher effectiveness can involve many different aspects of pedagogical practice. According to one research-based protocol, the Framework for Teaching (FFT), developed by Charlotte Danielson in 1996, teaching activity can be divided into 22 components and 76 smaller elements, which are clustered into four domains of teaching responsibility: planning and preparation, classroom environment, instruction, and professional responsibilities. Whether a so-called outstanding teacher should be defined as excelling at all of these aspects to the exclusion of other pedagogical attributes does not admit simple and straightforward answers, especially given the enculturating nature of teaching.

Teacher evaluation is complex because it serves a variety of purposes. This further exacerbates the difficulty of establishing a sound evaluation system. During the past decade, constant efforts have been attempted to establish a better teacher evaluation system and, more diversified factors have been included to evaluate teacher performance for various purposes. For example, principals and other school personnel conduct observations of teacher practice in order to make tenure and retention decisions. Teacher salary and pay decisions, conversely, are more based on their experience, degrees and some "value-added" scores produced from their student performance on state assessments. Other promotions or professional development responsibilities may depend on some combination of personality, motivation, classroom performance, academic degrees and some external credential such as National Board Certification (Hill et al., 2012).

Could one solution be to construct a better teacher evaluation system by incorporating multiple extensive indicators of teacher effectiveness? This plan seems to be not viable not only for technical reasons related to implementation. Among scholars it has been difficult to obtain consistent results regarding credible and reliable indicators of professional practice. The same inconsistency marking traditional measures of teacher effectiveness characterizes recent statistical studies. For instance, teacher experience and teacher test scores are asserted to be mostly consistently linked to student achievement in one study (Clotfelter, Ladd and Vigdor, 2007), while they were found to explain only a modest fraction of the variation in student outcomes in other studies (Kane, Rockoff and Staiger, 2008; Goldhaber and Brewer 1997; Hanushek 1996).

There could be other obstacles to establishing a good system of teacher evaluation. For example, many principals and assistant principals have to face the time issue while trying to balance their work between completing teacher evaluations and other tasks such as managing their other day-to-day operations and handling many other issues with more immediate timelines (Danielson interview, 2013). Provide a second or a third issue as well, for example, opposition from teachers' unions or the expenses entailed in ramping up assessment at the cost of other potentially reforms that might lift student achievement more rapidly, such as new curricula or better professional development.

2.2.2 Current Efforts to Change Teacher Evaluation

In recent years, states and districts have launched unprecedented efforts to develop more precise and useful systems to evaluate teacher performance in order to provide teachers with the feedback they need to improve their practice (Bill & Melinda Gates Foundation, 2012). Researchers and policy makers have started to explore more objective approaches to quantify the heterogeneity in students' test score trajectories and to use student achievement outcomes for measuring teacher effectiveness in an attempt to promote teacher quality and student learning.

When the No Child Left Behind Act (NCLB) was signed into law in 2001, Adequate Yearly Progress (AYP) was introduced as one of the cornerstones of NCLB and adopted to measure the progress of students nationwide. NCLB defined AYP as an indicator to signal how public schools and school districts in the country perform academically according to student achievement results on standardized tests. To evaluate

the AYP indicator, a school must compute for all students in a grade, as well as for various subgroups, the proportions meeting a fixed standard, and then compare these proportions with those obtained in the previous year (Braun, 2005). Although the NCLB accountability system may appear to focus on change, in many ways, it actually focuses on status (Linn, 2004). Therefore, by employing the AYP under NCLB Act, the judgments of students within schools are made on the basis of current status. Concerns about using the current status model for evaluating schools include that students entering with a higher level of achievement will have less difficulty meeting the proficiency standard than those who enter with a lower level do.

Most educational researchers and practitioners have recognized that reporting school test results as measured by the percentage of students who score at or above the proficient level using status model or cohort-to-cohort change model is unfair to some teachers and school administrators. This is due to the fact that students' current test scores are influenced by many factors beyond the control of the teachers or schools such as out-of-school experience, family and community inputs. Above all, student achievement is cumulative in nature, as it is the result of the input of past teachers, classroom peers, actions taken by administrators, and so on (Harris and Sass, 2005). As such, evaluating schools or teachers based on whether their students meet those proficiency standards will be neither accurate nor fair.

While largely holding schools accountable for the performance of their students, the NCLB Act intended to require more accountability for the achievement of students throughout the nation. With one focus of the legislation on the preparation of a quality teaching force that will provide students with the best education possible, it carried an

expectation that improvements in teachers' professional development will promote positive changes in teaching practice, which, in turn, will enhance student achievement. With the implementation of the Obama administration's Race to the Top initiative, participating states are required to make binding commitments to measuring teacher performance using student learning outcomes and placing more emphasis on teacher accountability through establishing statewide teacher evaluation systems geared toward improving teacher effectiveness.

Over the last decade, a series of statistical approaches have been developed to explore superior techniques to incorporate student learning outcomes into measures of teacher effectiveness, with many such approaches found wanting. For instance, single point-in-time analyses may reflect demographics more than effectiveness, and moreover, they cannot distinguish between schools or teachers that promote skill development and those that allow students to languish (McCall et al, 2004). Analyses employing status models or cross-sectional measures cannot account for students' prior status, such as whether students entered with high or low skills, or whether they have gained or lost ground as a result of instruction. The cross-sectional percent-proficient model, hence, has been characterized as one of the least valid evaluation methods (Flicek & Wong, 2003). Since 1994, the status models have been the school accountability paradigm embedded in Title I, and many state accountability system as well (Piche, 2007). Status models can be appropriate for making judgments about the achievement levels of students at a particular school for a given year, whereas cohort-to-cohort models are better at tracking whether a school is improving. However, both are less useful for comparing the effectiveness of teachers or instructional practices, either within or across schools (Braun, 2010).

While status models and cohort-cohort models were largely questioned, the student growth model is introduced in tandem as an alternative to measure the effectiveness of teachers and schools. It is based on the premise that meaningful and defensible judgments about teachers or schools should be informed by their contributions to the growth in student achievement and not based solely on the proportions of students who have reached a particular standard (Braun, 2005). A growth model should capture a student's score change over time and focus on the change itself. Student growth models, when used most accurately, require scores that can be mathematically compared from one occasion to another, be connected for the same students over two or more occasions, and show changes that indicate trait changes (O'Malley, 2011). In contrast to the status model while a single year's assessment is used, growth models can provide richer information on student learning by connecting multiple assessments. One common approach that this model utilizes is to measure student achievement by tracking the test scores of the same students from one year to the next to determine the extent of their progress.

While focusing on student learning by tracking test scores, the student growth model faces the challenge of the changing nature of the assessment construct over time. When constructs of assessments shift across grades, such as when mathematics assessments move from testing arithmetic skills in third grade to testing pre-algebra and geometry skills in later grades, the growth model results may lead to misleading longitudinal interpretations (Reckase, 2004; Martineau, 2006).

Besides, accountability systems built on growth models give teachers and schools credit as long as their students show improvement, regardless of whether they were high-performing or low-performing to begin with. However, growth models usually do not

control for student or school background factors, and thereof cannot address which factors are responsible for student growth (Braun, 2010). In addition, Willms (2008) indicated that depending on the design and psychometric characteristics of the assessment, students' rates of growth in achievement may be statistically related to students' socioeconomic status (SES), with those who start out with higher scores typically gaining at faster rates.

The effectiveness of schools or teachers can best be measured by following individual students over time and analyzing the changes in their achievement outcomes. (McCall et al, 2004; Doran & Izumi, 2004). One type of analytical procedure, commonly referred to as value-added analysis, has been widely used to estimate the school effects on student growth (Linn, 2004). VAMs seek to control for the influence of selected factors or the impact of an intervention on student performance, and therefore objectively isolate the contributions of teachers and schools to student learning. In this way, what each teacher or school makes in a given year can be compared to the performance measures of other teachers or schools.

2.2.2.1 What are States Doing about Teacher Evaluation?

Given encouragement and support from the Obama administration's Race to the Top program, as well as the NCLB waiver policy (U.S. Department of Education, 2012), a number of states and districts are launching new initiatives to improve their teacher evaluation systems. For example, New Hampshire, Ohio, New York, Massachusetts, as well as Dallas, Houston, Denver, and Washington, D.C. have begun to develop what is intended to be a more credible and comprehensive systems for measuring teacher

effectiveness. Most states are creating new teacher evaluation systems by including a variety of teacher performance indicators, which, by and large, center on the growth in student learning outcomes over time while retaining indicators of teacher practice based on classroom observations and other evidence as well. These newly developed teacher evaluation systems often seek to estimate the contribution that teachers make to that growth by tracking individual students' academic performance over several years. The quantitative evaluation of teachers based on an analysis of the test score gains of their students shows a new prospect and has gained many proponents in recent years.

2.3 Indicators of Teacher Effectiveness

2.3.1 Introduction to Indicators of Effective Teaching

With the increasing attention directed toward teacher evaluation, a range of teacher-linked factors that may influence student learning are being investigated as part of ongoing efforts to measure teacher effectiveness. The studies have identified a variety of teachers attributes that affect student learning outcomes, and meanwhile found alarming inconsistency regarding the relationship between teachers' characteristics and student achievement as well.

Darling-Hammond (2000) stated that the variables presumed to be indicative of teachers' competence that may link to student learning include measures of academic ability, years of education, years of teaching experience, measures of subject matter and teaching knowledge, certification status, and teaching behaviors in the classroom. Further, Berk (2005) identified 12 potential sources of evidence to measure teacher

effectiveness, which included (a) student ratings, (b) peer ratings, (c) self-evaluation, (d) videos, (e) student interviews, (f) alumni ratings, (g) employer ratings, (h) administrator ratings, (i) teaching scholarship, (j) teaching awards, (k) learning outcome measures, and (l) teaching portfolios.

Darling-Hammond and Youngs (2002) reviewed research on teaching qualifications and student achievement, and argued that student-teaching experience, as well as the characteristics of the teacher training program, such as the pedagogical coursework and subject matter knowledge, are at least as important in producing effective teachers as other commonly examined teacher characteristics. Moreover, this review also indicated that the associations between teacher qualifications and student learning are often mediated by the grade level and subject matter. Besides, some qualifications may matter more than others, at least in selected subjects and grades.

In a study of the relationship between teacher quality and student achievement using data from the Prospects National Longitudinal Study, Rowan, Correnti, and Miller (2002) used a multi-level hierarchical linear growth model for students from grade 1 to 6 to examine the effect of the "presage" variables. The characteristics that are discussed in the study include teacher certification status, advanced degrees, and experience, as well as the "process" variables, such as using active teaching methods and aligning content coverage with assessment. Results of the analyses show consistency across cohorts but differences by academic subject. For example, teacher experience was found to have significant effects on students' mathematics and reading growth, whereas the impacts of teachers' degree and certification on students' achievement growth were only evident for reading.

Hill, Rowan and Ball (2005) employed a linear mixed-model methodology to explore the relationship between teachers' mathematical knowledge for teaching, which focused on measuring the specialized mathematical knowledge and skills used in teaching mathematics, and the gains in student mathematics achievement. Consonant with findings from other educational production function literature, results of this study showed that teachers' mathematical knowledge is significantly related to student achievement gains in both first and third grades after key student- and teacher-level covariates are taken into account.

Wenglinksy (2002) found that after controlling for class size and socioeconomic status (SES), several aspects of teacher quality were still significantly related to student achievement. These included the teachers' college major, professional development in using higher-order thinking skills and in diversity, and hands-on learning. This study measured three aspects of teacher inputs (teachers' education level, their major in the relevant subject area, and years of teaching), and ten aspects of professional development (the amount of professional development teachers received last year and whether teachers received any professional development in the last five years in the topics of cooperative learning, interdisciplinary instruction, higher-order thinking skills, classroom management, portfolio assessment, performance-based assessment, cultural diversity, teaching special-needs students, and teaching limited-English-proficient (LEP) students), and concluded that teachers, through quality classroom practices, can contribute as much to student learning as the students themselves.

Wayne and Youngs (2003) concluded from several studies that student achievement is only weakly related to the ranking of teacher's undergraduate programs.

Additionally, in some subjects such as reading, students may benefit from teachers with higher verbal scores. Other results suggested that mathematics teachers' degrees and coursework may contribute to improved student achievement in mathematics, and their certification also matters. As this influence was only detected for mathematics, Wayne and Youngs speculated that it may be due to the fact that across the years there is a more substantial research base for this discipline.

In the addendum to a report about teacher preparation research, Wilson and Floden (2003) synthesized research on teacher professional characteristics and examined the factors and credentials that may be related to teacher effectiveness, such as teacher subject knowledge, advanced degrees, pedagogical theory and knowledge, field-based experience, and the teacher preparation programs. The authors claimed that there is an alarming inconsistency in the findings regarding the relationship between student achievement and teachers' characteristics.

Rice (2003) analyzed a variety of indicators of teacher characteristics in the literature that have been assumed to reflect teacher quality and categorized these characteristics into five broad groups of measurable and policy-relevant indicators: teacher experience, teacher preparation programs and degrees, teacher certification, teacher coursework, and teachers' own test scores. The study concluded that these five categories of indicators can all contribute positively to teacher effectiveness, although individual effects may differ depending on the subject areas, grade levels, and student populations.

Other perspectives such as the research from the Alliance for Excellent Education (2008) specified that teaching qualifications, such as teaching experience, certification status, or advanced degrees, have been used to reward teachers for years. These qualifications can serve as quality control and sometimes predict student achievement, but they are only weak proxies for teacher effectiveness, as opposed to indicators based directly on measures of student learning.

Teachers' opinions of good indicators of teacher effectiveness have also been surveyed (Coggshall et al, 2011). Interestingly, 56 percent of surveyed teachers believed that student performance on standardized tests is a good or excellent indicator of teacher effectiveness -- despite the fact that teachers' unions typically oppose using test scores to measure effectiveness. However, far higher percentages of teachers preferred other indicators of effectiveness. For example, 92 percent of teachers agreed that student engagement is a good or excellent indicator of teacher effectiveness and 72 percent emphasized that the comparison of "how well their students were learning" and "the learning of students in other schools" is a good or excellent indicator.

The 2013 report from the American Association of Colleges for Teacher Education (AACTE)'s Professional Education Data System suggested that teacher candidates' future success in classroom can be better assured by admitting academically competitive candidates, incorporating better clinical experience in the teacher preparation program, and utilizing the performance-based exit measures.

2.3.2 Indicators to be Addressed in This Dissertation

As more attention is directed toward teacher evaluation and accountability, diverse indicators of teacher effectiveness are being investigated. Each targets somewhat different aspects of teaching performance, and all are fallible and subject to bias.

As described in the previous section, earlier findings have shown that teacher characteristics, such as credentials and experience, can hardly fully reflect teacher effectiveness. The relationship between teachers' characteristics and student achievement varies substantially by academic subject, grade level and student and teacher population. Another traditional indicator of teacher effectiveness, classroom observation, can be influenced by factors unrelated to teacher performance, one of which, apparently, is the experience of the observers. The Measures of Effective Teaching (MET) study (2012) involved observers who were highly trained and had to pass an exam to demonstrate their skills; however, it is unlikely that this level of training can be available in everyday school settings. In addition, classroom context will likely affect observation measures; for example, it may be difficult to make valid comparison between the classroom management skills of a teacher who has emotionally impaired students, subject to frequent disruptions, to the skills of a teacher whose students are less disruptive (Harris, 2012).

At present, research efforts increasingly target value-added measures and student learning objectives. Value-added modeling (VAM) analyzes multiple years of students' test scores, decomposes them into components attributed to student heterogeneity and to teacher quality. On the other hand, with student learning objectives (SLOs), teachers work with their instructional supervisors to

create specific objectives and establish metrics to measure students' progress towards those objectives. In this study, these two types of indicators for teacher effectiveness are investigated: VAM estimates (representative of indicators based solely on student learning outcomes derived from standardized assessments) and SLOs quality (representative of indicators related to classroom practice and student learning). Both approaches are subject to bias, which will be discussed in the following sections.

Thus far, very limited evidence regarding the validity of different approaches to measuring teacher effectiveness is available. In particular, there is no evidence about the validity or reliability of SLOs (Harris, 2012). The Measures of Effective Teaching (MET) study (Bill and Melinda Gates Foundation, 2012) found statistically significantly positive correlations between value-added measures and classroom observation rubrics based on the Danielson Framework. The relationship was stronger for English Language Arts (ELA) than for mathematics. When student survey feedback was correlated with value-added measures in the further analysis, the relationship appeared to be stronger. This result is consistent with the findings of prior studies that investigate the correlation between value-added measures and principals' low-stakes evaluations of teachers (Harris & Sass, 2009; Jacob & Lefgren, 2008).

Clearly, more studies about value-added measures as well as other evaluation methods are needed to determine how valid they are for particular groups of teachers. The limited evidence about different approaches to measuring teacher effectiveness is a big problem since the information from available value-added studies cannot provide adequate support for decision making or give clear guidance for future research directions. Only when similar analyses are conducted to other measures can the best

options for measuring teacher effectiveness be selected from the alternatives (Harris, 2012). Therefore, through investigating how the value-added estimates relate to other indicators of teacher effectiveness, this dissertation focuses on the relational analysis among different indicators of teacher effectiveness measures, and thus contribute to the current research on teacher evaluation.

2.3.3 VAM as an Indicator of Teacher Effectiveness based on Student Achievement

McCaffrey and Lockwood (2008) indicated that although the origins of VAM of teacher effects date back over 30 years (Hanushek, 1972; Murnane, 1975), interest in relevant methods among researchers, policy makers, and educators grew precipitously following the publication of a technical report by William Sanders and June Rivers in 1996. This report argued that teacher effects estimated from student test score gains could predict student outcomes at least two years into the future, suggesting that teachers have persistent effects on their students' achievement and the accumulation of these effects could be substantial. After another paper from Sanders and his colleagues (1997) claiming that teachers are the most important school-related source of variation in student achievement, as well as the replication of the Sanders and Rivers results (Mendro et al., 1998; Rivers, 1999), interest in VAM continued to grow.

With the increasing availability of longitudinal student achievement data derived from standardized assessments since the NCLB Act (2001), value-added methods have gathered a great deal of attention among both policymakers and researchers. Currently, many states and districts such as Florida, North Carolina, Denver, Dallas, and Houston have begun to employ the VAM approach for their teacher evaluation system. Of the 65

member districts of the Council of Great City Schools, nearly one fourth have implemented some form of value-added based school or teacher rewards program (Hill, Kapitula & Umland, 2011). Other states and districts are either designing pilot VAM programs or are using VAM for lower-stakes purposes such as professional development to explore its viability.

In 2008, Ohio began using VAM as one component of its state accountability system, to show how much schools and districts were adding to their students' learning over the course of one or more school years (Public Impact, 2008). In addition, the most current Ohio system for evaluating teachers relies on two key evaluation components, each weighted at 50 percent: a rating of student academic growth, which requires that value-added data be included if available; and a rating of teacher performance based on classroom observations and other factors (Ohio Department of Education, 2013). Clearly the use of VAM for teacher accountability is on the rise (Soto et al, 2011).

Value-added models are a family of statistical models that attempt to attribute some fraction of student achievement growth over time to certain schools, teachers, or programs. They aim to address the problem of nonrandom assignment of students to teachers and schools (Braun, 2010). A related way of thinking about value-added models is that they are "an attempt to capture the virtues of a randomized experiment when one has not been conducted" (Organisation for Economic Co-operation and Development, 2008, p. 108). They are proposed for four main research purposes including school and teacher improvement, school and teacher accountability, program evaluation, and research. One of the reasons why VAM has attracted growing interest is that early VAM studies purport to show very large differences in the effectiveness among teachers. If

these differences can be substantiated and causally linked to specific characteristics of teachers, the potential for improvement of education would be great (McCaffrey et al, 2003).

Value-added models hold out the promise of isolating the effects of teachers or schools from that of other factors such as prior academic achievement, family background, poverty, or school leadership. By employing a collection of complex statistical techniques to analyze multiple years of students' test score data, VAM decompose the variances into components attributed to student heterogeneity and to teacher quality. VAM can provide estimates for the effects of individual schools or teachers so that the estimated contributions to student achievement growth from different teachers or schools can be compared to each other or compared with that of the average teacher or school (Braun, 2010).

A value-added estimate is meant to approximate the (causal) contribution of the school, teacher, or program to student performance (Braun, 2010). It is playing an important role in many high-stake decisions and policies regarding teacher evaluation. For example, in a recent bill drafted by the General Assembly of Pennsylvania, VAM-based estimates of teacher and school effects have affected salaries and career advancement as well as contract renewal not only for teachers but also for school and district administrators (McCaffrey et al, 2003).

There are a variety of VAM models depending on the specific statistical techniques applied and the factors selected to be accounted for. For example, some VAM models calculate the difference between observed scores of the students and the expected

scores after controlling for other factors that might be related to differenced student academic achievement. A summary of the aggregated differences serve as the school or teacher value-added estimates.

The limitations of using VAM for estimating teacher effectiveness have also been widely analyzed and discussed. The most fundamental limitation is that the use of VAM results requires a causal interpretation of the estimates of teacher effectiveness. Although randomized experiments are widely considered the gold standard in scientific work, it is generally the case that students are not assigned at random to different classes or schools. "If making causal attributions is the goal, then no statistical model, however complex, and no method of analysis, however sophisticated, can fully compensate for the lack of randomization" (Braun, 2005). Consequently, the teacher effectiveness estimates obtained from VAM likely represent a combination of many factors in addition to the actual teacher contributions. Therefore, attributing observed differences solely to true differences in teacher effectiveness can undermine the fairness of the teacher evaluation process.

Indeed, there are many factors that are not -- or cannot -- be taken into account in VAM. For example, teachers may not be randomly assigned to classes and thus may be inappropriately credited or penalized for their students' growth. Moreover, student learning could be influenced by many factors that are not related to teacher instruction and performance, such as school resources, parent involvement, peer interactions and so on. The variables that are controlled in the VAM models cannot fully adjust for all pre-existing differences among classes. As a result, simply attributing student learning and

growth to the contribution of the teacher, indicated by the VAM estimates that are just residuals from a regression, may introduce bias.

Braun (2005) discussed the problems of bias with VAM model assumptions and imprecision with VAM model estimates. For example, a VAM model may assume that a teacher's effect is essentially the same for all of that teacher's students in a given subject and year and that this effect persists undiminished into the future for those students. However, such assumptions may not hold so that VAM produces biased estimates of teacher effects. In addition, the precision of the VAM model estimates is reduced by the uncertainty involved in the estimation process due to the fact that there are a limited number of students contributing to the estimated effect for each teacher.

Reardon and Raudenbush (2009) also examine the plausibility of the VAM assumptions in practice, as well as the consequences of violating those assumptions for practitioners and ultimately for students. First, comparing the effectiveness of different schools (teachers) entails the comparison of the entire distributions of the potential outcomes in those schools (teachers), which implies that if a good school is good for one subset of students it is good for all other subsets. Obviously, there could be many reasons to challenge that this assumption, for one, not all schools have sufficient numbers of students at all skill levels to support precise estimates of mean achievement gains at each skill level.

Second, the assumption of the stable unit treatment value (SUTVA) (Rubin, 1986) implies that each student possesses one and only one potential outcome in each school (teacher), which may seem to be implausible since peer effects are real, particularly given

the reality of school segregation with regard to student demographic characteristics (family socio-economic background, ethnicity, linguistic background, and prior achievement) and student composition for the organization and delivery of instruction. There can be many potential outcomes with a particular teacher depending on the other students in the class.

Further, most analyses compare school (teacher) effectiveness by comparing their means, overall or for sub-groups, which implies that the quantity of interest is the mean difference in potential outcomes associated with any comparisons. Reliance on the mean contains the assumption that the unit of the student test score distribution are on an interval scale of social interest. However, in measuring cognitive skill, it is unclear what the reference metric should or could be (Ballou, 2008). Thus, how one can determine whether a given test metric should be considered interval-scaled remains unclear.

Besides, practical limitations of value-added measures are discussed in several other studies. One included the "test ceiling effect" meaning teachers whose students start off with high achievement will receive lower performance ratings than they deserve (Koedel & Betts, 2009). In addition, Harris and Anderson (2012) pointed out in their study that almost all the evidence about value-added validity is based on studies in elementary schools and that typical value-added measures are biased in middle and high school. Another publicized study concluded that value-added measures are probably highly sensitive to the context of teachers' classrooms, including behavioral issues and the school culture (Reardon & Raudenbush, 2009).

Currently, Student Growth Percentiles (SGPs) (Betebenner, 2011) have often been considered an alternative to the VAMs. Initially devised to provide useful descriptions of student growth by contextualizing current performance, SGPs compare a student's current test score with those of her academic peers – those students with the same or similar prior test score trajectories. When SGPs are used for the purpose of teacher accountability, for example a median growth percentile can be used as an indicator for the growth of a class, the concerns raised with VAMs are appropriately raised with SGPs as well (Braun, 2012).

2.3.4 SLOs as an indicator of Teacher Effectiveness based on Classroom Practice

2.3.4.1 SLOs Introduction: Concept, Accountability, and Application

As teacher evaluation attracts greater attention, various approaches to measuring teacher effectiveness are emerging. Student Learning Objectives (SLOs) has become an increasingly acknowledged approach, and is designed specifically to tackle the issue of measuring teacher effectiveness for the non-tested subjects and grades.

The US Department of Education (ED) defines tested grades and subjects as those covered by the state's assessment under Elementary and Secondary Education Act (ESEA) and non-tested grades and subjects as those falling outside that coverage. Prince et al (2009), while discussing rewarding the performance of teachers of non-tested subjects and grades, report one of their findings as "the other 69 percent", which refers to the percentage of teachers whose contributions to student learning cannot currently be measured by test-based approaches (e.g., value-added models) because the subjects or grades they teach are not assessed with state-wide standardized tests. Indeed, state

standardized assessments are mostly designed for the subjects related to mathematics and English but not for others. However, the data show that only 31 percent of Florida classroom teachers taught reading and mathematics during the 2004-2005 school year and only 15 percent of the staff in large high schools in Alaska were responsible for teaching reading, writing, and mathematics during 2005-2006.

A good teacher evaluation system must include all teachers and gauge their effectiveness regardless of the subject area or grade they teach, or their particular student group characteristics. If the eligibility for teachers to receive performance awards were restricted only to those who teach the subjects that are assessed on state-mandated achievement tests, apparently only a very small percentage of public school teachers would qualify. As emphasized by the Race to the Top (RTTT) program initiated by the federal government (2009), the teacher evaluation systems are expected to differentiate teacher effectiveness by incorporating student academic growth as a significant factor. While defining student growth, RTTT particularly stressed the clear distinction between "tested grades and subjects" and "non-tested grades and subjects".

The RTTT Technical Assistance Network defines SLOs as "a participatory method of setting measurable goals, or objectives, based on the specific assignment or class, such as the students taught, the subject matter taught, the baseline performance of the students, and the measurable gain in student performance during the course of instruction" (2010). The North Carolina Department of Education (2010) noted in its RTTT application that through the SLOs design and implementation process, teachers and administrators work together to identify specific Standard Course of Study-related areas of focus for each class. With SLOs, teachers establish learning objectives for

individual students, a particular student group or the class as a whole, based on their knowledge of the students. Once the learning objectives are created for students, the extent to which these objectives have been achieved during a particular learning period can be used to evaluate the teachers' effectiveness as well as the students' growth.

The development of student learning objectives is uniquely a teacher activity, and is particularly suitable for individual teacher evaluations. Setting student learning objectives is a process that starts with something teachers should know and be able to do well. It builds on their strengths and then extends teachers' opportunities for further thoughts and analyses about their teaching practice, which capitalizes on teacher professionalism (CTAC, 2005).

Goal setting will affect performance by directing attention and effort toward activities that are relevant; energizing or generating greater effort; impacting effort and arousing task-relevant knowledge and strategies (Locke & Latham, 2002). Athletic coaches and trainers speak of setting goals that reach for one's "personal best," but specific and personal goals or objectives can produce even better outcomes. According to the findings of the CTAC Denver study (2001), writing objectives for students requires teachers to collect better information and obtain greater precision than the customary approach of the planning of teaching that is based on lesson plans. This could lead to dramatic transformations in teachers' work. Schools may be influenced by the new, objective-based approach and become more precise, open and reflective about student outcomes.

Characterized as being highly flexible and directly tied to teacher's classroom practices, SLOs holds evident advantages over many other approaches to measuring teacher effectiveness. For example, SLOs have been largely regarded as a substitute for standardized assessments for the non-tested grades and subjects, and thus can be implemented across all grades and subjects.

In addition, as SLOs are often directly tied to the regular practices of teachers' work, they can increase the credibility of the objectives designed for student learning and growth. Meanwhile, the teachers' understanding of what must be done in order to meet a given performance target will be promoted as well. Further, SLOs can potentially create greater teacher buy-in of the teacher evaluation system (Buckley & Marion, 2011). Besides, SLOs do not impose specific teaching models or conflict with state or district standards, and can be implemented with a variety of assessment formats, such as the nationwide standardized test, state or district assessment, and even teacher-developed measures. By now, SLOs have been widely adopted as a component of teacher evaluation systems in many states and school districts including Rhode Island, North Carolina, Ohio, Virginia and many more.

At the same time, the SLOs approach to quantifying teacher effectiveness also faces many challenges. First, every student learning objective must be approved by a school administrator, which requires thorough understanding of the objective and, therefore, significantly increases the school administrator's time commitment to supervision. Moreover, the SLOs implementation process, such as providing guidance to teachers while developing the student learning objectives and facilitating to substantiate the outcomes, is highly resource intensive as well.

2.3.4.2 SLOs Quality[7] as an Indicator of Teacher Effectiveness Based on One Aspect of Classroom Practice

A common approach to SLOs is to treat them as the goals that students are expected to attain within a certain period of time, and the extent to which they are achieved can be used as an indicator of the teacher effectiveness. Denver, Colorado and the district of Charlotte-Mecklenburg, North Carolina are at the forefront of this approach (Buckley & Marion, 2011). This study, however, focuses on one aspect of SLOs and adopts a different approach to analyzing SLOs. In the TIF-LEAP project, teachers are required to develop the SLOs for their students, either individually, by focusing on a group of students, or by developing them for the class as a whole. The quality of these objectives for student learning presented by the teachers is then evaluated and considered as the indicator for the effectiveness of these teachers.

Simply put, SLOs are treated as the written objectives created by teachers, and the quality of these written SLOs is employed as a proxy for teacher quality in this study. When SLOs are tied to high-stakes decisions, special cautions and, perhaps, an audit process is needed since, though the SLOs are aligned with state curriculum frameworks, creating individual learning objectives can still be a highly subjective process that can affect whether students are likely to achieve the objectives. The problems of using this SLOs approach in high-stakes accountability system are further discussed in later chapter.

[7] Other studies using SLOs quality as indicator of teacher effectiveness have not been found.

In order to create high quality learning objectives for each student, a particular student group or the class as a whole, and make those goals attainable yet rigorous, teachers need to have good academic instructional plans, strong pedagogical skills and effective professional development practice. A key advantage of the SLOs approach over the traditional test-centered approaches to accountability is the active involvement of teachers in their own assessment. SLOs are designed to reflect and incentivize good teaching practices, such as setting clear learning targets, differentiating instruction for students, monitoring students' progress toward these targets, and evaluating the extent to which students have met the targets (Marion et al, 2012). The quality of the teacher-developed SLOs should, to a significant degree, reflect the quality of the teacher's classroom practice, which serves as the main assumption behind this study.

Findings from the Community Training and Assistance Center (CTAC) Pay for Performance study in Denver have already suggested that student learning objectives are regarded as a significant element in measuring teacher practice, connecting student achievement with teacher compensation. One of the compelling results from the Denver study is that teachers who developed high-quality SLOs produced better student achievement growth. Additionally, the quality of the student learning objectives correlated positively with student gains as well. More specifically, students of teachers with the highest quality objectives (rubric score 4) showed increases in achievement on both the Colorado Student Assessment Program (CSAP) and the Iowa Test of Basic Skills (ITBS) (CTAC, 2001).

The SLOs approach to measuring teacher effectiveness faces challenges as well. The practical fairness is an issue that arises when using the quality of the written SLOs as

an indicator of teacher effectiveness. For a variety of reasons, some teachers may be superior at developing the objectives than others, although they do not differ as much in actual classroom teaching and in fostering students' improvements. In such cases, treating the quality of SLOs as teacher effectiveness indicator may suffer from bias.

Moreover, developing and monitoring SLOs could be difficult and time-consuming especially when they are still relatively new practices that require much support from states and school districts. In addition, it is hard to ensure the quality of SLOs and the assessments used to measure student learning. Goe (2011) reviewed SLOs that have been used in different states and districts, and identified important challenges that have arisen. For example, in the Austin Independent School District, student learning objectives are developed by individual teachers who enjoy a wide range of options for the choice of assessments and objectives while obeying the state curriculum frameworks under the highly centralized educational system in Texas. The availability of the various assessments that may be used to establish student growth results in a lack of comparability across classrooms. Further, the autonomy that teachers have to set their own objectives with SLOs can be viewed as a weakness, as the measures are unlikely to be comparable across teachers and may be too easily manipulated to give the appearance of high performance (Harris, 2012).

2.3.4.3 SLOs in Charlotte-Mecklenburg School District: Introduction

Charlotte-Mecklenburg Schools (CMS) is a large school district located in and around Charlotte, North Carolina. There are 159 schools throughout the cities and towns of Mecklenburg County and more than 141,100 students from kindergarten through 12th

grade. Students in CMS have diverse cultural and ethnic backgrounds from 160 different countries. As one of the largest employers in Mecklenburg County, CMS currently employs approximately 18,800 teachers, support staff and administrators.

The Teacher Incentive Fund-LEAP (TIF-LEAP) project in Charlotte-Mecklenburg is a five-year project that began in 2007 and ended in 2012. SLOs have been part of the project ever since its inception. The SLOs project in Charlotte-Mecklenburg School District comprised six components that served as the basic guidelines for teachers to develop the objectives, each addressing a specific aspect of a process that is intended to assist teachers in building quality SLOs.

The first component is population: Teachers must specify the student population for SLOs participation; that is, describe why particular students are selected to participate in the SLOs project. The second component is learning content: Teachers need to develop appropriate content of the instructional objectives for the SLOs. Typically the instructional objectives are selected from the list of objectives from the North Carolina Standard Course of Study (NCSCOS). The next component is interval: Teachers must describe the length of time during which the SLOs will be implemented and completed. Another component is the assessment: Teachers need to introduce and explain the test instruments employed in developing the SLOs. Teachers may have a variety of options in choosing the pre- and post-assessments for their students.

The remaining two components are growth expectation and strategies: Teachers were required to indicate their expectations for how students will progress over the time interval, as measured by the pre-assessments and post-assessments. Moreover, the

instructional techniques and approaches that will be used to facilitate students achieving the growth objectives should be provided. Each teacher is expected to develop one or more complete SLOs that should incorporate all of the components described above.

In addition to the multiple components of the SLOs, there are three types of SLOs designed and implemented in the Charlotte-Mecklenburg schools: Class SLOs, Target SLOs and Team SLOs. Teachers were afforded these options to develop different types of SLOs depending on their situations. The primary difference among the three types of SLOs centers on the student population teachers intend to address. For example, Class SLOs are designed for an entire class of students, while the population for Target SLOs is only a subset of students in a class that are singled out by the teacher for certain specific needs. Team SLOs are relatively new and are designed by and for a group of teachers of the same grade and subject who share similar learning content within a comparable interval of time, but may have different expectations of growth for their students.

The SLOs project in Charlotte-Mecklenburg school district was designed as an integrated system that was intended to guide and facilitate teachers in developing an objective, measuring a starting point for student learning progress and then striving to achieve the goals set for both their teaching and their students' learning. Since all the objectives developed by teachers for student learning were designed to benefit individuals or groups of students, the ultimate goal of improving teaching and student learning would be accomplished through the successful implementation of the SLOs project.

In general, the main purpose of this study is to explore the relationship between the two approaches of measuring teacher effectiveness. The areas of concentration include focusing on the indicators of VAM and SLOs, as well as how the correlations between the two indicators vary by factors including year, grade and type of school. In the next chapter, the methodology to be used in this dissertation is presented.

CHAPTER 3. RESEARCH DESIGN

3.1. Data Description

3.1.1. Student Achievement Data

This study employs value-added estimates as an indicator of teacher's effectiveness in improving student learning outcomes. In contrast with status measures (e.g. percentage of students at or above certain achievement levels), value-added estimates, through accounting for student prior achievement scores, measure the process of students' academic growth across years. Likewise, SLOs are designed to measure students' progress during a period of learning time, with the objectives typically constructed for an academic semester or year by teachers and principals. Therefore, in order to explore the relationship among teacher effectiveness indicators, it is more appropriate to compare the SLOs with VAM estimates rather than with status measures since they both measure the progress of student growth.

Data in this study include two major components: (1) the student academic achievement data on North Carolina state tests from school years 2007-2008 through 2010-2011 in Charlotte-Mecklenburg school district; and (2) the quality scores of the SLOs from teachers of the TIF-LEAP project in Charlotte-Mecklenburg school district in North Carolina.

In North Carolina public schools, End-of-Grade (EOG) tests are administered to all students from grade 3 through 8, and students from grade 9 to 12 are required to take End-of-Class (EOC) tests. Both EOG and EOC tests cover an assortment of subjects each

year. This study analyzes data on the EOG test scores of students from grade 4 to 8 in both mathematics and reading from school years 2007-2008 through 2010-2011.

Designed and scored differently from the typical school tests that classroom teachers often devise to measure student learning on a limited number of goals and objectives, the EOG and EOC tests serve the purposes of the North Carolina statewide testing program: state and school system accountability. Since the state testing programs aim to measure what students have learned over an entire academic year, multiple test forms are used with the intention of incorporating broader learning content.

Each test form contains a sample of items measuring different aspects of the objectives of the North Carolina Standard Course of Study (NCSCS). While the different test forms are built to the same blueprints, each contains different items representing a different random domain sample of the curriculum. Items on the North Carolina mathematics and reading EOG tests are four-foil, multiple-choice items. The test forms are designed to be parallel and comparable so that all tests are kept equivalent in difficulty and the scores can be compared across forms. The tests are statistically equated at the total test score level. Students' raw test scores are converted to scale scores using software[8] implementing an IRT model with three parameters (threshold, slope and lower asymptote) (North Carolina Assessment Brief, 2008).

3.1.1.1 EOG Test reliability

The technical reports for the North Carolina statewide tests indicate that internal

[8] The software is developed by the psychometric laboratory at the University of North Carolina at Chapel Hill.

consistency coefficient analyses have been periodically conducted to document the reliability for the North Carolina EOG tests. The results of one reliability analysis are shown in Table 3.1. The alpha coefficients imply that the EOG Mathematics and Reading tests are both highly reliable. Likewise, similar reliability analyses have also been performed for a variety of subgroups based on student gender, ethnicity, disability, and LEP (Limited English Proficiency) status, and the results consistently suggest that the high degree of reliability extends across all these subgroups (North Carolina Assessment Technical Reports).

Table 3.1 North Carolina EOG Tests Reliability Indices, Averages by Grade and Subject[9]

Grade	Average Coefficient Alpha Mathematics	Average Coefficient Alpha Reading
Grade 3	0.96	0.88
Grade 4	0.96	0.91
Grade 5	0.95	0.90
Grade 6	0.96	0.91
Grade 7	0.95	0.91
Grade 8	0.94	0.90

3.1.1.2 EOG Test Validity

Test validity is investigated by the North Carolina Department of Public Instruction (NCDPI) and addressed in the annual technical reports as well. Two approaches to validity analysis are employed: content validity and criterion validity.

Evidence of content validity begins with an explicit statement of the constructs or concepts measured by the test. Almost all the test items are developed by North Carolina

[9] Reliability indices averaged across North Carolina EOG and EOC Tests of Mathematics forms, 2006.

teachers and other educators, and every item generated is reviewed by at least two content-area teachers from North Carolina. This process is intended to ensure that all test questions not only match the course standard of their particular grade but also are comprehensible to the students at that level.

Criterion-related, concurrent validity analysis of the state tests reveals moderate to strong correlations between scale scores and teachers' judgment of student achievement, expected grade and assigned achievement levels. In addition, low correlations are found between the scale scores and demographic variables such as gender, limited English proficiency, and disability status from grades 3 to 8 (less extreme than ± 0.10 for gender or limited English proficiency, less extreme than ± 0.30 for disability status). Another type of concurrent validity is investigated based on the trend analysis between students' progress on the National Assessment of Education Progress (NAEP) and their progress on EOG scores. While acknowledging that the EOG scores cannot and should not be compared with NAEP scores directly, and it is not valid to compare the percent "proficient" on each test, the North Carolina assessment team examines the trends of the two test results and find corresponding increases in both NAEP mathematics scores and scores on the North Carolina EOG tests in mathematics across multiple years. As for the North Carolina Reading Comprehension Tests, teachers' judgments of student achievement, expected grade, and assigned achievement levels all serve as sources of concurrent validity evidence. The results of the correlation coefficient analyses range from 0.49 to 0.65, implying a moderate to strong correlation between scale scores and the variables listed above.

3.1.2 Student Learning Objectives (SLOs) Data

The TIF-LEAP project of Charlotte-Mecklenburg school district in North Carolina is a five-year endeavor and has been ongoing since 2007-2008 school year. The program is named Leadership for Educators' Advanced Performance" (LEAP) and CMS has partnered with the Community Training and Assistance Center (CTAC) to support this initiative with three goals: To create a compensation and evaluation system for teachers and principals; To build teacher and principal capacity to foster student achievement; To support recruiting and retaining qualified teachers and principals. There are 21 elementary schools, 12 middle schools and 16 high schools that have participated in this project as either experimental schools or control schools. Among the schools in the project, approximately 22 schools comprising elementary, middle and high schools participated as experimental schools, and implemented the entire SLOs developmental process and practice. A total of over 1,000 teachers and more than 10,000 students from the Charlotte-Mecklenburg school district have been involved in the SLOs project across three years.

SLOs have been an important component of the integrated mission across years. They are designed to provide guidance and support for teachers to develop unique objectives for individual student's learning status and expected growth, which then may inform teacher's instruction and performance to further promote student learning. Teachers may make decisions on the assessment used to measure students' progress with support and supervision from principals. The evaluation of the quality of the SLOs is accomplished by a CTAC team of educators and researchers who anchor, read, and rate all of the SLOs using a holistic scoring procedure. Each SLO is read and rated by at least

two evaluators as meeting the requirements of Level 1, 2, 3, or 4. In the TIF-LEAP project, CTAC has evaluated and scored the quality of the SLOs for 438 teachers in 2008-2009, 943 teachers in 2009-2010 and 826 teachers in 2010-2011. This study investigates the SLOs quality scores for teachers of mathematics and reading from grade 4 through grade 8 across these three years.

The inter-rater reliability is analyzed by the CTAC evaluation team based on SLOs quality score data of school year 2010-2011. The percent of exact agreement is 83.5% while that of adjacent agreement is 100%. The inter-rater reliability coefficient, Cohen's Kappa, is 0.53 ($p < .0.001$; 95% CI: 0.387, 0.663) (CTAC, 2013).

3.1.3 Procedures for Scoring the Quality of Teacher SLOs in the Charlotte-Mecklenburg Schools Project

In order to measure the rigor and overall quality of the SLOs, a four-point rubric has been developed based on the desired criteria or traits, which include learning content, completeness, cohesion and expectations. The standards used to gauge SLOs were derived from a review of teacher planning guides found in the Education Resource Information Center (ERIC) database, the district scope and sequence (which contains subject standards for grades K-12), and the elements that were provided by the CTAC Design Team to the teachers (CTAC, 2005).

The most fundamental feature of the SLOs process is the set of objectives that the teachers develop. Accordingly, the quality of the SLOs is a critical factor in the process. Quality instructional objectives are expected to display four key traits.

The first trait is related to learning content. Quality learning content of the objectives should be closely related to the subject or discipline, appropriate to the student level, and rigorous in thought and application. The choices of content in the SLOs should be aligned with the state standards for the subject and the students' grade level. In addition, high quality SLOs are expected to be comprehensive and include elements from various perspectives. A complete expression of an educational objective should embrace all components of the basic guideline, such as student population, learning content, assessment, baseline data, teaching strategies, expectations, and evidence of whether the objectives are achieved.

Another aspect of quality SLOs is cohesion, which refers to the logic and unity among the elements of the objective and demonstrates that rigorous thought and careful planning have taken place in the development of the objectives. Though incorporating a variety of elements, each SLO should display an impression of being complete and integrated. The final trait of quality SLOs is expectations. With quality objectives for student learning, the teachers should show an understanding of both the population and individuals to be directed and hold high but realistic expectations for each student as well as for him/herself.

In the TIF-LEAP project, schools in Charlotte-Mecklenburg district are assigned to either treatment group or control group. Each teacher in the treatment school is asked to develop two SLOs, which can be Class SLOs, Team SLOs or Target SLOs. These SLOs are approved by the school's principal and then they form the basis for measuring teacher effectiveness in terms of their classroom practice. Establishing SLOs is considered a central component of the project.

The four-point scale rubric employed to rate all SLOs in the CMS project has been validated and field-tested. The scores demonstrate the quality of the SLOs, with 4 indicating "Excellent", 3 "Acceptable", 2 "Needs Improvement" and 1 denoting "Too Little to Evaluate". Below is a detailed description of the 1-4 point rubric:

4 (Excellent): The teacher states clearly what students will learn, expressing completely and coherently all elements of the objective, including the assessment, and demonstrating high expectations for students. There is a strong sense of the whole.

3 (Acceptable): The teacher refers (i.e., from a skill section in a book or test or a program acronym) to what the student will learn but may lack thoroughness in addressing the elements or in making clear the relationship or unity among the elements. The student expectations may seem somewhat conditional or low.

2 (Needs Improvement): The teacher has attempted to address most of the elements of the objective but may not have stated the learning content, showing a lack of understanding about what is expected or confusing the elements (i.e., stating the objective as an assessment goal rather than a learning goal). Expectations for students may be low.

1 (Too Little to Evaluate): The teacher does not address the objective in a manner that shows either an understanding of the task at hand or an effort to complete the task as requested. Objective may place too many conditions or exclude too many students to be reliably assessed.

SLOs are scored holistically by multiple evaluators. Each objective is first scored by two different evaluators individually, and then reviewed by the two evaluators jointly. Any disagreement in the ratings are discussed and analyzed. If the conformity cannot be

reached after further discussion, a third team evaluator is invited to make a final judgment. The holistic scoring procedure enhances inter-rater reliability and also provides a metric with a descriptive referent that can be exploited for further comparisons and analyses over time. In previous analyses a positive correlation is found between the quality of SLOs as measured with the rubric, and student gains in achievement from two independent assessments. Once the SLOs are rated, results are also employed for additional analyses, such as the relationship between the quality of SLOs and whether or not these SLOs have been achieved, and so on.

3.1.4 Data Source and Sample

Data containing the two sets of teacher effectiveness indicators, VAM estimates and SLOs quality, are matched through an anonymous teacher ID. Since the SLOs quality score is only accessible for the teachers in the experimental schools of the TIF-LEAP project in school years 2008-2009, 2009-2010 and 2010-2011, this study draws upon the corresponding teachers' student achievement data in mathematics and reading EOG tests in the three years, as well as these students' prior years achievement scores, to construct HLM models. The VAM estimates are first calculated for all teachers in the particular grades and subjects, both from the experimental and control schools in the TIF-LEAP project. Subsequently, the target teachers' VAM scores are then extracted from the previous results and ranked.

The numbers of teachers employed for analyzing the relationship between teachers' VAM estimates and SLOs quality scores are summarized in Table 3.2. The

groups identified by subject, school year and grade level with less than 10 teachers are not be included in the analyses.

Table 3.2 Number of Teachers with Both VAM and SLOs by Grade and Subject

	2008-2009		*2009-2010*		*2010-2011*	
	Mathematics	Reading	Mathematics	Reading	Mathematics	Reading
Grade 4	17	23	19	17	23	19
Grade 5	18	21	21	18	20	21
Grade 6	17	16	9	22	20	12
Grade 7	13	16	13	14	18	16
Grade 8	11	15	14	10	15	14
Total	76	91	76	81	96	82

In order to obtain more stable estimates, this study obtains the target teachers' VAM estimates using larger groups of teachers and students from the TIF-LEAP project. Data of teachers and students from the experimental schools of TIF-LEAP project[10] are employed for fitting the HLM models that generate teachers' VAM estimates.

The number of teachers and students for the VAM analysis is illustrated in the Tables 3.3-3.4 for mathematics and reading respectively. During data investigation, it was found that the current dataset included a number of teachers who were linked to very few students. In the beginning of the project CTAC has required that the teachers should have at least 5 students in the study, and those teachers with very few students (equal or less than 4) were identified as the school facilitators or coaches. They were certified teachers but not formally registered in the system, and they mainly helped with small

[10] SLOs were only implemented in the experimental schools of the TIF-LEAP project. Teachers in the comparison schools were not of interest to this study.

groups of students, or provided one-on-one support[11]. Therefore, these teachers were not included for analyses in this dissertation.

Table 3.3. The Number of Students and Teachers for VAM Analysis -- Mathematics

	2008		2009		2010	
	N of students	N of teachers	N of students	N of teachers	N of students	N of teachers
Grade 4	1808	164	1178	72	1239	65
Grade 5	1864	164	1963	149	1287	59
Grade 6	1853	57	2053	52	2196	33
Grade 7	2322	97	2007	46	2282	33
Grade 8	2312	97	2215	89	1893	33
Total	10159	579	9416	408	8897	223

Table 3.4. The Number of Students and Teachers for VAM Analysis -- Reading

	2008		2009		2010	
	N of students	N of teachers	N of students	N of teachers	N of students	N of teachers
Grade 4	1786	164	1163	73	1214	65
Grade 5	1850	162	1929	148	1263	60
Grade 6	1815	60	2000	57	2128	36
Grade 7	2307	97	1964	49	2219	35
Grade 8	2561	103	2435	96	2159	34
Total	10319	586	8262	423	8243	230

[11] Based on recent communication with senior CTAC researcher and relevant teachers in July 2014.

Besides, what is also worth noting from the tables is that the numbers of teachers varied by year and grade. The number of teachers in later years and higher grades was noticeably reduced while the number of students showed little change. In order to further investigate the data eligibility in this regard, the student-teacher ratio was thoroughly analyzed and the comparisons between 2009-2010 and 2010-2011 mathematics results were shown in Table 3.5 as an example. First, more than 60% of the teachers taught fewer than 30 students in both years. There were more teachers teaching multiple sections of a class in 2010 than in 2009; for example, nearly 30% of the teachers taught more than 60 students in 2010, while this number was only around 3% in 2009. This is where the main difference between the two years was located, and these were the middle school teachers. In addition, there was a cluster of teachers (18.4%) who each taught 60-79 students in 2010 while a similar cluster (10.8%) in 2009 happened with teachers who each taught 40-59 students.

Further, the overall average class/section size is 20 in 2010 and 15 in 2009. Most middle school teachers taught more than 30 students and 3-4 sections in both years. The mean class/section size for teachers with more than 30 students was 21 in 2010 and 17 in 2009.

Therefore, it can be concluded that more middle school teachers taught multiple sections of a class with larger class size in 2010 than in 2009. There could be multiple reasons for this change. The major one was probably the economic impact and school restructure in the school district across years; for example, nine schools were closed during this time and the rest of the school district was consolidated. The overall number of teachers was noticeably reduced in 2010. Moreover, there could be changes in

teachers' contracts with schools. Some teachers, for example, could teach a block of students for 80 minutes everyday instead of having multiple short sections. Most middle school teachers taught 3-4 sections with 20-30 students in each section. These teachers were retained in the dataset for the VAM analysis.

Table 3.5. Distribution of Students Linked to Each Teacher in Two Years

N of Students Linked to each teacher *	**2010**		**2009**	
	N of teachers	Average class size	N of teachers	Average class size
101-110	2 (0.7%)	26	0	--
90-99	12 (4.0%)	24	0	--
80-89	3 (1.0%)	21	3 (0.6%)	21
70-79	30 (10.0%)	23	3 (0.6%)	19
60-69	25 (8.4%)	22	11 (2.2%)	18
50-59	10 (3.3%)	18	35 (7.1%)	18
40-49	8 (2.7%)	18	18 (3.7%)	15
30-39	6 (2.0%)	18	30 (6.1%)	13
20-29	66 (22.1%)	22	28 (5.7%)	12
10-19	133 (44.5%)	14	238 (48.6%)	12
5-9	4 (1.3%)	8	124 (25.3%)	7

*Teachers with four or fewer students were not included.

Due to the hierarchical structure of the datasets, multilevel models are built to answer the research questions. Analyses are conducted using HLM 6.08 software (Raudenbush, Bryk, & Congdon, 2004). One characteristic of this software is worth noticing: HLM does not allow any missing data at the second level; this implies that missing data at the teacher level need to be removed or replaced. Fortunately in this

study, most teacher level variables are aggregated from those at the student level, and therefore, little missing data at the second level should be expected.

3.2 Variables

3.2.1 Outcome Variables

For research question one (How do value-added estimates of teacher effectiveness based on student test score trajectories compare to the practice-based estimates of teacher effectiveness based on the SLOs quality scores?), and research question two and three (to what extent is the relationship between value-added estimates of teacher effectiveness and practice-based estimates of teacher effectiveness moderated by student, and, teacher level characteristics?), two-level conditional HLM models are employed to obtain the value-added estimates of teacher effectiveness. Interest centers on the achievement growth parameters, and particularly the differences between the observed and expected achievement growth.

For research question four -- To what extent does the relationship between value-added estimates and practice-based estimates of teacher effectiveness vary by year, by grade, by subject and by type of school? – A Weighted Least Square (WLS) regression analysis is conducted. The criterion variable is the correlation between the value-added estimates, attained from the HLM models results of research questions one through three, and the practice-based estimates of teacher effectiveness indicated by the SLOs quality scores. A normalizing transformation is applied to the estimated correlation coefficients prior to analysis.

For research question five -- To what extent do the SLOs quality scores correspond to the status of how well the SLOs were achieved? -- A logistic regression analysis is performed. Since teachers may have chosen to develop different types of SLOs including Class, Target and Team SLOs, this section of analysis focuses on the type of SLOs that has been chosen by majority of teachers: Class SLOs. Hence, the parameter of interest for the logistic regression analysis focuses on the variable indicating the status of whether the Class SLOs were achieved.

For research question six -- To what extent do teachers' VAM estimates agree with the achievement status of the SLOs? – A point-biserial correlational analysis is conducted. Similar to research question five, the attainment status of Class SLOs is adopted for the analysis. The interest of this correlational analysis centers on both the VAM scores and the SLO achievement status.

3.2.2 Student Level Predictors

At the student level, the student prior achievement scores are first entered in the HLM models. Depending on the quality and accessibility of the dataset, the previous achievement scores that can be drawn on and accounted for in the models are presented by school year and grade level in Table 3.6.

Table 3.6 Student Prior Achievement for Analyses by Grade and Subject

School Year 2010-2011		School Year 2009-2010		School Year 2008-2009	
Grade Level	Prior Achievement Scores Accounted for	Grade Level	Prior Achievement Scores Accounted for	Grade Level	Prior Achievement Scores Accounted for
Grade 8	Grade 5, 6, 7	Grade 8	Grade 6, 7	Grade 8	Grade 7
Grade 7	Grade 4, 5, 6	Grade 7	Grade 5, 6	Grade 7	Grade 6
Grade 6	Grade 3, 4, 5	Grade 6	Grade 4, 5	Grade 6	Grade 5
Grade 5	Grade 3, 4	Grade 5	Grade 3, 4	Grade 5	Grade 4
Grade 4	Grade 3	Grade 4	Grade 3	Grade 4	Grade 3

Including more than one year of student prior test scores could reduce bias and increase precision in the teacher VAM estimates by reducing the variance of the error term in the model, however, the standardized tests in most states, including North Carolina, are not administered until the 3rd grade. Therefore, in this study, including multiple years of student prior score would eliminate the possibility of estimating VAM for an entire grade of teachers because only one year of student prior scores could be available to evaluate the 4th grade teachers. In addition, some students missed the additional prior year of scores for other reasons such as absence on the test day, student transfer, or record missing. Therefore, in order to obtain more complete results of the analyses, students with only one year of prior score are included in the study.

In addition, five student-level demographic variables drawn from the CMS TIF-LEAP project are included in order to examine the extent to which the relationship

between value-added estimates and practice-based estimates of teacher effectiveness is moderated by student contextual characteristics. The student demographic variables are:

- Gender,
- Ethnicity,
- LEP (Limited English Proficiency) status (indicating if a student was ever so designated),
- Gifted status (indicating whether a student was ever so designated), and
- SWD (Student with Disability) status (indicating whether a student was ever so designated and received special education services).

3.2.3 Teacher Level Predictors

At the teacher level, class size[12] is added to the HLM models to explore the extent to which the relationship between value-added estimates and practice-based estimates of teacher effectiveness is moderated by the class characteristics. Due to the limited number of teachers in each grade and year of the dataset, and that teachers' SLOs cannot be adjusted in the similar manner as teachers' VAM estimates if more class characteristics are included in the models, class size is added into the model as the only teacher-level variable at this time. In order to further explore the relationship between VAM and SLOs indicators of teacher effectiveness, a second-stage Weighted Least Square regression analysis is conducted and the influence of other variables including year, grade, and type of school is further studied.

[12] The variable is calculated based on the number of students with the same class ID in the dataset.

3.3 Analytical Strategies

3.3.1 Preliminary Descriptive Analyses

The preliminary descriptive analysis section includes results at both the student level and the teacher level. First, it reports the distributions of the student achievement outcome variables in 2008-2009, 2009-2010 and 2010-2011 school years, respectively. Student achievement scores in school year 2007-2008 are regarded as baseline data and presented as well. The descriptive analyses should compare and contrast the statistical characteristics of the baseline data with those in later years when SLOs were implemented. Results are displayed by subject and grade level.

Further analyses of the distributions of all other student level demographic variables by year constitute the next section of the descriptive analyses. These variables include gender, ethnicity, LEP (Limited English Proficiency) status, SWD (Student with Disability) status and GIFTED status. Data for all TIF-LEAP schools used for teacher VAM analyses, and for the treatment schools where SLOs are implemented, are displayed separately.

Next, preliminary descriptive analyses of all variables at the teacher level are presented. Specifically, the distributions of teachers' SLOs quality scores are summarized for each year, and for mathematics and reading respectively. Some teachers may have developed more than one SLO, either for the entire class or a selection of target groups of students; in some cases the teachers may have worked as a team and established multiple team SLOs for specified groups of students. Therefore, the teachers may have more than one SLOs quality score available. As teachers were required to develop at least one SLO

and most teachers chose to develop Class SLO, the quality scores of the Class SLO are used for all related analyses in this study. The Class SLOs quality scores are analyzed for its statistical descriptive characteristics.

Lastly, descriptive analysis of the status variable indicating whether the SLOs were achieved is conducted. Due to the limited availability and accessibility of the dataset, only Class SLOs are adopted for the logistic regression analysis and the point-biserial analysis although teachers may have developed multiple SLOs of various types and numbers. Consequently, the status of whether teachers' Class SLOs were achieved is elaborated and reported.

3.3.2 Research Question One

The first research question asks: How do value-added estimates of teacher effectiveness based on student test score trajectories compare to the practice-based estimates of teacher effectiveness based on the SLOs quality scores? Because of the multilevel structure of the data, with students nested within teachers, two-level HLMs are constructed to accommodate the dependency among students of the same teacher (Raudenbush & Bryk, 2002). Employing the dataset with multiple years of student achievement records, HLMs are carried out in stages to predict student achievement and obtain the value-added estimates of teacher effectiveness. For a given year and subject, the statistical models generally can be expressed as:

Level 1 (Student level):

$$Y_{ij} = \beta_{0j} + \beta_{1j}X_{1ij} + \cdots + \beta_{pj}X_{pij} + r_{ij}$$

Level 2 (Teacher level): (3.1)

$$\beta_{0j} = \gamma_{00} + u_{0j}$$
$$\beta_{1j} = \gamma_{10} + u_{1j}$$
…
$$\beta_{pj} = \gamma_{p0} + u_{pj}$$

where *i* denotes students within teachers, and *j* indicates teachers;

Y_{ij} is the academic achievement outcome for student *i* within teacher *j*;

X_{1ij}, ..., X_{pij} are *p* student prior achievement scores for student *i* within teacher *j*;

β_{0j} is the mean of the student achievement for teacher *j*, adjusted for the student prior achievement scores X_1, ...,X_p;

β_{1j}, ..., β_{pj} are the regression coefficients for teacher *j*, associated with the covariates X_1,...,X_p;

r_{ij} is the random error (or residual) in the level 1 equation, where $r_{ij} \sim N(0, \sigma^2)$ and σ^2 is the variance of the student-level residuals;

γ_{00} is the intercept for the level 2 equation which is the grand mean of the adjusted teacher-level achievement means across all teachers;

$\gamma_{10}, \ldots, \gamma_{p0}$ are constants representing the common values of the *p* regression coefficients across all teachers; and

$u_{0j}, \ldots, u_{pj}$ are random effects in the level 2 equations, where $u_{pj} \sim N(0, \sigma^2)$ and σ^2 is the variance of the teacher-level residuals.

The value-added estimates for a teacher is represented by the difference between the observed adjusted mean for that teacher and the grand mean of adjusted means, indicated by u_{0j} in the equation.

The first research question focuses on the association between two sets of teacher effectiveness indicators: value-added estimates and the SLOs quality scores; hence the teacher contextual characteristics are not included while student background characteristics will be entered at a later stage. In the process of building models, each model is carried out in stages. In the first stage, unconditional models are developed and displayed, which allows partitioning of the total variability in student achievement into within and between teacher variance components.

In the second stage, student achievement scores measured in the prior years are added to the models of different years respectively. It is hypothesized that a student's achievement scores across different years are highly correlated; therefore, this model should explain a large percent of the available variance in current student achievement outcomes. The value-added estimates of teacher effectiveness are calculated from the models at this stage, and the relative ranks of teachers in the experimental group are compared to their SLOs quality scores. Subsequently, a Spearman correlation is calculated based on the two sets of teacher effectiveness indicators.

The Spearman correlation coefficient is defined as the Pearson correlation coefficient between the ranked variables (Myers & Well, 2003). The correlation coefficient takes the statistical form of:

$$\rho = \frac{\sum_i (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_i (x_i - \bar{x})^2 (y_i - \bar{y})^2}}$$

A simpler procedure using differences between the ranks of each observation on the two variables can also be employed to calculate ρ when duplicate values (ties) are known to be absent. The statistical form of this procedure is:

$$r = 1 - \left(\frac{6 \sum d^2}{n(n^2 - 1)} \right)$$

where d is the differences between the ranks of each observation on the two variables;
n is the number of pairs of cases.

One statistical form of the correlation coefficient is adopted in this study depending on the actual data condition (i.e. whether there are ties in the data). In addition, the standard errors of the Spearman's correlation coefficients are reported as well. The equation provided by Glass and Hopkins (1995, p350) is adopted to estimate the standard errors:

$$s_r = \sqrt{\frac{1 - r^2}{n - 2}}$$

where s_r is an estimate of the standard error of r,
n is the number of pairs of scores.

3.3.3 Research Question Two

The second research question asks: To what extent is the relationship between value-added estimates of teacher effectiveness and practice-based estimates of teacher effectiveness affected by student characteristics? To answer this question, similar HLM models are constructed from the final model in the first research question (i.e., the models that control for student prior achievement scores).

In this model, additional student-level covariates (e.g., gender, race, LEP status, SWD status, GIFTED status) are added one by one to explore the associations between student achievement outcome and students' characteristics. Since all student-level demographic variables have been recoded as dichotomous dummy variables, they are placed into the models uncentered during the analysis. The final model at this stage only includes variables that are statistically significant at the .05 level[13].

In addition, the relationships between the level-1 predictors and the student achievement outcome should be examined across teachers. If there is no significant variation in the level-1 slopes across teachers, the level-1 slopes are fixed. Otherwise, the level-1 slopes are allowed to vary. Therefore, the final model may be different from the one at the original and intermediate steps in terms of incorporating certain fixed level-1 slopes. The value-added estimates of teacher effectiveness derived from the final model at this stage are used to obtain the relative rankings of teachers within their peer group. Likewise, a Spearman correlation is computed between teacher's rankings based on their value-added estimates and their SLOs quality scores. Results from this analysis are

[13] The significance level is set at .05 throughout the analyses in this dissertation.

compared with those from the first research question.

3.3.4 Research Question Three

The third research question asks: To what extent is the relationship between value-added estimates of teacher effectiveness and practice-based estimates of teacher effectiveness affected by teacher-level contextual characteristics? To answer this question, new HLMs are constructed from the final model in research question two (i.e., the models that control for student prior achievement and student-level covariates). These statistical models take the general form of:

Level 1 (Student level):

$$Y_{ij} = \beta_{0j} + \beta_{1j}X_{1ij} + \cdots + \beta_{pj}X_{pij} + r_{ij}$$

Level 2 (Teacher level): (3.2)

$$\beta_{0j} = \gamma_{00} + \gamma_{01}W_{1j} + \cdots + \gamma_{0q}W_{qj} + u_{0j}$$

$$\beta_{1j} = \gamma_{10} + \gamma_{11}W_{1j} + \cdots + \gamma_{1q}W_{qj} + u_{1j}$$

…

$$\beta_{pj} = \gamma_{p0} + \gamma_{p1}W_{1j} + \cdots + \gamma_{pq}W_{qj} + u_{pj}$$

where W_{1j}, …, $W_{qj}$are q teacher-level covariates for teacher j; and

γ_{01}, ..., γ_{pq} are the q(p+1) regression coefficients associated with the teacher level covariates W_{1j}, …, W_{qj}.

Comparing equations 3.1 and 3.2, the only difference is that teacher-level contextual variables are accounted for at level-2. Teacher level covariates W_{1j}, …, W_{qj} are entered to explore the associations between the level-1 intercept (i.e., student achievement outcome), level-1 slopes, and teacher/class characteristics. The model building process involves the following new stages: Teacher level covariates, to start with, are added at level-2 to predict the intercept of level-1. In the next stage, teacher-level covariates are included to explain the variance in the level-1 slopes, when a level-1 slope is allowed to vary across teachers. It is worth noting that the HLM models in the above equations can represent a full model that takes into account all possible covariates and interactions (i.e., an intercepts- and slopes-as-outcomes model). Depending on the results of the analyses, the final model probably takes a much simpler form from what is presented here. The value-added estimates of teacher effectiveness, thereafter, are attained from the concluding model at this stage, which then produces the relative positions of teachers within the same peer group. Correspondingly, teachers' VAM rankings are compared with the their SLOs quality scores in the Spearman's correlation analysis, results from which are compared and contrasted with those from the second research question, and the impact of incorporating the teacher level covariates into the models can be scrutinized at this point.

3.3.5 Research Question Four

The fourth research question asks: To what extent does the relationship between value-added estimates of teacher effectiveness and practice-based estimates of teacher effectiveness vary by year, by grade, and by type of school? To answer this question, a

Weighted Least Square (WLS) analysis is performed. This method takes the statistical form of:

$y_{i,j} = \mu_j + \varepsilon_{i,j}$

where

$y_{i,j}$ are observations;

μ_j is the group mean of the observations;

$\varepsilon \sim N(0, \sigma^2)$, ε are random errors.

One assumption in standard linear regression models is the constant variance within the population under study. However, the group size of teachers in this study greatly varies by year and grade, which probably generates the issue of unequal sample sizes in ANOVA that may cause different standard errors estimated for the different groups of teachers. Under such circumstances, Ordinary Least Squares (OLS) no longer provides optimal model estimates. Therefore, the Weighted Least Squares (WLS) procedure is performed to compensate for different precisions of measurement. The Weight Estimation procedure can test a range of weight transformations and indicate which one may give the best fit to the data.

In order to stabilize the variance of the correlation coefficients (Fisher, 1915), Fisher Z-transformation $\left(z = \frac{1}{2} ln\left(\frac{1+r}{1-r}\right)\right)$ is applied to the results of the Spearman's correlation analysis from the first three research questions. Transformed Spearman correlation coefficients, along with the estimated variance of z obtained from the transformation process, are treated as the outcome variables in this WLS analysis. Factors

comprise year (i.e., 2007-2008, 2008-2009, 2009-2010), grade (i.e., grade 4 through 8), the type of school (i.e., elementary and middle schools) at this stage.

Results from the main effects and interactions WLS analyses demonstrate the extent to which the correlations between the two types of teacher effectiveness indicators vary by these characteristics. Analyses are conducted for mathematics and reading respectively. A post-hoc analysis is conducted when a significant difference is detected.

3.3.6 Research Question Five

The fifth research question asks: To what extent do the SLOs quality scores correspond to whether the SLOs have been achieved? To answer this question, a logistic regression analysis is conducted. This method takes the statistical form of:

$$\mathrm{Y} = \ln(o) = \ln\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 \mathrm{X} + \varepsilon$$

where

Y= (0, 1);

X is the SLOs quality score;

p is the probability that the event Y occurs, *p*(Y=1);

o or $\frac{p}{1-p}$ is the "odds ratio";

$\ln\left(\frac{p}{1-p}\right)$ is the log odds ratio, or "logit";

ε are random errors.

Given that teachers may have developed different types and numbers of SLOs in this study, sufficient data capable of supporting complicated statistical analysis is not available for every type of SLOs. Consequently, the most commonly developed SLOs - Class SLOs are employed in the logistic regression analysis to predict the probability of whether the SLOs were achieved. The status of whether the SLOs were met is treated as the outcome variable, while the SLOs quality scores serve as the predictor in the logistic regression model. This analysis reveals the association between teachers' ability to develop the written SLOs and their aptitude to facilitate students achieving those SLOs. Results of the analysis should also suggest the extent to which the quality of teachers' SLOs can be indicative of teachers' effectiveness in fostering student learning.

3.3.6 Research Question Six

The sixth research question asks: To what extent do teachers' VAM estimates agree with the achievement status of the SLOs? To answer this question, a point-biserial correlational analysis is conducted in which the SLO achievement status is considered a dichotomous variable and the variable of teachers' VAM estimates is treated as a continuous one. The statistical form of this method is as follows.

$$r_{xy} = \frac{M_1 - M_0}{s_{n-1}} \sqrt{\frac{n_1 n_0}{n(n-1)}}$$

The standard deviation for data sample, s_{n-1}, is obtained using the statistical form of:

$$s_{n-1} = \sqrt{\frac{1}{n-1} \sum_{i=1}^{n} (X_i - \bar{X})^2}$$

where

M_1 is the mean value on the continuous variable;

M_0 is the mean value on the dichotomous variable;

n_1 is the number of data points in group 1;

n_0 is the number of data points in group 2;

n is the total sample size.

Compared to the rankings of teacher's VAM estimates used in the previous research questions, the actual teachers' VAM scores are employed in the last research question for the correlational analysis with the SLO achievement status. This analysis reveals the association between teachers' ability to foster student's academic achievement and their aptitude to facilitate students achieving the SLOs. Results of the analysis should help address the triangular relationship among teacher's VAM estimates, their SLO quality and SLO attainment status.

CHAPTER 4. RESULTS

This chapter presents results from the analyses outlined in chapter three. It is organized into seven sections. The first section reports results from the descriptive analyses, comprising the distributions of the student achievement outcomes including the prior test scores that were accounted for in the models, a summary of the students' background variables, as well as class level characteristics. In addition, this section further reports the distributions of SLOs quality scores and whether the SLOs have been achieved. The results are presented by year, grade, and subject, respectively.

Each of the next six sections answers one of the research questions in this study. In sections two, three and four, hierarchical linear models were carried out in stages and the results reported at the end of each stage. Because there were only limited number of teachers who had mathematics and reading SLOs quality scores within each grade and year, the HLM models were first built among all teachers (including those with and without SLOs quality scores) in the experimental schools of the TIF-LEAP project so as to obtain more stable VAM estimates. The VAM rankings of the target teachers were then extracted from those results to be analyzed in conjunction with the SLO quality scores. Section five reports the results from the weighted least square analyses, presenting the extent to which the relationships between teachers' VAM rankings and SLOs quality scores vary by year, grade and the type of school. In addition, section six reports the results from the logistic regression analysis, which quantified the relationship between teachers' SLOs quality scores and whether these objectives were achieved. Finally, the

last section reports the results from the point-biserial correlation analysis examining the relationship between teachers' VAM estimates and the attainment status of their SLOs.

4.1 Descriptive analyses

4.1.1 The Value-added Analyses

4.1.1.1 Student Achievement Outcomes

The descriptive analyses of student achievement were conducted at three levels of aggregation – overall, by year, and by grade, for mathematics and reading, respectively. At the overall level, the descriptions of student achievement were based on the scores of the entire student group, while at the other levels, the calculations were based on the aggregated (average) scores of different years and grades for both subjects, respectively. The descriptions at various levels provide a comprehensive view of the student achievement outcomes in this study.

Overall level

In this study, there are 48950 student records for mathematics across five years (school years 2006-2007, 2007-2008, 2008-2009, 2009-2010, and 2010-2011) and six grades (Grade 3 through 8), and 50542 records for reading through the same years and grades. In the dataset there were students with incomplete records across multiple school years. The reasons for the missing records were varied; for example, students may have

moved out of the state during that period, or some students did not take the test. Since the number of records missing was small for each year and grade (less than 5 percent), only the valid scale scores were used for the analysis. No vertical scaling was applied to the school district's test scores, and the achievement scores were separately scaled by year and grade for both mathematics and reading (see section 3.1). Table 4.1 and 4.2 summarize the means and standard deviations of student achievement outcomes by year and grade for mathematics and reading, respectively.

Table 4.1 The Means and Standard Deviations of Student Achievement Scores by Year and Grade – Mathematics

	2006			2007			2008			2009			2010		
	N	Mean	s.d	N	Mean	s.d.	N	Mean	s.d.	N	Mean	s.d.	N	Mean	s.d.
Grade 3	1	328.00	-	1545	338.46	8.75	1151	339.85	8.99	1221	340.17	9.25	1435	339.70	9.37
Grade 4	-	-	-	1656	345.41	8.59	1808	346.27	8.45	1178	347.33	8.83	1239	347.80	8.76
Grade 5	7	350.71	8.94	1624	351.51	8.29	1864	351.83	8.28	1963	352.59	8.49	1287	353.40	8.35
Grade 6	2318	351.21	8.57	2183	351.79	8.63	1853	351.45	8.19	2053	352.51	8.50	2196	352.60	8.53
Grade 7	1722	351.79	7.34	2465	354.65	8.94	2322	355.63	8.89	2007	355.19	8.49	2282	355.59	8.39
Grade 8	1369	354.69	6.61	1781	355.36	6.77	2312	357.83	7.34	2215	359.05	7.04	1893	357.76	6.85

Table 4.2 The Means and Standard Deviations of Student Achievement Scores by Year and Grade -- Reading

	2006[14]			2007			2008			2009			2010		
	N	Mean	s.d.	N	Mean	s.d.	N	Mean	s.d.	N	Mean	s.d.	N	Mean	s.d.
Grade 3	1	240.00	-	1528	332.08	10.35	1147	333.28	11.01	1206	333.09	10.75	1419	333.25	10.55
Grade 4	-	-	-	1626	339.76	9.08	1786	340.26	8.84	1163	340.83	9.24	1214	340.94	9.06
Grade 5	6	252.33	6.38	1592	345.11	8.34	1850	345.60	8.23	1929	345.88	8.15	1263	346.74	8.19
Grade 6	2275	255.51	7.69	2168	348.47	9.31	1815	347.92	8.78	2000	348.86	8.50	2128	349.23	8.02
Grade 7	2116	257.89	8.29	2449	351.43	8.49	2307	352.29	8.60	1964	351.80	8.54	2219	352.85	8.14
Grade 8	1950	261.48	7.45	2266	354.66	8.00	2561	355.27	8.04	2435	356.28	7.89	2159	355.33	7.89

By year and grade

As indicated in Table 4.1 and 4.2, the means and standard deviations of the scale scores were, in general, quite stable across years for both mathematics and reading. In addition, depending on the scaling[15], the mean scores demonstrated an overall rising trend from lower to higher grades for both subjects, and across all years. The distribution of the standard deviations indicated slight variations among different grade levels; for example, the standard deviations in grade 8 were lower than those in other grades for both mathematics and reading, and across different years. The focal analyses in this study were conducted using student achievement scores from grade 4 through 8 of school year 2008-2009, 2009-2010, and 2010-2011 (referred to as 2008, 2009 and 2010 in the following sections) for the value-added analysis. Scores in other years and grades were

[14] The TIF-LEAP research project focused on evaluating teachers in 2008, 2009 and 2010, and student achievement data in early grades of 2006 was incomplete. The reading scores in 2006-2007 used a different scale.

[15] The developmental scales in the EOG test are based on IRT estimates of differences between adjacent-grade means and ratios of adjacent-grade standard deviations. See Chapter 3, p71.

used as student prior achievement covariates that were incorporated in the HLM models.

4.1.1.2 Student-level Variables

As described in chapter three, five types of student-level variables were included in this study: gender, race ethnicity, LEP (Limited English Proficiency) status (indicating if a student was ever so designated), GIFTED status (indicating whether a student was ever so designated), and SWD (Student with Disability) status (indicating whether a student was ever so designated and received special education services).

Table 4.3 provides a summary describing all the student-level variables in the HLM analysis. Results shown in the table were averaged across grades 4 through 8 and displayed by year and subject. In terms of the gender distribution, there was approximately the same number of male and female students in different years for both subjects. Across years and subjects over 60% of the students were African American, followed in descending proportions by Hispanic, White, Asian, Multi-race, and American Indian students. The three most numerous race/ethnicity classes (African American, Hispanic, and White) contained approximately 90% of the whole population in the dataset. One thing to note about the race distribution was that the number of African American students increased slightly across three years for both subjects, while the number of White students, on the contrary, decreased noticeably during this time period.

Table 4.3 Description of Student-level Variable: Average Gender Percentage across Grades

Mean % of students across Grades 4-8	Mathematics			Reading		
	2008	2009	2010	2008	2009	2010
By Gender						
Female	49.2%	49.1%	48.7%	49.3%	49.5%	49.3%
Male	50.8%	50.9%	51.3%	50.7%	50.5%	50.7%
By Ethnicity						
American Indian	0.6%	0.6%	0.6%	0.6%	0.6%	0.6%
Asian	3.9%	4.7%	4.8%	3.9%	4.1%	4.4%
African American	60.6%	60.7%	64.1%	60.3%	60.8%	64.8%
Hispanic	22.7%	23.7%	22.4%	22.2%	23.1%	21.7%
Multi-race	2.4%	2.4%	2.3%	2.4%	2.6%	2.5%
White	9.7%	7.8%	5.9%	10.5%	8.9%	6.0%
By Special Types						
Gifted	4.1%	3.9%	3.5%	4.6%	4.7%	3.9%
LEP	13.4%	16.6%	15.1%	12.2%	14.3%	14.0%
SWD	8.3%	7.8%	8.4%	7.7%	7.1%	7.5%

On average, there were approximately 7-8% of SWD students and 4% of GIFTED students among all years and subjects. About 15% of mathematics students and 13% of reading students had limited English proficiency across different years, and this number increased from 2008 through 2010 for both subjects.

4.1.1.3 Teacher-level Characteristics

As explained in chapter 3, due to the limited number of level-2 records available in the dataset, class size was adopted as the only teacher-level covariate in the HLM

models. Table 4.4 summarizes the means and standard deviations of this variable across all years and grades for both mathematics and reading. The summaries illustrated in the table reflect the average class size for the teachers teaching different subjects in various years and grades.

Generally, class size[16] ranged from 5 to 30 with an overall mean of approximately 15 across all grades and years for both subjects. The class size has been roughly consistent across grades for all three years. When compared across years, as shown in Table 4.4, the class size increased over time, especially in 2010. For example, the mean class size for mathematics teachers at each grade increased from an average of 11.7 in 2008 to 13.5 in 2009, and to 20.1 in 2010. Possible reasons for this increase were explained in chapter 3 (section 3.1.4).

Table 4.4 Description of Teacher-level Variable: Means and Standard Deviations of Class Size

	Mathematics			**Reading**		
	2008	2009	2010	2008	2009	2010
Grade 4	11.12(3.06)	13.43(3.49)	18.00(2.03)	11.06(3.05)	13.22(3.46)	17.89(2.26)
Grade 5	9.72(2.76)	11.14(3.42)	19.14(3.34)	9.72(2.76)	11.50(3.20)	18.86(3.40)
Grade 6	14.30(3.70)	16.80(2.96)	23.07(3.60)	12.01(4.01)	14.23(3.97)	18.65(6.04)
Grade 7	11.45(2.18)	14.85(4.38)	22.18(3.53)	10.17(2.48)	14.11(5.16)	20.21(4.28)
Grade 8	11.93(2.46)	11.14(4.38)	22.90(4.41)	10.73(2.48)	10.32(4.20)	20.79(4.73)

[16] The class size refers to the number of students that contributed to the teacher's VAM scores in this study. Smaller class sizes could indicate substantial student mobility.

4.1.2 The SLO Analysis

4.1.2.1 SLO Quality Scores

As introduced in chapter 3, three types of SLOs (Class SLO, Team SLO, and Target SLO) were implemented in the TIF-LEAP project. They served different purposes for evaluating teacher performance based on various groupings of students. Class SLO was designed for the entire class of students, while teachers typically built a Target SLO for a small group of students based on their characteristics and requirements, and Team SLO was devised to develop goals for a class or group of students through the collaborations among multiple teachers.

During the project implementation process, teachers were required to create at least one SLO for each student, and most teachers chose to establish Class SLOs. Therefore, considerably more data was available for the Class SLOs than for the Team and Target SLOs. In order to minimize issues with insufficient data and to avoid potential confusion with combining different types of SLOs, this study only employed teachers' Class SLO quality scores and attainment status for further analyses.

Table 4.5 Overall Descriptions of SLO Quality Scores

	N	Minimum	Maximum	Mean	s.d.
Mathematics	243	2.0	4.0	3.43	.62
Reading	259	2.0	4.0	3.39	.62

Table 4.5 displays statistical summaries for the SLO quality scores used for this study. The SLO quality scores ranged from 2 to 4, with the overall mean of 3.43 for

mathematics and 3.39 for reading. A frequency distribution of the numbers of SLO scores across grades and years is provided in Table 4.6. These are the actual numbers of teachers available for the relationship analysis among VAM estimates, SLO quality scores, as well as the SLO attainment status.

Table 4.6 Frequency of SLO Quality Scores by Subject, Grade, and Year

	Mathematics			Reading		
	2008	2009	2010	2008	2009	2010
Grade 4	17	23	19	17	23	19
Grade 5	18	21	21	18	20	21
Grade 6	17	16	9	22	20	12
Grade 7	13	16	13	14	18	16
Grade 8	11	15	14	10	15	14
Total	76	91	76	81	96	82

4.1.2.2 SLO Attainment Status

Figures 4.1 and 4.2 summarize the overall SLOs attainment status in each grade and year for mathematics and reading, respectively. Cell entries denote the percent of SLOs that were achieved. The tables indicate that the SLOs across all years were more likely to have been achieved in lower grades than in higher grades for both mathematics and reading. In addition, greater percentages of SLOs were attained in 2009 than in the other two years across all grades and subjects, which was probably due to the impact of ongoing project progress and other school related factors during the implementation of the project. More details were explained in chapter 3 (section 3.1.4).

Figure 4.1 SLOs Attainment by Year and Grade for Mathematics

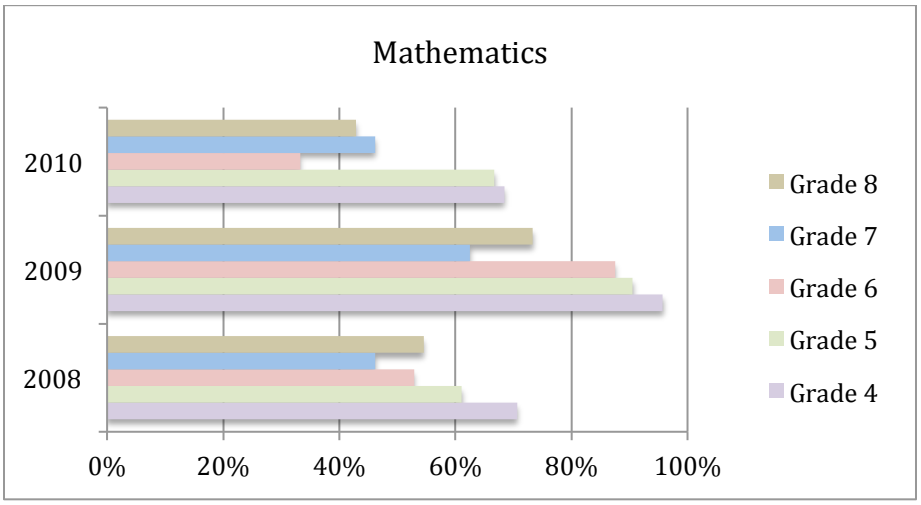
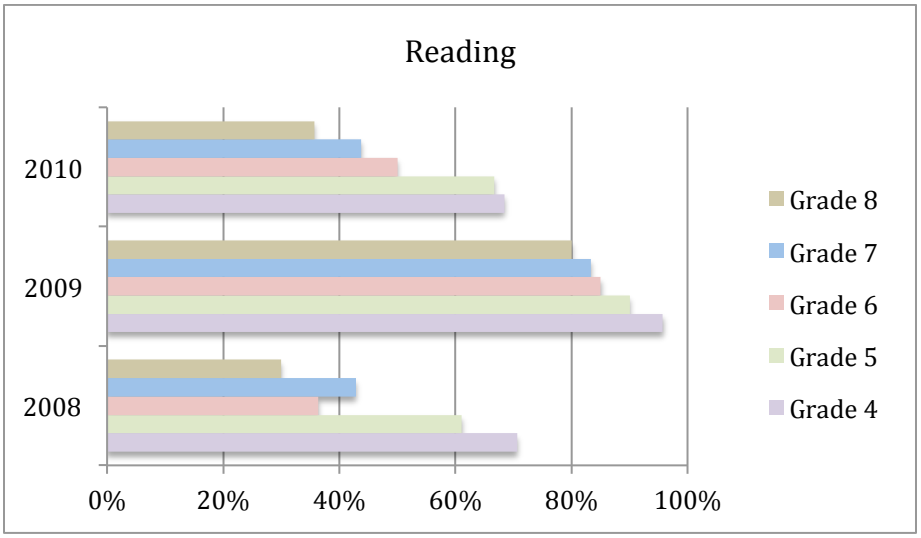


Figure 4.2 SLOs Attainment by Year and Grade for Reading



4.2 Research Question One

In order to explore the associations between the estimates of teacher effectiveness based on VAM and SLO quality, the longitudinal dataset with students' achievement records was first used for building hierarchical linear models and obtaining the value-added estimates for teachers in different subjects, years and grades. Estimation of hierarchical linear models was carried out in two stages. In the first stage, an unconditional model with no predictors at either the student or teacher level was built. The intra-class correlation coefficient (ICC= $\frac{\tau_{00}}{\sigma^2+\tau_{00}}$) was calculated for each year and grade, for both mathematics and reading. The results are summarized in Table 4.7. Taking the model for mathematics-2008-Grade 4 as an example, ICC= $\frac{\tau_{00}}{\sigma^2+\tau_{00}} = \frac{11.94}{11.94+56.99} = 0.17$.

Table 4.7 Summary of Intra-class Correlation Coefficients by Year and Grade for Mathematics and Reading

	Mathematics			Reading		
	2008	2009	2010	2008	2009	2010
Grade 4	0.17	0.15	0.13	0.09	0.10	0.08
Grade 5	0.19	0.17	0.06	0.15	0.12	0.04
Grade 6	0.23	0.18	0.20	0.25	0.21	0.11
Grade 7	0.37	0.18	0.14	0.26	0.25	0.25
Grade 8	0.24	0.21	0.07	0.24	0.27	0.21

As indicated in Table 4.7, the overall mean ICC across years and grades was 0.18 for both mathematics and reading, which means that 18% of the total variance in student achievement was due to the variance between teachers. Because students were nested within teachers/classes, multilevel models were needed throughout the study because of their ability to account for the hierarchical structure of the data, as well as the ability to provide more appropriate standard error estimates and the statistical significance levels of the results.

At stage two, the students' achievement scores in the same subject from the prior years (one to three prior test scores for each student depending on the grade and year) were group-mean centered and added to the models. As expected, student prior achievement scores explained large proportions of variance in student achievement scores. Table 4.8 displays the amount of variance explained by student prior achievements compared to the unconditional models ($\frac{\sigma^2_{unconditional}-\sigma^2_{conditional}}{\sigma^2_{unconditional}}$). An average of 63% of the variance in student achievement scores for both mathematics and reading across years and grades were explained by their prior test scores. In other words, taking the model for mathematics-2008-Grade 4 as an example, 55% of the total variance within teachers was explained in this model. Models with multiples years of student prior scores are expected to account for more variance in student achievement; however, the increasing number of prior scores as covariates for later grades (e.g. grade 7 and 8) does not add explanatory power overall. This is somewhat unexpected and there could be multiple reasons for it, for example, more tracking at higher grade levels may lead to more between-class variations.

Table 4.8 Amount of Variance Explained by Adding Student Prior Achievement into the Models

	Mathematics			Reading		
	2008	2009	2010	2008	2009	2010
Grade 4	55%	61%	64%	60%	64%	61%
Grade 5	63%	68%	73%	57%	65%	68%
Grade 6	56%	70%	69%	59%	63%	69%
Grade 7	55%	67%	69%	56%	67%	68%
Grade 8	56%	54%	63%	62%	61%	71%

Depending on the data availability, different numbers of student prior test scores were included in the models. An example of the final model (mathematics-2008-Grade 4) developed in this section can be expressed as

Level-1 Model (Student level) (4.1)

$$Math\_2008\_G4\_Score_{ij} = \beta_{0j} + \beta_{1j}*(Math\_2007\_G3\_Score_{ij}) + r_{ij}$$

Level-2 Model (Teacher level)

$$\beta_{0j} = \gamma_{00} + u_{0j}$$

$$\beta_{1j} = \gamma_{10}$$

where *i* denotes students within teachers, and *j* indicates teachers.

This is an example of the base model used in the next section to answer research question two. The complete model equations specifications for all years, grades and subjects are given in Appendix A.

In addition, the relationships between the level-1 predictors and the student achievement outcome were examined across teachers. If there were no significant variation in the level-1 slopes across teachers, the level-1 slopes were fixed. Otherwise, the level-1 slopes were allowed to vary.

Teachers' value-added estimates, as well as their relative rankings, were obtained through calculating the residuals at Level-2 of the models at this stage to answer research question one. Those records that could be matched with teachers' SLOs quality scores were extracted and used for the correlation analysis between VAM rankings and SLOs quality scores. Table 4.9 summarizes the results of estimating the relationships between teachers SLOs quality scores and their VAM estimates, using Spearman's correlation coefficient, after controlling for student prior achievement scores. The standard error of the correlation coefficients were calculated based on the equation $s_r = \sqrt{\frac{1-r^2}{n-2}}$ (Glass and Hopkins, 1995).

Results are displayed by year and grade for both mathematics and reading. The correlation coefficients ranged from -.38 to .46 for mathematics and -.35 to .51 for reading. Both positive and negative correlations were found for different years and grades though most coefficients were positive. Specifically, of the 30 estimated coefficients, there were 5 negative correlations for mathematics and 4 for reading. No noticeable patterns for the correlations were found. Given the limited number of records in each cell for the correlation analysis, only one estimated correlation coefficient reached statistical significance.

Table 4.9 Summary of Spearman's Correlation Coefficients (standard errors) between VAM (based on models with Student Prior Achievement Accounted for) and SLO Quality by Year and Grade for Mathematics and Reading

	Mathematics			Reading		
	2008	2009	2010	2008	2009	2010
Grade 4	0.12(.26)	0.12(.22)	-0.07(.24)	0.03(.26)	0.25(.21)	-0.35(.23)
Grade 5	-0.30(.24)	0.44*(.21)	0.14(.23)	-0.18(.25)	0.37(.23)	0.35(.22)
Grade 6	-0.32(.24)	-0.14(.26)	0.31(.36)	0.00(.22)	-0.33(.22)	0.03(.32)
Grade 7	0.46(.27)	0.10(.27)	0.29(.29)	0.44(.26)	-0.19(.25)	0.03(.27)
Grade 8	-0.38(.31)	0.24(.27)	0.18(.28)	0.00(.38)	0.51(.24)	0.00(.30)

*Correlation is significant at the 0.05 level (2-tailed).

4.3 Research Question Two

In order to investigate how the inclusion of student-level characteristics impacts the relationship between teachers' SLOs quality scores and their VAM estimates after controlling for the prior achievement scores, student-level covariates were first added by group into the final models from the first research question. After examining all the variables by group, all retained variables were added to the next models and only the statistically significant (at the .05 level) variables were retained. If there was significant amount of variations in the level-1 slope estimates and the slopes could be reliably

estimated[17], they were allowed to vary across teachers. Example results from the final models to answer research question two are presented in Table 4.10, and the complete models for all years, grades, and both subjects are contained in the Appendix B.

Table 4.10 shows an example of the final model results at this stage: 2008 Grade 4 models for both mathematics and reading. The fixed regression coefficients are at the top of the table. After adjusting for all other student-level covariates, students' 4th-grade scores in 2008 and their prior achievement (3rd-grade scores in 2007) were still strongly associated (0.68, $t(1039) = 31.35$, $p<.001$, deviance = 7581 for mathematics; 0.62, $t(1015) = 33.30$, $p<.001$, deviance = 7458 for reading). Compared with the coefficients from the base model in which only student prior test scores included (0.74, $t(1047) = 30.44$, $p<.001$, deviance = 7647 for mathematics; 0.67, $t(1023) = 36.27$, $p<.001$, deviance = 7525 for reading), this model is effective and demonstrates better model fit[18].

Among the five types of student-level covariates, race/ethnicity subgroups, GIFTED status, LEP status and SWD status were all statistically significant for both mathematics and reading. Gender was the only student demographic variable that was not significantly associated with students' achievements in Grade 4 of 2008 after controlling for their prior achievement in 2007. Among race/ethnicity subgroups, the mean difference between Hispanic and African American students was 1.97 points for mathematics and 1.20 for reading, which suggests that in a typical situation, Hispanic

[17] Because the number of units at level 2 is much smaller than that at level 1, the slope estimates can be far less reliable. As suggested from past experiences (Raudenbush & Bryk, 2002), a slope would be fixed when the reliability of a random level-1 coefficient drops below 0.05 (see Raudenbush & Bryk, 2002, p.125).

[18] Model deviance can indicate the level of model fit. The deviance of 7581 for mathematics for the current model is lower than 7647 for the base model; similarly the deviance of 7458 for this model is lower than 7525 for base model for reading, which indicates the models at this stage are better fit for both mathematics and reading.

students scored higher in both 2008-Grade 4 mathematics and reading EOG tests than their African American counterparts after adjusting for their achievement scores in the previous year (Grade 3 in 2007), and other student characteristics (1.97, $t(1039) = 4.29$, $p<.001$ for mathematics; 1.20, $t(1015) = 2.51$, $p<.001$ for reading).

It is interesting to note that the White students scored 0.77 points lower than the African American students in this model, which is not consistent with most literature findings. Regarding the comparisons between the test scores of White and African American students in the entire study across all models based on grade and year (see Appendix A), the scores of White students were found to be higher than their African American counterparts in most models. As this finding in the model of 2008-Grade 4 reading was not statistically significant ($p=0.40$), it could simply be due to chance.

Regarding the other three types of statistically significant student background covariates, the mean score difference between students who were ever considered gifted by their schools and those who were not was 3.72 for mathematics and 3.33 for reading. This suggested that the gifted students scored higher in the 2008-Grade 4 mathematics and reading EOG tests than those who were never considered gifted, even after adjusting for their achievements in the previous year (Grade 3 in 2007), and other student characteristics (3.72, $t(1039) = 3.90$, $p<.001$ for mathematics; 3.33, $t(1015) = 3.49$, $p<.001$ for reading). Similarly, students with limited English proficiency (LEP) scored lower in the tests than students who were fully proficient in English, after controlling for their 3rd Grade achievements in 2007 and other student characteristics. The mean difference was 1.56 points in mathematics and 2.84 points in reading (-1.56, $t(1039) = -2.80$, $p<.001$ for mathematics; -2.84, $t(1015) = -4.78$, $p<.001$ for reading).

Table 4.10 Results of HLMs for Research Question Two: 2008-Grade 4

	Mathematics			Reading		
Level-1 Fixed Effect[a]	Coef.	s.e.	*p*-value	Coef.	s.e.	*p*-value
Intercept, γ_{00}	346.19	0.37	**<0.001**	340.58	0.37	**<0.001**
2007_Grade3_Score, γ_{10}	0.68	0.03	**<0.001**	0.62	0.02	**<0.001**
Student Race/Ethnicity[b]						
American Indian, γ_{20}	-0.65	3.84	0.87	3.61	3.05	0.24
Asian, γ_{30}	-0.14	0.81	0.86	0.40	0.82	0.63
Hispanic, γ_{40}	1.97	0.36	**<0.001**	1.20	0.49	**<0.05**
Multi-race, γ_{50}	1.09	0.67	0.10	0.14	0.93	0.88
White, γ_{60}	0.30	0.78	0.70	-0.77	0.92	0.40
Gifted status, γ_{70}	3.72	0.85	**<0.001**	3.33	0.81	**<0.001**
LEP status, γ_{80}	-1.56	0.52	**0.003**	-2.84	0.58	**<0.001**
SWD status, γ_{90}	-2.49	0.57	**<0.001**	-2.92	0.58	**<0.001**

Random Effect	Variance Component	df	χ^2	*p*-value	Variance Component	df	χ^2	*p*-value
Intercept, u_0	14.77	163	872.28	**<0.001**	12.86	156	667.51	**<0.001**
Level-1 effect, γ_{ij}	24.80				27.05			
Reliability estimate								
Intercept	0.786				0.753			

Note. Bolded values are significant at .01

[a] Fixed effects at level 2 is not reported here, since they are not of research interests in research question one; [b] Reference group is African American.

Moreover, the model results indicated that students' disability status (SWD) was a significant predictor as well. The mean difference between SWD students and their counterparts was 2.49 points for mathematics and 2.92 for reading, which implied that in the 4th-grade mathematics and reading EOG tests in 2008, SWD students scored lower than non-SWD students after adjusting for their achievement scores in the previous year,

and other student characteristics (-2.49, $t(1039) = -4.26$, $p<.001$ for mathematics; -2.92, $t(1015) = -4.55$, $p<.001$ for reading).

Table 4.10 also lists the variance components in the mean student achievement scores of the 4th grade in 2008 (i.e. 14.77 for mathematics and 12.86 for reading). The significant χ^2 statistics associated with mean 4th-grade achievements in 2008 ($p<.001$) indicate that significant differences existed among the teachers' means on their students' 4th-grade mathematics and reading achievement levels in 2008. No statistical significance was found with the mean slopes, meaning that the relationships between the students' 4th grade achievement in 2008 and any student-level covariates did not vary across all teachers. Therefore, all slopes were fixed in this model.

Compared to the model specifications in research question one (eq. 4.1), the final model developed at this step can be expressed as (taking mathematics-2008-Grade 4 as an example):

Level-1 Model (Student level) (4.2)

$$\begin{aligned} Math\_2008\_G4\_Score_{ij} = {} & \beta_{0j} + \beta_{1j}*(Math\_2007\_G3\_Score_{ij}) \\ & + \beta_{2j}*(AmericanIndian_{ij}) + \beta_{3j}*(Asian_{ij}) + \beta_{4j}*(Hispanic_{ij}) \\ & + \beta_{5j}*(Multi\text{-}race_{ij}) + \beta_{6j}*(White_{ij}) + \beta_{7j}*(GIFTED_{ij}) \\ & + \beta_{8j}*(LEP_{ij}) + \beta_{9j}*(SWD_{ij}) + r_{ij} \end{aligned}$$

Level-2 Model (Teacher level)

$$\beta_{0j} = \gamma_{00} + u_{0j}$$

$$\beta_{1j} = \gamma_{10}$$

$\beta_{2j} = \gamma_{20}$

$\beta_{3j} = \gamma_{30}$

$\beta_{4j} = \gamma_{40}$

$\beta_{5j} = \gamma_{50}$

$\beta_{6j} = \gamma_{60}$

$\beta_{7j} = \gamma_{70}$

$\beta_{8j} = \gamma_{80}$

$\beta_{9j} = \gamma_{90}$

This is an example of the base model used in the next section to answer research question three. The complete model specifications for all years, grades and for both subjects are presented in the Appendix B.

Table 4.11 Amount of Variance Explained by Adding Student-level Demographics into the Models (Numbers in parentheses indicate the % of variance explained by the models with Student Prior Scores only)

	Mathematics			**Reading**		
	2008	2009	2010	2008	2009	2010
Grade 4	56% (55%)	63% (61%)	65% (64%)	61% (60%)	65% (64%)	62% (61%)
Grade 5	65% (63%)	68% (68%)	73% (73%)	59% (57%)	65% (65%)	68% (68%)
Grade 6	58% (56%)	71% (70%)	69% (69%)	61% (59%)	63% (63%)	69% (69%)
Grade 7	56% (55%)	68% (67%)	69% (69%)	57% (56%)	67% (67%)	68% (68%)
Grade 8	57% (56%)	56% (54%)	63% (63%)	62% (62%)	62% (61%)	71% (71%)

Table 4.11 summarizes the amount of variance explained by adding the student-level covariates into the models. Compared with the explained variance from the models in the first research question (see Table 4.8), models at this stage explained slightly greater amounts of variance in student achievement.

Table 4.12 Summary of Spearman's Correlation Coefficients (standard errors) between VAM (based on Models with Student Prior Achievement and Student-level Covariates Adjusted for) and SLO Quality by Year and Grade for Mathematics and Reading

	Mathematics			**Reading**		
	2008	2009	2010	2008	2009	2010
Grade 4	0.18(.25)	0.09(.22)	-0.02(.24)	0.08(.26)	0.25(.21)	-0.37(.23)
Grade 5	-0.30(.24)	0.43(.21)	0.19(.23)	-0.17(.25)	0.43(.22)	0.33(.22)
Grade 6	-0.24(.25)	-0.14(.26)	0.31(.36)	0.03(.22)	-0.30(.22)	0.03(.32)
Grade 7	0.46(.27)	-0.15(.26)	0.29(.29)	0.48(.25)	0.22(.24)	0.06(.27)
Grade 8	-0.43(.30)	0.21(.27)	0.23(.28)	0.00(.38)	0.51(.24)	0.00(.30)

New sets of teachers' VAM estimates were obtained from the final models at this stage. Table 4.12 displays the correlations between teachers' SLOs quality scores and their new VAM estimates after controlling for students' prior achievement and all the student-level covariates. Similar to the correlational results from the models in research question one (see Table 4.9), for both subjects correlation coefficients varied across all years and grades. Teachers' SLOs quality scores were positively correlated with VAM in most grades and years, while a few negative coefficients were found as well. Similar to the results from research question one, no noticeable patterns of the correlation

coefficients distributions were found among different years and grades. Due to the limited availability of level-2 records, it is not unexpected that no significant coefficients were found at this stage.

4.4 Research Question Three

In order to investigate the associations between teacher-level variables and students' achievement scores after adjusting for their prior achievement scores and student-level characteristics, teacher-level characteristics were added to the final multilevel models from research question two. In this study, due to the limited availability of level-2 records (see section 3.3.2), class size was the only level-2 variable. Since the regression coefficients at level-2 are related to the focus of this research question, the level-2 variables were grand-mean centered. The continuous variables at level-1 (e.g. student prior achievement scores) were also grand-mean centered, and the categorical variables remained uncentered. After teacher-level variables were entered into the model, the covariates no longer significant were removed. As class size was the only level-2 covariate, it was remained in the model even when no statistical significance was found.

Table 4.13 summarizes the results of the final models. Class size was found to be significant in some years and grades. It is interesting to note that class size had mixed associations with student achievement; for example, on average, in the model for mathematics-2008-Grade-7, students in larger classes scored significantly higher than those from smaller classes (0.18, $t(79)=3.30$, $p<.001$) given the same student prior

academic achievement and other student background characteristics. On the other hand, in the model for mathematics-2010-Grade-8, students in larger classes scored significantly lower than those with smaller class size (-0.30, $t(30)=-3.78$, $p<.001$), though the difference in scores was not substantively meaningful.

Table 4.13 Results of HLMs for Research Question Three: Estimated Regression Coefficients for Class-size

	Mathematics			**Reading**		
	Coef.	s.e.	*p*-value	Coef.	s.e.	*p*-value
2008-Grade-4	-0.02	0.09	0.81	-0.03	0.06	0.62
2008-Grade-5	-0.03	0.05	0.57	0.00	0.04	0.95
2008-Grade-6	-0.03	0.10	0.78	0.10	0.09	0.27
2008-Grade-7	0.18	0.05	**<0.001**	0.07	0.04	0.08
2008-Grade-8	0.06	0.06	0.34	-0.01	0.03	0.80
2009-Grade-4	-0.16	0.07	**0.03**	0.01	0.07	0.89
2009-Grade-5	0.05	0.05	0.32	0.14	0.05	**<0.01**
2009-Grade-6	-0.15	0.10	0.15	-0.07	0.09	0.49
2009-Grade-7	0.11	0.08	0.22	0.05	0.06	0.47
2009-Grade-8	0.01	0.04	0.89	0.04	0.02	0.12
2010-Grade-4	-0.10	0.07	0.15	-0.02	0.06	0.74
2010-Grade-5	-0.09	0.09	0.35	0.07	0.07	0.30
2010-Grade-6	0.24	0.11	**0.03**	0.05	0.07	0.41
2010-Grade-7	-0.13	0.12	0.30	-0.01	0.04	0.79
2010-Grade-8	-0.30	0.08	**<0.001**	-0.04	0.02	**<0.05**

With the teacher-level covariate added to the model, an example of the final models (mathematics-2008-Grade 4) developed at this stage can be expressed as below. The complete model specifications for all years, grades and subjects are contained in Appendix B.

Level-1 Model (Student level) (4.3)

$$\begin{aligned} Math\_2008\_G4\_Score_{ij} &= \beta_{0j} + \beta_{1j}*(Math\_2007\_G3\_Score_{ij}) \\ &+ \beta_{2j}*(AmericanIndian_{ij}) + \beta_{3j}*(Asian_{ij}) + \beta_{4j}*(Hispanic_{ij}) \\ &+ \beta_{5j}*(Multi\text{-}race_{ij}) + \beta_{6j}*(White_{ij}) + \beta_{7j}*(GIFTED_{ij}) \\ &+ \beta_{8j}*(LEP_{ij}) + \beta_{9j}*(SWD_{ij}) + r_{ij} \end{aligned}$$

Level-2 Model (Teacher level)

$$\beta_{0j} = \gamma_{00} + \gamma_{01}*(Class\_size_j) + u_{0j}$$

$$\beta_{1j} = \gamma_{10}$$

$$\beta_{2j} = \gamma_{20}$$

$$\beta_{3j} = \gamma_{30}$$

$$\beta_{4j} = \gamma_{40}$$

$$\beta_{5j} = \gamma_{50}$$

$$\beta_{6j} = \gamma_{60}$$

$$\beta_{7j} = \gamma_{70}$$

$$\beta_{8j} = \gamma_{80}$$

$$\beta_{9j} = \gamma_{90}$$

Table 4.14 Amount of Variance Explained by Final Models

	Mathematics			Reading		
	2008	2009	2010	2008	2009	2010
Grade 4	59%	64%	66%	62%	65%	65%
Grade 5	66%	69%	73%	59%	65%	69%
Grade 6	58%	71%	69%	61%	62%	69%
Grade 7	58%	68%	69%	60%	67%	68%
Grade 8	57%	56%	64%	64%	64%	71%

Table 4.15 Example of Variance Decompositions in Different Models (Mathematics-2008-Grade-4)

			Within teachers		Between teachers	
Model	Level 1 covariates	Level 2 covariates	Variance	Percent of variance in model 1 accounted for	Variance	Percent of variance in model 1 accounted for
1	None	None	56.99	-	21.94	
2	Student prior achievement	None	25.44	55%	17.00	27%
3	Student prior achievement and other covariates	None	24.80	56%	14.77	33%
4	Student prior achievement and other covariates	Teacher-level covariate	24.80	56%	8.40	59%

With the amount of variance explained by adding teacher-level covariate into the models displayed in Table 4.14, Table 4.15 summarizes the variance decompositions in the final models at each stage and the percent of variance in the baseline model accounted for both within and between teachers. Taking mathematics-2008-Grade 4 as an example, Model 1 is the unconditional model or the baseline model. Apparently, most of the variance is among students within teachers, with 28% of total variance between teachers. In model 2, because of the strong association between students' achievement in grade 4 and in grade 3, adding student prior achievement into the model accounted for 55% of the variance among students within teachers and 23% of the variance among teachers. In other words, 46% of the total variance in student mathematics achievement at 2008 Grade 4 (i.e. $\frac{TotalVariance-ResidualVariance}{TotalVariance} = \frac{(56.99+21.94)-(25.44+17.00)}{(56.99+21.94)} = 46\%$) was accounted for by their prior achievement at Grade 3 in 2007. In model 3, student-level covariates were added into the model and this model explained 56% of the total variance within teachers and 33% of the total variance between teachers. In comparison to model 2, an extra 1% (i.e. 56% - 55%) of the total variance within teachers and 6% (i.e. 33% - 27%) of the total variance among teachers were accounted for. In model 4, a teacher-level covariate was included in the model and this model explained 56% of the total variance within teachers and 59% of the total variance among teachers. That is 58% of the total variance was accounted for in the final model.

Table 4.16 Summary of Spearman's Correlation Coefficients (standard error) between VAM (based on Models with Student Prior Achievement, Student and Teacher Level Covariates Adjusted for) and SLO Quality by Year and Grade for Mathematics and Reading

	Mathematics			Reading		
	2008	2009	2010	2008	2009	2010
Grade 4	0.06(.26)	0.25(.21)	0.20(.24)	0.22(.25)	0.51*(.19)	-0.14(.24)
Grade 5	-0.05(.25)	0.28(.22)	-0.04(.23)	-0.14(.25)	0.35(.23)	0.12(.23)
Grade 6	-0.06(.26)	0.08(.27)	0.73*(.26)	0.16(.22)	-0.01(.24)	0.31(.30)
Grade 7	0.35(.28)	0.01(.27)	0.33(.28)	0.27(.28)	0.02(.25)	-0.06(.27)
Grade 8	-0.33(.31)	0.18(.27)	0.10(.29)	0.08(.38)	-0.01(.28)	0.21(.30)

* Correlation is significant at 0.05 level (2-tailed).

With the full model completely established, another set of teachers' VAM estimates were obtained for different years and grades. Table 4.16 displays the correlation coefficients between teachers' SLO quality scores and their new VAM estimates after controlling for student prior achievement, student and teacher level covariates. Similar to the correlation analysis results from research questions two, for both subjects, positive correlations were found in most grades and years while there were a few negative ones in some years and grades. In particular, the correlation coefficient for the model of mathematics-2010-Grade 6 was statistically significant (0.73, $p<.05$). Similarly, for the model of reading-2009-Grade 4, teachers' VAM estimates were also significantly correlated with their SLOs quality scores (0.51, $p<.05$). It is worth noting that as multiple analyses for different subject/year/grade combinations have been conducted, there could

be the problem of multiplicity (Benjamini & Braun, 2002) and the significant findings may be simply due to chance.

The correlations between the VAM rankings and SLO quality were further analyzed across models. The correlation results from Model 1 adjusting for student prior achievement were compared and contrasted against those from Model 2 while both prior achievement and student level covariates were included. Further, the correlations from Model 3 with teacher level covariate incorporated in addition to student prior achievement and student level covariates were added to the cross-model analysis. Figure 4.3-4.4 illustrate the comparisons of the correlations from the groupings based on different years and grade levels for mathematics and reading, respectively. Correlations from Model 1 were considered the baseline for the comparisons in the figure.

Figure 4. 3 Comparisons of the Correlations from Models 1 through 3 - Mathematics

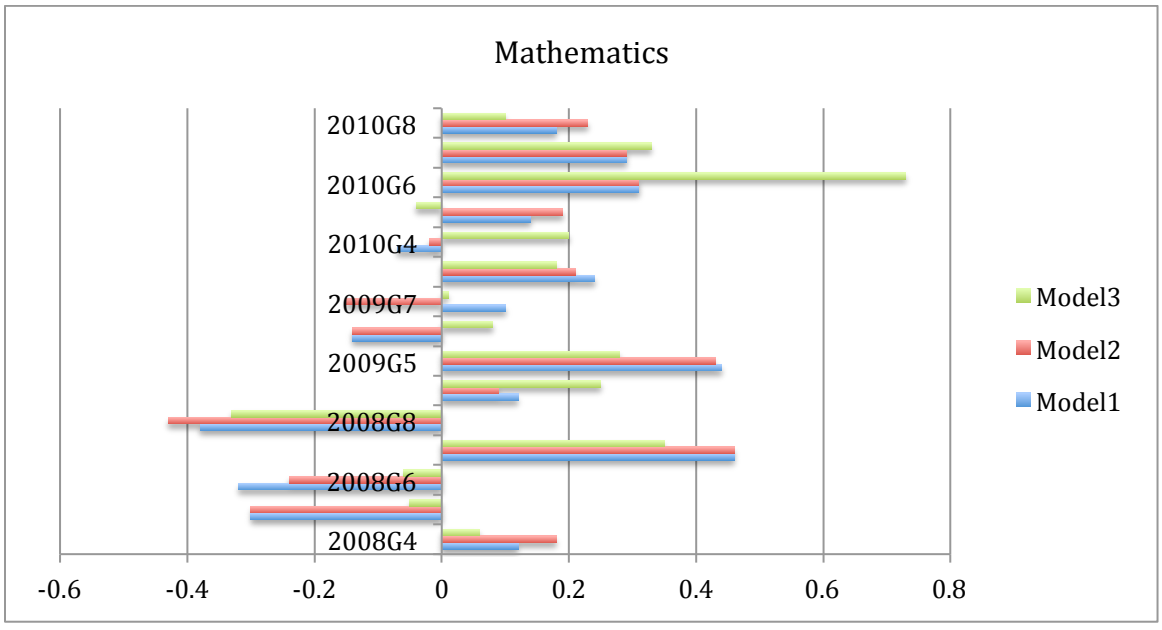
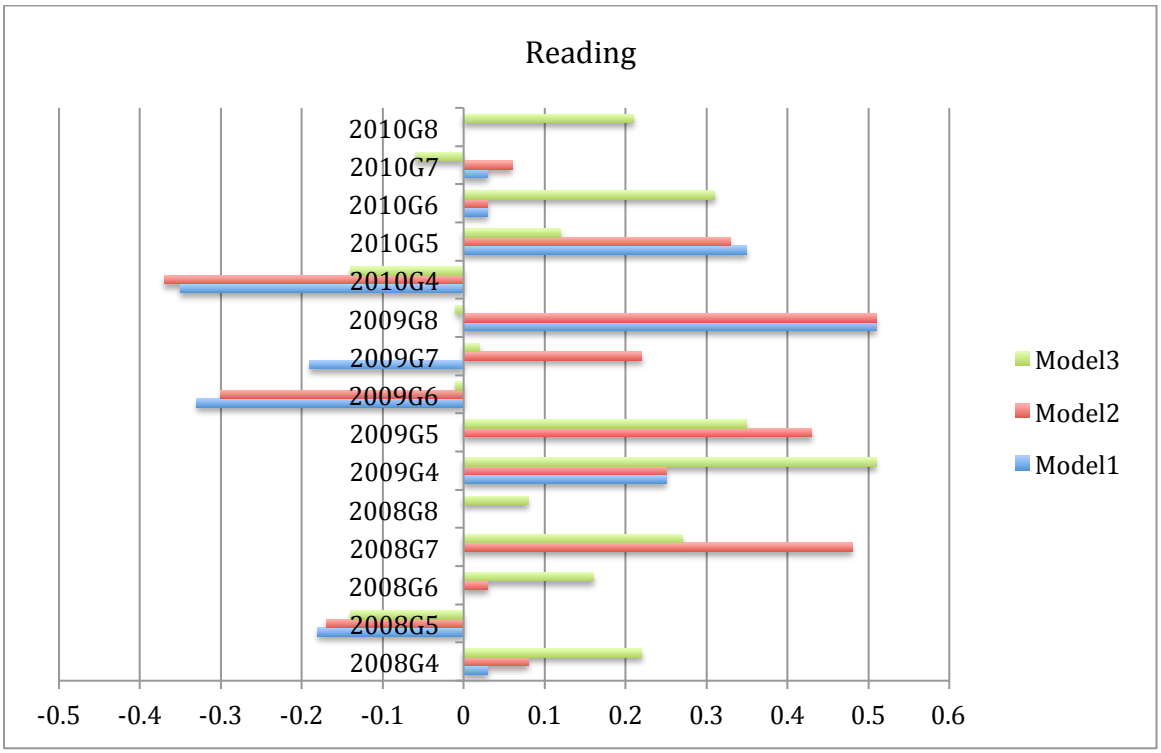


Figure 4. 4 Comparisons of the Correlations from Models 1 through 3 - Reading



In general, for both mathematics and reading, correlations in Model 2 were higher than those in Model 1 while the correlations in a few year/grade combinations were lower. Most correlations in Model 3 were also higher than in Model 1 as well. There were several groups that had exceptional high correlations (r>0.5) such as mathematics-2010-Grade 6, reading-2009-Grade 4, and reading-2009-Grade 8. For mathematics, the correlations from Model 3 seemed to be higher in 2010 while the correlations from Model 2 appeared higher in 2009. As for reading, the differences in the correlations between Model 2 and Model 3 did not yield a clear pattern. It is worth noting that most estimates were not statistically significant, and therefore these are likely noise-to-noise comparisons.

The between-model agreement of the correlations was further examined among the three models that estimated teachers' VAM with different levels of covariates adjusted for. The rankings of the correlations between the VAM estimates from different models and SLO quality were examined for consistency among models. The correlations from each of the three models were first categorized into four quartile groups. An index of consistency on the correlation rankings was then calculated using the percent of correlations that were consistently categorized into each quartile group. These consistency indices are presented in Table 4.17.

Table 4.17 Comparisons of Quartile Group Rankings in Three Models Estimating the Correlations between VAM and SLO Quality

		Quartile			
		1st	2nd	3rd	4th
Mathematics	Model 1 & Model 2	75%	75%	100%	100%
	Model 2 & Model 3	75%	50%	75%	67%
Reading	Model 1 & Model 2	75%	50%	50%	100%
	Model 2 & Model 3	75%	25%	25%	33%

Similar to the findings based on the correlation tables from the first three research questions, for mathematics, Model 1 and Model 2 showed the highest consistency in the third and fourth quartiles with all correlations completely matching in the top half of the correlation rankings. This implies that adding student level covariates to the models did not change the top half correlation rankings, while there was a 25% change in the lower half of the rankings. With the teacher level covariate included in Model 3 in addition to student-level covariates, Model 3 and Model 2 had the least agreement in the second

quartile followed by the fourth quartile. As for reading, Model 1 and Model 2 demonstrated higher agreement than Model 2 and Model 3, which implies that adding student-level covariates did not change the rankings as much as including teacher level covariate. Except for the first quartile, Model 2 and Model 3 had lower agreement across the rest of the distribution. In general the overall consistency across models is higher for mathematics than for reading.

4.5 Research Question Four

To investigate the extent to which the correlations between teachers' VAM rankings and their SLOs quality vary by year, grade, and type of school, a Weighted Least Square (WLS) regression was employed in order to adjust for the variation in the numbers of level-2 cases across years and grades (see section 3.1).

Figure 4.5 Correlation Distributions by Grade and Year for Mathematics

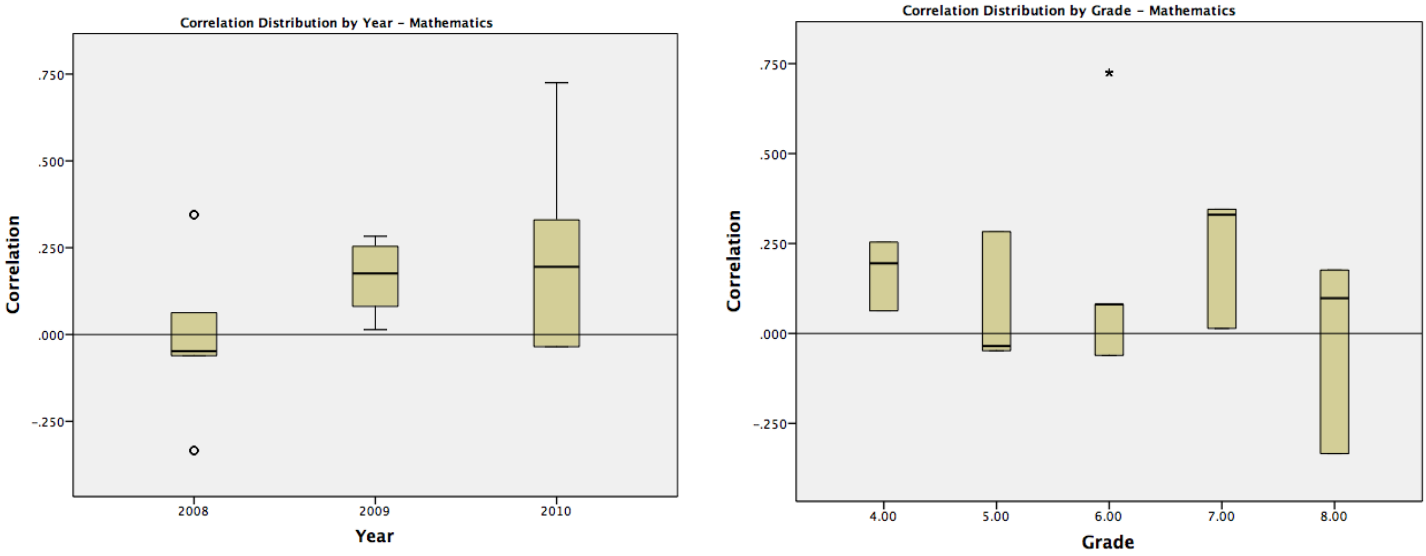
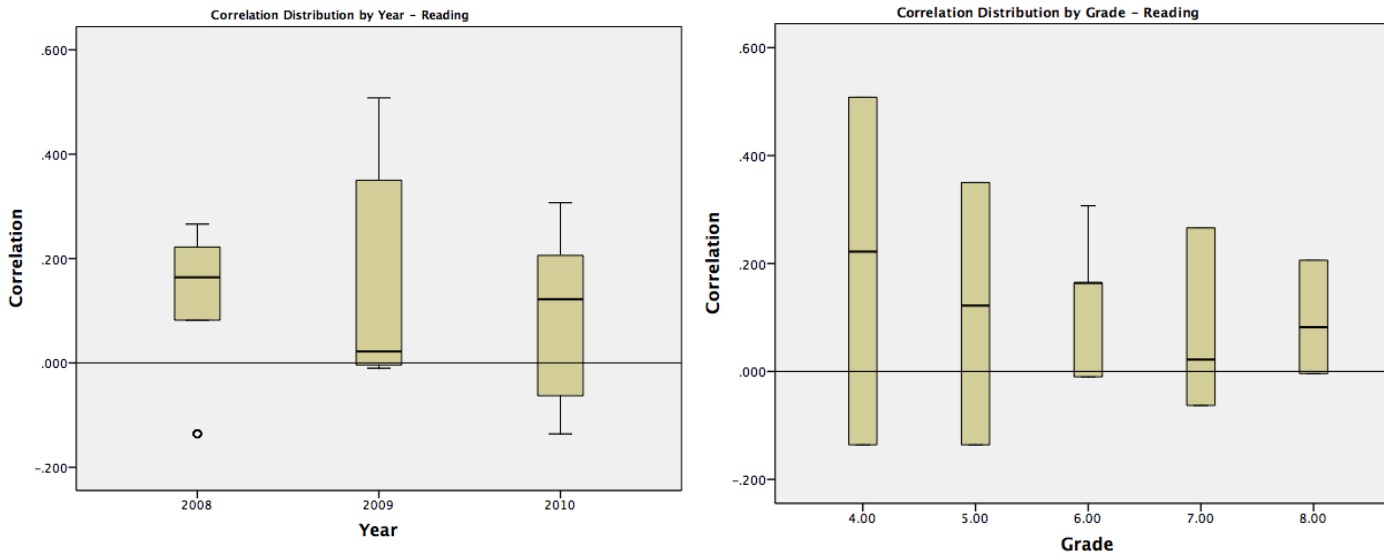


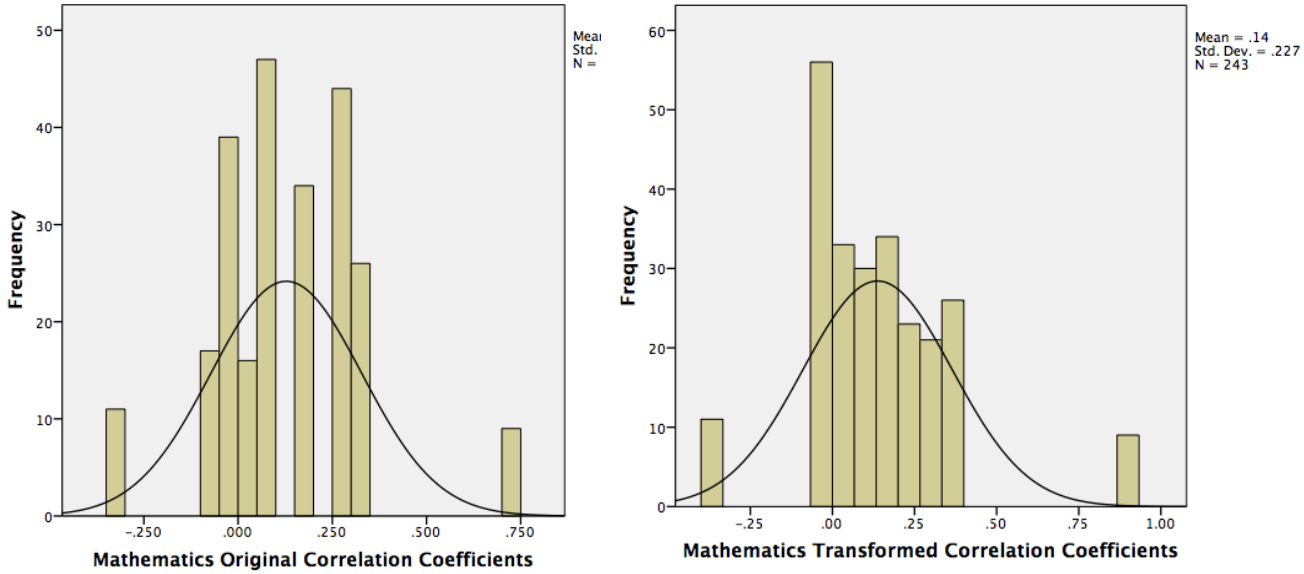
Figure 4.6 Correlation Distributions by Grade and Year for Reading



The distributions of the correlation coefficients by year and grade are displayed using box-and-whisker diagram in Figure 4.5 and 4.6 for mathematics and reading, respectively. Unsurprisingly, the distributions varied considerably across grades and years. Most correlation coefficients, for both mathematics and reading, were positive in all three years. With regard to mathematics, the distribution of 2010 showed a wider range than that of the other two years. In addition, the correlation coefficients distribution displayed greatest range in the eighth grade and the modest range in the sixth grade. For reading, the correlations in 2009 showed larger value and range than the other years. In addition, the distribution in lower grades indicated greater range than higher grades.

In order to use the correlation coefficients as the dependent variables in the WLS analysis, the Fisher-Z transformation[19] was employed to stabilize the variance of the correlations. Figure 4.7 and 4.8 display the comparison of the original and transformed correlation coefficients for the final models with student prior achievement, student-level and teacher-level covariates accounted for in both subjects. As expected, the transformed coefficients displayed greater conformity to normality than the original correlations.

Figure 4.7 Comparisons of Correlation Coefficients and Their Fisher-Z Transformations for Mathematics



[19] The behavior of Fisher-Z transformation has been extensively studied and found to provide normal distribution by many researchers. Though the sample sizes across years and grade levels in this study are similar, the transformation can still help because the variance also depends on the value of the correlation.

Figure 4.8 Comparisons of Correlation Coefficients and Their Fisher-Z Transformations for Reading

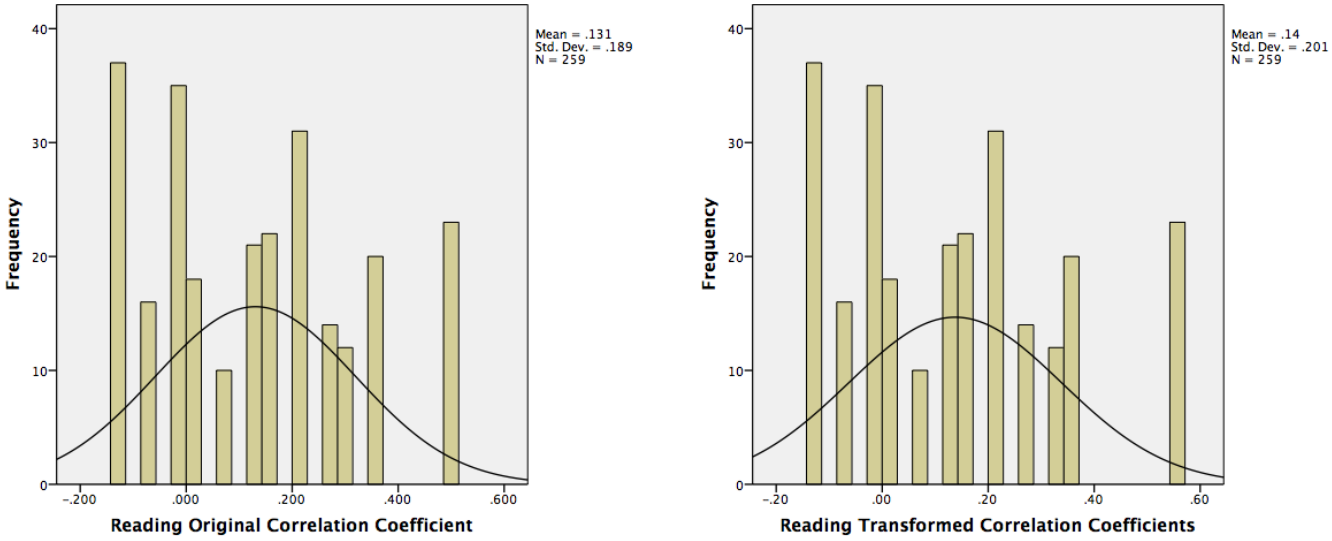


Table 4.18 Overall WLS Results for Research Question Four

	Mathematics		Reading	
Variables	*F*	*p*-value	*F*	*p*-value
Year	27.81	**< .001**	22.08	**< .001**
Grade	7.20	**< .001**	9.55	**< .001**
Type of School	.69	.41	18.93	**< .001**

Table 4.18 summarizes the results of the weighted least square regression. The transformed Spearman's correlations between mathematics teachers' VAM estimates and SLO quality varied significantly by year (F=27.81, p<.001) and grade (F=7.20, p<.001). With respect to reading, the correlations between teachers' VAM estimates and SLO quality varied significantly across years (F=22.08, p<.001) and grades (F=9.55, p<.001) as well.

It is worth noting that the (transformed) correlation coefficients between the SLO quality score and VAM estimates obtained from the models of grade and year combinations are used as the dependent variable in this regression analysis. Since most correlations coefficients are not statistically significant, this is an analysis that, by combining results across years and grade levels, hopes to uncover potentially meaningful patterns, despite high levels of noise.

Table 4.19 displays the results of the post-hoc comparisons of the transformed correlation coefficients among different years and grade levels for both mathematics and reading. For mathematics, the transformed correlation coefficients were highest in 2009 followed by 2010 and 2008. The correlations between teachers' VAM and SLO quality in 2009 were statistically significantly higher than that in 2008 (d=.19, $s.e.$=.03, p<.001)[20], and similarly the correlations in 2010 were significantly higher than that in 2008 (d=.15, $s.e.$=.03, p<.001). In other words, teachers' VAM estimates and SLO quality had a significantly stronger correlation in 2009 than in 2008. Similarly, the two indicators were better correlated in 2010 than in 2008. The difference between 2009 and 2010 was not significant. With regard to reading, the correlations in general were higher in 2009 than in 2010 and 2008. The difference between 2009 and 2010 was significant (d=.19, $s.e.$=.03, p<.001), and so was the difference between 2009 and 2008 (d=.12, $s.e.$=.03, p<.001). No significant difference was found between 2008 and 2010.

[20] d denotes the mean difference from the comparison.

Table 4.19 Comparisons of Transformed Correlation Coefficients among Different Years Holding Grade Constant

	Mathematics			Reading		
Year	Mean Difference	Std. Error[21]	*p*-value	Mean Difference	Std. Error	*p*-value
2008 vs 2009	-.19	.03	< .001	-.12	.03	< .001
2009 vs 2010	.04	.03	< .001	.19	.03	< .001
2008 vs 2010	-.15	.03	< .001	.07	.03	.11

Table 4.20 Comparisons of Transformed Correlation Coefficients among Different Grade Levels Holding Year Constant

	Mathematics			Reading		
Grade	Mean Difference	Std. Error	*p*-value	Mean Difference	Std. Error	*p*-value
Grade 4 vs 5	.11	.03	<.01	.14	.03	<.01
Grade 4 vs 6	.09	.04	.19	.15	.04	<.01
Grade 4 vs 7	-.01	.04	1.00	.21	.04	<.001
Grade 4 vs 8	.15	.04	<.01	.18	.05	<.01
Grade 5 vs 6	-.02	.04	.98	.02	.04	.99
Grade 5 vs 7	-.12	.04	<.05	.08	.04	.46
Grade 5 vs 8	.04	.04	.88	.04	.05	.94
Grade 6 vs 7	-.10	.04	.27	.06	.04	.69
Grade 6 vs 8	.06	.04	.71	.03	.05	.99
Grade 7 vs 8	.16	.04	<.05	-.03	.05	.98

[21] Standard error estimates were calculated using the equation $s_e = \sqrt{\frac{SSE}{n-2}}$.

The transformed correlation coefficients were further compared and contrasted between different grade levels, and the results are summarized in Table 4.20. Regarding the variations among grade levels for mathematics teachers, the correlations across years were greater in the 4th and 7th grades, and lower in the 5th and 8th grades. There was a statistically significant difference between 4th grade and 5th grade (d=.11, $s.e.$=.03, p<.01), as well as between 4th and 8th grade (d=.15, $s.e.$=.04, p<.01). In addition, the correlations in 7th grade were greater than those in the 5th grade (d=.12, $s.e.$=.04, p<.01), and in the 8th grade (d=.16, $s.e.$=.04, p<.05). As far as the variations among grade levels were investigated for reading, the correlations in the 4th grade were found to be the highest, and they were statistically significantly higher than in any other grades (against the 5th grade: d=.14, $s.e.$=.03, p<.01; against the 6th grade: d=.15, $s.e.$=.04, p<.01; against the 7th grade: d=.21, $s.e.$=.04, p<.001; against the 8th grade: d=.18, $s.e.$=.05, p<.01). No significant difference was found among grade levels 5 through 8.

Table 4.21 Comparisons of Transformed Correlation Coefficients between School Types Holding Year and Grade Constant

	Mathematics			**Reading**		
Type of School	Mean Difference	Std. Error	p-value	Mean Difference	Std. Error	p-value
Elementary vs Middle	-.02	.03	.408	.11	.03	< .001

Further, the relationship between teachers' VAM estimates and SLO quality was examined by the type of school (elementary schools versus middle schools), and the comparisons of the transformed correlation coefficients are summarized in Table 4.21.

The correlations in elementary schools were compared with those in middle schools across the years. Results indicated that statistically significant difference was only found among reading teachers (d=.11, $s.e.$=.03, p<.001). In other words, teachers' VAM and SLO quality were more highly correlated in elementary schools than in middle schools for reading.

4.6 Research Question Five

To investigate the associations between the SLOs quality scores and the corresponding achievement status, a logistic regression model was used for mathematics and reading, respectively, in which whether the SLOs were attained served as a dichotomous dependent variable and the SLO quality scores were used as a continuous predictor.

With regard to the findings from the logistic regression analysis, no significant results were found from the overall analysis of the relationship between SLOs quality and their attainment status for either mathematics or reading. In other words, the SLOs quality scores in this study are not a good predictor of whether the SLOs were achieved.

The logistic regression analysis was also conducted for subgroups based on different grade/year combinations. Significant results were only found in the subgroup of the 7^{th} grade mathematics teacher in 2009 (B=3.00, $s.e.$=1.35, p<.05). The model results are summarized in Table 4.22, which suggests that for the Grade-7 mathematics in 2009,

with one unit increase in teachers' SLO quality score, the log-odds[22] of whether these SLOs were achieved is estimated to increase by 3 units. In other words, the odds of the teachers, who had higher SLOs quality scores, to achieve their SLOs, were 20 times higher than those of teachers with lower SLOs quality scores. It is worth noting that as multiple analyses for different subgroups have been conducted, there is a problem of multiplicity and this significant finding may be simply due to chance.

Table 4.22 Logistic Regression Results for Research Question Five

Mathematics - Grade 7 - 2009			
Coef.	s.e.	*p*-value	Exp (B)
3.00	1.35	**.027**	20.000

4.7 Research Question Six

To investigate the associations between the achievement status of the SLOs and the teachers' VAM estimates, a point-biserial correlational analysis was employed in which the achievement status was treated as a dichotomous variable and the VAM estimates as a continuous variable.

The overall point-biserial analysis results showed that the correlations between VAM estimates and SLO attainment status were only .04 for both mathematics and

[22] Log-odds unit is the original coefficient from the logistic regression analysis outcome. The exponentiation of this coefficient, or the odds ratio, is often used for easier interpretation. In this analysis (output shown in Table 4.22), the log-odds coefficient is 3 and the odds ratio is 20.

reading. Due to the limited number of records in the dataset, no significant results were found. When examined by grade/year combination, the correlations between the fifth grade mathematics teachers' VAM estimates and their SLO achievement status were found to be statistically significant (r=.27, p<.05) across years. This suggested that if these teachers' effectiveness in helping students with their academic achievement were high, their SLOs were likely to be achieved as well. In other words, if teachers' SLOs were not achieved, their ability in improving student achievement was probably low as well.

CHAPTER 5. CONCLUSIONS

5.1. Summary of Findings

Over the last decade, schools and teachers increasingly have been held accountable for student learning outcomes. The differential effectiveness of teachers in improving student progress as measured by test performance has been demonstrated by many research studies. Clearly, the substantial differences in teacher effectiveness have meaningful consequences for student performance and growth. Therefore, it is of critical importance to identify both effective and ineffective teachers through the development and implementation of a teacher evaluation system. As the methods and practices of teacher evaluation currently employed in public school districts nationwide are often based on simplistic criteria with marginal relevance to what teachers need to do to enhance student learning (Danielson & McGreal, 2000), discussions regarding more relevant indicators of teacher quality are ongoing. With the increasing availability of longitudinal student achievement data, more quantitative strategies have been explored to exploit the heterogeneity in students' test score trajectories in order to measure a key aspect of teacher effectiveness. In this context, the present study extends current efforts to examine different approaches to measuring teacher effectiveness and explores the relationship of two indicators of teacher effectiveness.

In particular, this study examined the value-added estimates based on patterns in student test performance, and the quality of the student learning objectives that were developed by teachers. Based on students' end-of-grade academic achievement in mathematics and reading across grades and years, different hierarchical linear models

were fit in order to estimate the associations between student achievement and their background characteristics, after adjusting for the prior test scores in previous years. Teachers' value-added estimates were calculated based on the models with different sets of factors accounted for. While the VAM scores are normative, the SLO quality scores are considered criterion-referenced. The relationship between teachers' value-added estimates and their SLO quality was examined at each stage of the developed models, and further compared and contrasted across grades and years, as well as against SLO attainment status. The following section summarizes findings from the six main research questions.

5.1.1 Relationship between SLO Quality and VAM Estimates based on Models with Student Prior Achievement Adjusted for

In the analysis for the first research question, students' achievement scores from prior years were taken into account. As expected, all estimated regression coefficients of student prior scores were found to be highly significant and, for both mathematics and reading, they explained approximately 63% of the variance in current student achievement, averaged across grades and years.

The relationship between SLO quality and VAM estimates based on HLM models with student prior achievement adjusted for varied across years and grades in both mathematics and reading. 67% of the correlation coefficients across years and grades in mathematics were positive, and 33% were negative. Similarly, 73% of the correlations

coefficients across years and grades in reading were positive, and 27% were negative. No other noticeable patterns for the correlation distribution were found.

Among all the correlation coefficients for difference years, grades, and subjects, only the one for mathematics-2009-grade 5 was found to be statistically significant. This means that the VAM estimates of the 5th-grade mathematics teachers in 2009, after controlling for their students' prior achievements, were significantly correlated with their SLOs quality scores. The positive coefficient means that the higher the teachers were ranked based on their students' achievement growth, the greater their SLOs quality scores. Given that the correlation analyses were conducted for a range of teacher groups based on the subject, year, and grades they taught, there is a problem of multiplicity and the only significant finding is likely due to chance.

5.1.2 Influence of Student-level Covariates on the Relationship between SLO Quality and VAM Estimates

In the second stage of the HLM analysis, student-level covariates were added by group into the models. Different student-level variables were identified as being associated with students' achievement in different models based on subjects, years and grades, even after adjusting for their prior achievement scores. Taking the model of mathematics-2008-Grade 4 as an example, students' race/ethnicities, gifted status, LEP status and SWD status were all found to be significantly associated with the outcome variable. In particular, Hispanic students were estimated to score higher in both 4th grade mathematics and reading EOG tests than their African American counterparts, even after adjusting for their prior achievement in the 3rd grade in 2007 and other student

background characteristics. Similarly, partial regression coefficients indicated that gifted students were estimated to score higher than non-gifted students, LEP students were estimated to score lower than non-LEP students, and SWD students were expected to score lower than non-SWD students. In general, indicators of race/ethnicity were found to be significant predictors in the models of most years and grades for mathematics and reading. Gender, gifted, LEP and SWD were significant in various models respectively depending on the subject, year and grade combinations.

Since different sets of VAM estimates were obtained from the final models at this stage of HLM analysis with all student prior achievement and student-level factors adjusted for, new sets of correlation coefficients were calculated with teachers' SLO quality scores. Similar to the correlation analysis results for research question one, for both mathematics and reading, the correlation coefficients at this stage were found to be positive in most grades and years, while a few negative coefficients were found as well. Sixty percent of the correlation coefficients across years and grades in mathematics were positive, and forty percent were negative. Eighty percent of the correlation coefficients across years and grades in reading were positive, and twenty percent were negative. No statistically significant coefficients were identified at this stage, and no noticeable patterns of the correlation coefficient distribution were found among different years and grades for both mathematics and reading.

In comparing the findings with regard to the correlation results from the research questions one (with student prior scores only in the models) and two (both student prior scores and student-level covariates in the models), it is worth noting that the correlation

coefficients from the second research question did not demonstrate substantial changes from those of the first research question. Most correlation results across years and grades were similar to the results for the first research question or, in other words, adding student-level covariates into the models after controlling for student prior achievement did not substantially change the statistical relationships between teachers' SLO quality and VAM estimates.

5.1.3 Influence of a Teacher-level Covariate on the Relationship between SLO Quality and VAM Estimates

Due to the limited availability of level-2 records, class size was adopted as the only teacher-level covariate and was kept in the models even when no statistical significance was found. It was interesting to note the relationship of class size to student achievement from the analysis results at this stage. For example, in the model of Mathematics-2008-Grade 7, students from larger classes on average scored statistically significantly higher than those from smaller classes, given the same student prior achievement scores and other student background characteristics. However, the opposite pattern was found in the model of Mathematics-2010-Grade 8, where students with larger class sizes were estimated to score statistically significantly lower than those with smaller class sizes.

With both student- and teacher-level covariates included in the models at this stage, new sets of teachers' VAM estimates were calculated. Based on the new VAM results, another set of correlation coefficients between teachers' SLO quality scores and their VAM estimates were obtained. Similar to previous results, both positive and

negative coefficients were found depending on different years and grades for both subjects: seventy-three percent of the correlation coefficients across years and grades in mathematics were positive, and twenty-seven percent were negative. Likewise, sixty-seven percent of the correlation coefficients across years and grades in reading were positive, and thirty-three percent were negative.

In particular, the correlation coefficients from Mathematics-2010-Grade 6 and Reading-2009-Grade 4 models were found to be significant. The higher the SLO quality of the teachers in these two subject/year/grade combinations, the higher their VAM rankings were estimated to be even after controlling for their students' prior achievement, student background characteristics, and class size. Compared with the final model results in research question two, when the correlation coefficients in these two models were not significant, adding teacher-level variable brought about some changes to these coefficients. In other words, after controlling for class size in these two models in addition to student prior achievement and background characteristics, the relationship between teachers' SLO quality and VAM estimates became statistically significant. Again it is worth to note that there is the problem of multiplicity due to the multiple groups of analyses and this significant finding is likely due to chance.

5.1.4 Variation of the Relationship between SLO Quality and VAM Estimates by Year, Grade, and Type of School

Significant heterogeneity was found in the relationships between SLO quality and VAM estimates across years, grade levels, and the type of schools. With respect to year, the relationship between SLO quality and VAM estimates varied significantly for both mathematics and reading. Teachers' VAM estimates were more strongly correlated with their SLO quality in 2009 than in other years for both mathematics and reading. The differences in the transformed correlation coefficients between 2009 and 2008 was found to be statistically significant for both subjects, while the difference between 2009 and 2010 was found significant only for reading and the difference between 2008 and 2010 was found significant only for mathematics.

With respect to heterogeneity across grade levels, the relationship was stronger for the 4^{th} grade teachers of both subjects. The correlations in other grades were all statistically significantly lower than those in the 4^{th} grade for reading, and for mathematics the correlation coefficients in the 5^{th} and 8^{th} grades were significantly lower than those in the 4^{th} grade as well; however, the 6^{th} and 7^{th} grade teachers had correlations similar to those of the 4^{th} grade. Furthermore, the relationship between teachers' VAM estimates and SLO quality in elementary schools was stronger than that in middle schools across years.

5.1.5 Associations between SLO Quality and SLO Attainment Status

Logistic regression models were employed to analyze the associations between SLOs quality and whether the SLOs were achieved after a learning period. Results indicated that at the overall level, with all years and grades included in the analysis, no significant relationships were found. In other words, the SLOs quality score did not appear to be a good predictor of whether the SLOs were attained for either mathematics or reading.

When the logistic regression analysis was conducted at the subgroup level, significant results were found for the 7th grade mathematics teachers in 2009, which suggested that the odds of the teachers, who had higher SLOs quality scores, to achieve their SLOs, were 20 times higher than those of teachers with lower SLOs quality scores.

5.1.6 Associations between VAM estimates and SLO attainment status

Point-biserial analyses were employed to further investigate the associations between the achievement status of the SLOs and the teachers' VAM estimates. Similar to the previous research question, the results from the overall analysis, with all grades and years included, showed no statistically significant findings of the relationship between VAM estimates and SLO attainment status. By subgroup, the correlation between the fifth grade mathematics teachers' VAM estimates and their SLO achievement status was found to be statistically significant across years. This suggested the expected result - that if these teachers' effectiveness in improving students' academic achievement growth was high, their SLOs were likely to be estimated as achieved. Conversely, if teachers' SLOs

failed to be attained, their ability in helping students with achievement growth would probably be predicted to be low. However, given the number of analyses conducted at this step, the results will also face the challenge of multiplicity. In other words, the only statistically significant finding from one group of teachers at this stage of analysis could be a chance result. As both VAM estimates and SLO achievement status intend to measure teacher's aptitude in fostering student's academic growth over time, the weak relationship between VAM scores and SLO attainment status from this analysis might be surprising.

5.2 Policy Implications

The quality of teaching is an important determinant of student learning and progress. Therefore, a successful teacher evaluation system is urgently needed. With the increasing availability of longitudinal student achievement data, researchers and policy makers have started to explore more objective approaches to quantifying the heterogeneity in students' test score trajectories and to use patterns in student achievement outcomes as the basis for indicators of teaching effectiveness.

Value-added models have been intensively studied and widely employed in many states and districts to examine the effectiveness of teachers in facilitating students' academic progress. However, as emphasized by the NRC report (2011) and the Measures of Effective Teaching (MET) study (Bill & Melinda Gates Foundation, 2012), validation of value-added estimates of teacher effectiveness remains an important area of research. This study, investigated how value-added estimates relate to other indicators of teacher effectiveness and could produce one type of validity evidence[23] for using the value-added approach, as well as SLOs approach, to measuring teacher effectiveness and, thus, contributes to the current research on teacher evaluation.

Regarding the methodology of value-added modeling, it is worth noting the impact of different specifications of the models. The same VAM models with variations in accounting for factors at different levels, or the same models using data from different years and grades can produce substantially different results. This study produced teacher rankings at different stages of HLM models for data from different subjects, years, and

[23] Strictly speaking this is not a validity study of VAM or SLO quality scores as we are not able to suggest whether VAM or SLO should be used.

grade levels. Covariates at student and teacher levels were added sequentially to the models, which, to different extent, impacted the value-added estimates of teachers in each subgroup based on year and grade levels. While the overall pattern in correlations across different models remained essentially the same, depending on the model specifications, teachers' VAM estimates were ranked differently and the fluctuations altered the correlations with teachers' SLO quality. Since the value-added estimates of teachers could be easily impacted by a range of factors, it should not be used as the sole or principal indicator in making high-stakes decisions on teachers' rewards or sanctions so as to avoid potential injustices.

Results from the last three research questions also have policy implications. First, the relationship between teachers' VAM estimates and their SLO quality showed statistically significant variation across years and grades for both mathematics and reading. It is still worth noting that since multiple analyses were conducted based on subject/year/grade combinations, there is a problem of multiplicity and even the significant results could be mostly noise. In addition, the relationship in elementary schools and middle schools was statistically significantly different for reading. This finding suggests that teacher effectiveness measured by their students' achievement progress may not be consistently strongly correlated to that measured by an aspect of teacher classroom practice, as indicated by SLO quality. The relationship between the two types of indicators may be stronger in some years, grade levels, or subjects, and weaker in other circumstances.

Moreover, there may be good reasons why the relationship between the VAM estimates and SLO quality is not so strong. For example, teachers who excel at improving

student academic achievement may not be good at developing and writing the learning objectives for students. Similarly those teachers with superior ability in classroom practice and establishing SLOs may not be able to help with students' test performance so effectively. Were that generally the case, measuring teacher effectiveness will never be a simple task. It may involve a variety of strategies for teachers at different grade levels or teaching different subjects. Other indicators of teacher effectiveness, in addition to the VAM estimates and SLO quality, as well as the relationship among different indicators, need be investigated in order to establish a teacher evaluation system that can help to identify different dimensions of teacher quality and, ideally, lead to instructional improvement.

Second, results from the logistic regression analysis of the relationship between teachers' SLO quality scores and their achievement status did not yield significant results. This implied that teachers who excel at developing SLOs may not be superior at the real classroom performance and providing exceptional support for their students to achieve those objectives. Conversely, there may be teachers who granted extraordinary help in encouraging student learning and improving their achievement outcomes but failed to establish satisfactory SLOs in the initial evaluation process. Therefore, the quality of SLOs may not be a good predictor of whether those SLOs can be attained. Given that the SLO quality in this study was assessed by multiple expert evaluators through a reliable process, it may still be considered a good indicator of teacher effectiveness in that they can reflect some aspects of teachers' classroom performance. Therefore, SLO quality scores could still be used as a component of the teacher evaluation system and inform effective classroom instruction.

Third, results from the analyses of VAM estimates and SLO attainment status in the last research question also did not yield significant results, suggesting that teachers who were outstanding at improving student achievement progress may not have achieved their SLOs. Given that students' scores in standardized state assessment were used to produce teachers' VAM estimates while other measures[24] may instead be employed to measure teachers' SLO attainment, this result could indicate the differences between the outcomes (and how they are measured) for the SLOs and the outcomes used to produce the VAM scores. More indicators should be investigated and explored in order to better encompass the comprehensive construct of teacher evaluation.

In addition, the lack of correlation between SLO attainment and VAM estimates, and similarly between SLO attainment and SLO quality, further questions the use of SLO (both the SLO quality and the attainment) as well as VAM estimates in teacher evaluation. As emphasized in the study, both SLO quality and attainment status are designed to measure one aspect of teacher effectiveness through their classroom practice while VAM estimates intend to evaluate a different aspect of teacher effectiveness in improving student achievement outcomes. Teacher quality is a complicated construct and needs multiple indicators to fully represent its various facets. Simply using one or two indicators is inadequate for a defensible evaluation process.

In summary, measuring teacher effectiveness is a more complex task and multiple indicators are required to accurately evaluate the capability and competence of different teachers. This study investigated two types of indicators of teacher effectiveness that

[24] While developing SLOs, teachers may decide on the objectives for their students and the ways to measure student growth. Different assessments may be used (see chapter 3, p75).

represent teacher performance in classroom practice and in fostering student academic progress. As these two indicators were designed to measure different aspects of teacher attributes, they were not expected to be highly correlated. The findings of this study reinforced that teacher evaluation is a multifaceted task, and a portfolio of teacher quality indicators is required in order to have a comprehensive understanding of teacher effectiveness and to measure it effectively.

5.3 Limitations

There are a number of limitations to this study. These are mainly related to the nature of the data. The original dataset was from one research project based on a section of one school district. The two subjects of interest in this study also limited the availability of records for analysis. Therefore, the findings are limited in their generalizability.

First, the small numbers of teachers for the relational analyses between VAM estimates and SLO quality scores is evidently problematic. Larger pools of teachers were used in the beginning of the study to ensure that more reliable VAM estimates for the target teachers could be extracted and further analyzed with their SLOs quality scores. However, the limited number of teachers with SLOs quality scores and attainment status for both mathematics and reading further restricts the generalizability of the conclusions from this study. In addition, the correlation coefficients obtained from many models rarely attained statistical significance (perhaps due to the limited sample size), which also restricts the interpretation of the findings.

In addition, the quality of SLOs that was employed as the indicator of teacher effectiveness is based on one aspect of teacher classroom practice. These SLOs were created by the teachers and treated as a proxy for teachers' effectiveness in this study. The evaluation scores for the quality of the SLOs were regarded as an indicator of teacher effectiveness and used for the relationship analyses with VAM estimates. Given that this is not a typical use of SLO, the results from this study must be interpreted with caution.

Thirdly, regarding the model specifications, due to the limited number of teacher-level records in this dataset, only one teacher-level variable (class size) was adopted in the HLM models in order to keep an ideal case-predictor ratio and preserve the model reliability. Therefore, under the circumstances that there was no statistical significance found for this variable, it was nevertheless included in the models for the value-added analyses.

5.4 Measuring Teacher Effectiveness -- Looking forward

There has been a general consensus on the need to develop a more useful teacher evaluation system, in view of the existing problems with the traditional methods and practices of measuring teacher effectiveness. Evidently, more indicators of teacher effectiveness should be explored and investigated. In this regard, much more research like the MET study is needed to examine the relationship among various indicators of teacher effectiveness before a successful teacher evaluation system can be established. The first area of research needed is in obtaining greater consensus on the components of an ideal teacher evaluation system from the current efforts such as Danielson Framework for teaching, Pianta's model, National Board for Professional Teaching Standards (NBPTS), etc. This study has shed some light on the utility and value of teachers' effectiveness in improving student achievement outcomes progress and in one aspect of their classroom practice. Other important aspects of teacher effectiveness need to be identified to fully capture the construct of teacher quality and the function of teacher evaluation. Given that different aspects of teacher practice may be more important in different contexts, and that there are diverse opinions of the elements that should be included in the perception of teacher effectiveness, as well as their importance, this will be a challenging but rewarding task.

Second, using the quality of teacher-developed student learning objectives to measure teacher effectiveness is a comparatively new approach. Most studies used the proportion of students who achieved the SLOs as an indicator of the teachers' capability in helping students obtain those goals after a learning period. In future research, other

approaches to measuring the SLOs with a focus on evaluating teachers' involvements and contributions may need to be developed to explore more meaningful indicators of teacher quality.

Thirdly, examining the relationship among various indicators of teacher effectiveness is an important way to study the quality and relevance of the indicators. This study focused on investigating the relationship between VAM estimates and SLO quality for two subjects that were broadly tested with standardized assessments. However, a great number of teachers teach subjects with no standardized tests available. Therefore, more research will be needed to investigate the relationship between teachers' ability to help students improve achievement outcomes and their performance in classroom practice in a variety of non-tested subjects.

Fourthly, in comparison to the design of SLOs, value-added models have the advantage of being able to account for students' achievement in prior years. Therefore, SLOs may need further improvements in design, such as the objectives of the SLO may take into account students' prior and current status. In future more research will be needed to explore better ways to evaluate the quality and attainment of SLOs, and to obtain more comprehensive as well as thorough understanding of teachers' classroom practices.

Moreover, results of value-added models vary with respect to controlling for different covariates at different levels of the model and using multiple years of data. Since only one teacher-level covariate was controlled for in the models of this study,

future studies may explore the influence of more teacher-level covariates on the relationship among different teacher effectiveness indicators.

Finally, the idea of measuring teacher effectiveness in fostering student learning progress needs careful consideration. Current initiatives from collecting teachers' credentials, school administrators' observations, students and parents' feedback are all moving the evaluation of teachers in a more quantitative and objective direction. As Skyes (1985) described teaching as a natural, spontaneous and organic human activity, the classroom atmosphere or in other words, students' learning environment, may largely depend on a teacher's personality or cultural background. Therefore, as the new classroom observational rubrics attempt to achieve, more efforts to measure teacher effectiveness should also be focused on evaluating teachers' creativity, teaching style, or in general, their classroom practice to promote student learning; for example, developing more indicators like SLOs that may provide information about teachers' practice and inform classroom instructions. Thus studies of the relationship among more indicators will be needed. As the small steps achieved in this study suggest, teacher effectiveness is complex by nature and a portfolio of indicators measuring different aspects of teaching will be needed to build a successful teacher evaluation system.

REFERENCES

Aaronson, Daniel, Lisa Barrow, and William Sanders. 2007. "Teachers and Student Achievement in the Chicago Public High Schools." Journal of Labor Economics 25(1): 95-135.

Alicias, E. R. Jr. (2005, May 6). Toward an objective evaluation of teacher performance: The use of variance partitioning analysis, VPA. Education Policy Analysis Archives, 13(30). Retrieved [date] from http://epaa.asu.edu/epaa/v13n30/.

Alliance for Excellent Education. (2008) Measuring and Improving the Effectiveness of High School Teachers.

Baker, E. l., Barton, P. E., Linda Darling-HammonD, E. H., Ladd, H. F., Linn, R. l., Ravitch, D., … Shepard, L. a. (2010). Problems with the Use of stUdent test scores to evaluate teachers. Economic Policy.

Ballou, D. (2008). Test Scaling and Value-Added Measurement. Paper presented at the National Conference on Value-Added Modeling.

Benjamini, Yoav; Braun, Henry. John W. Tukey's contributions to multiple comparisons. Ann. Statist. 30 (2002), no. 6, 1576--1594. doi:10.1214/aos/1043351247. http://projecteuclid.org/euclid.aos/1043351247.

Berk, R. A. (2005). Survey of 12 Strategies to Measure Teaching Effectiveness. Strategies, 17(1), 48-62.

Bill & Melinda Gates Foundation (2010). Learning About Teaching: Initial Findings from the Measures of Effective Teaching Project. Seattle: Author.

Bill and Melinda Gates Foundation, Gathering Feedback for Teaching (2012)

Braun, H. I. (2005). Using Student Progress To Evaluate Teachers : A Primer on Value-Added Models. ETS, 16.

Braun, H., Chudowsky, N., & Koenig, J. (2010). Getting Value Out of Value-added. Social Sciences.

Buckley, K., & Marion, S. (2011). A survey of approaches used to evaluate educators in non-tested grades and subjects. Dover, NH: National Center for the Improvement of Educational Assessment. Retrieved February, 21, 2012.

Chait, R. 2010. "Removing Chronically Ineffective Teachers." Washington: Center for American Progress.

Clotfelter, C. T., Ladd, H. F., & Vigdor, J. L. (2007). How and Why do Teacher Credentials Matter For Student Achievement? National Bureau of Economic Research.

Coggshall, J.G., Ott, A., & Lasagna, M., (2011). Convergence and Contradictions in Teachers' Perceptions of Policy Reform Ideas

CTAC. (2013). It's More Than Money - TIF-LEAP, Charlotte-Mecklenburg Schools.

CTAC. (2005). Catalyst for Change.

CTAC. (2001). Pay for Performance. Training, (December).

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

Darling-Hammond, L. (2010). The flat world and education: How America's commitment to equity will determine our future. Teachers College Press.

Darling-Hammond, L., & Richardson, N. (2009). Research review/teacher learning: What matters. Educational leadership, 66(5), 46-53.

Darling-Hammond, L. (2008). Teacher learning that supports student learning. Teaching for intelligence, 92-93.

Darling-Hammond, L. 2007. Recognizing and enhancing teacher effectiveness: A policymaker's guide. In L. Darling-Hammond and C. D. Prince (eds.), Strengthening teacher quality in high-need schools—policy and practice. Washington, DC: The Council of Chief State School Officers.

Darling-Hammond, L.. and Sykes, G.. (2003, September 17). Wanted: A national teacher supply policy for education: The right way to meet the "Highly Qualified Teacher" challenge? Education Policy Analysis Archives, 11(33). Retrieved [Date] from http://epaa.asu.edu/epaa/v11n33/.

Darling-Hammond, L., & Youngs, P. (2002). Defining" highly qualified teachers": What does" scientifically-based research" actually tell us?. Educational Researcher, 31(9), 13-25.

Darling-Hammond, L. (2000). Teacher Quality and Student Achievement : A Review of State Policy Evidence Previous Research. Education, 8(1), 1-44.

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, 285–328.

Danielson, C., & McGreal,T. L. (2005).Teacher Evaluation to Enhance Professional Practice. E-book.

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

Doran, H. C., & Izumi, L. T. (2004). Putting education to the test: A value-added model for California. San Francisco: Pacific Research Institute.

Fisher, R.A. (1915). "Frequency distribution of the values of the correlation coefficient in samples of an indefinitely large population". Biometrika (Biometrika Trust) 10 (4): 507–521. JSTOR 2331838.

Flicek, M. & Wong, K. (2003). The challenge of using large-scale assessment to hold schools accountable. Submitted.

Glass, G., Hopkins, K. (1995). Statistical Methods in Education and Psychology. Needham Heights, MA: Allyn & Bacon.

Glazerman, S., Goldhaber D., Loeb, S. U., Raudenbush, S., Staiger, D. U., & Whitehurst, G., (2011). Passing Muster: Evaluating Teacher Evaluation Systems. Policy.

Glazerman, S., Loeb, S., Goldhaber, D., Staiger, D., Raudenbush, S., & Whitehurst, G. (2010). Evaluating Teachers: The Important Role of Value-Added.

Goe, L. (2007). The Link Between Teacher Quality and Student Outcomes : A Research Synthesis, (October).

Goe, L. (2011). Student Growth in Non-Tested Subjects and for At-Risk Students, powerpoint presentation.

Goldhaber, D., & Hansen, M. (2010). Assessing the Potential Estimates of Teacher for Making Tenure Decisions. CALDER Working Paper No. 31, 0, 57.

Goldhaber, D., & Hansen, M. (2008). Assessing the Potential of Using Value-Added Estimates of Teacher Job Performance for Making Tenure Decisions. Brief 3. National Center for Analysis of Longitudinal Data in Education Research.

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

Goldhaber, D.D. and D.Brewer (2000) "Does Teacher Certification Matter? High School Teacher Certification Status and Student Achievement." Educational Evaluation and Policy Analysis v.22 pp.129-45.

Goldhaber, Dan D, and Dominic J. Brewer (1997), "Why Don't Schools and Teachers Seem to Matter? Assessing the Impact of Unobservables on Educational Productivity," Journal of Human Resources. Forthcoming, 32(3).

Halverson, R., Kelley, C., & Kimball, S. (2004). Implementing teacher evaluation systems: How principals make sense of complex artifacts to shape local instructional practice. In W. Hoy & C. Miskel (Eds.), Research and theory in educational administration. (Vol. 3). Greenwich, CT: Information Age Publishing.

Hanushek, E. A. (2011). The economic value of higher teacher quality. Economics of Education Review, 30(3), 466-479.

Hanushek, E.A., Kain, J.F., O'Brien, D.M., Rivkin, S.G. (2005). "The market for teacher quality". Working Paper 11154 (February).

Hanushek, Eric. (2003). The Failure of Input Based Schooling Policies. Economic Journal, 113, no.485: F64-F98.

Hanushek, E. A., Kain, J. F., & Rivkin, S. G. (1998). Teachers, schools, and academic achievement (Working Paper No. 6691). Cambridge, MA: National Bureau of Economic Research.

Hanushek, E. A. (1996). A more complete picture of school resource policies. Review of Educational Research, 66(3), 397-409.

Hanushek, Eric A. (1992). "The trade-off between child quantity and quality." Journal of Political Economy 100,no.1 (February):84-117.

Hanushek, E. (1972). Education and Race. Lexington, MA: D.C. Heath and Company.

Hargreaves, A., & Braun, H. (2013). DATA–DRIVEN IMPROVEMENT AND ACCOUNTABILITY. Boston College: National Education Policy Center. Retrieved October, 24, 2013.

Douglas N. Harris and Andrew Anderson, Bias of Public Sector Worker Performance Monitoring: Theory and Empirical Evidence from Middle School Teachers, (Paper presented at the Association for for Policy Analysis and Management 2012.)

Harris, D.N., and Sass, T. (2005). Value-added models and the measurement of teacher quality. Paper presented at the annual conference of the American Education Finance Association, Louisville, KY, March 17-19.

Harris, D. N., Sass, T. R., "What makes for a good teacher and who can tell?" (Working Paper no. 30, National Center for the Analysis of Longitudinal Data in Education Research (CALDER), Urban Institute, Washington, DC, 2009.)

Harris, D. N., (2012) "How Do Value-Added Indicators Compare to Other Measures of Teacher Effectiveness?" CARNEGIE KNOWLEDGE NETWORK. http://www.carnegieknowledgenetwork.org/briefs/value-added/value-added-other-measures/

Haycock, K. (1998). Good teaching matters: How well-qualified teachers can close the gap. Washington, DC: Education Trust.

Hill, H C.; Umland, K; Litke, E; and Kapitula, L R., "Teacher Quality and Quality Teaching: Examining the Relationship of a Teacher Assessment to Practice" (2012).Peer Reviewed Articles. Paper 2. http://scholarworks.gvsu.edu/sta_articles/2

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

Hill, H. C., Rowan, B., & Ball, D. L. (2005). Effects of teachers' mathematical knowledge for teaching on student achievement. American educational research journal, 42(2), 371-406.

Jacob, B. & Lefgren, L. (2008). Can principals identify effective teachers? Evidence on subjective performance evaluation in education. Journal of Labor Economics. 26(1), 101-36.

Johnson, B.L. (1997). An organizational analysis of multiple perspectives of effective teaching: Implications for teacher evaluation. Journal of Personnel Evaluation in Education, 11(1), 69-88.

Jordan, H., Mendro, R., & Weerasinghe, D. (1997). Teacher effects on longitudinal student achievement. Dallas, TX: Dallas Independent School District.

Kane, T.J., Taylor, E., Tyler, J., and Wooten A. (2011). "Evaluating Teacher Effectiveness: Can classroom observations identify practices that raise achievement?" Education Next, Summer 2011, Vol. 11, No. 3.

Kane, T. J., Rockoff, J.E., & Staiger, D.O. (2008). What does certification tell us about teacher effectiveness? Evidence from New York City. Economics of Education Review, 27(6), 615-31.

Koedel, C., & Betts, J. "Does Student Sorting Invalidate Value‐Added Models of Teacher Effectiveness? An Extended Analysis of the Rothstein Critique," Unpublished manuscript, 2009.

Kuh, G., & Ikenberry, S. (2009). More than you think, less than we need: Learning outcomes assessment in American higher education. Urbana-Champaign: University of Illinois and Indiana University, National Institute for Learning Outcomes Assessment.

Kuppermintz, H. (2003). Teacher effects and Teacher Effectiveness: A Validity Investigation of the Tennessee Value Added Assessment System. Educational Evaluation and Policy Analysis, Fall 2003, Vol,25, No. 3, pp. 287-298.

Linn, R. L. (2004). Rethinking the No Child Left Behind Accountability System Rethinking the No Child Left Behind Accountability System.

Locke, E.A., & Latham, G.P. (2002). Building a practically useful theory of goal setting and task motivation: A 35-year odyssey. American Psychologist, 57, 705–717.

Marion, S., Depascale, C., Domaleski, C., Gong, B., & Diaz-bilello, E. (2012). Considerations for Analyzing Educators' Contributions to Student Learning in Non-tested Subjects and Grades with a Focus on Student Learning Objectives. Center for Assessment.

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

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

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

Mendro, R., H. Jordan, E. Gomez, M. Anderson, and K. Bembry (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.

Meyer, R. H. (2001). Estimation of teacher and school performance in the Denver public schools: Afeasibility study. Madison: University of Wisconsin-Madison, Wisconsin Center for Educational Research.

McCaffrey, D. F., & Lockwood, J. R. (2008). Value-Added Models : Analytic Issues.

McCaffrey, D. F., Lockwood, J. R., & Hamilton, L. S. (2003). Evaluating Models for Teacher Accountability. Distribution (p. 35).

McCall, M. S., Kingsbury, G. G., & Olson, A. (2004). Individual Growth and School Success. Evaluation, (April). Measuring Student Growth for Teachers in Non-Tested Grades and Subjects : A Primer. Network, 1-9.

Milanowski, A. T., (2004) The Relationship Between Teacher Performance Evaluation Scores and Student Achievement: Evidence From Cincinnati. Peabody Journal of Education. Volume 79, Issue 4, 2004

Murnane, R.J. (1975). The Impact of School Resources on the Learning of Inner City Children. Cambridge, MA: Ballinger Publishing Co.

Myers, Jerome L.; Well, Arnold D. (2003). Research Design and Statistical Analysis (2nd ed.). Lawrence Erlbaum. p. 508. ISBN 0-8058-4037-0.

National Research Council. (2011). Incentives and test-based accountability in public educa- tion. Committee on Incentives and Test-Based Accountability in Public Education. M. Hout, N. Chudowsky, and S. W. Elliott (Eds.). Board on Testing and Assessment, Center for Education, Division of Behavioral and Social Sciences and Education. Washington, DC: The National Academies Press. Organisation

Nguyen, P., Terlouw, C., & Pilot, A. (2006).Culturally appropriate pedagogy: The case of group learning in a Confucian heritage culture context. Intercultural Education, 17(1), 1-19.

Nye, B, Konstantopoulos, S. and Hedges, L. (2004). "How Large Are Teacher Effects?" Educational Evaluation and Policy Analysis, 26 (3): 237-257. 25

Odden, A., Borman, G., & Fermanich, M. (2004). Assessing Teacher , Classroom , and School Effects , Including Fiscal Effects, 79(4), 4-32.

Olson, DR (2003). Psychological theory and educational reform: How school remakes mind and society. Cambridge, England: Cambridge University Press.

O'Malley, K.J., Murphy, S., McClarty, K.L., Murphy, D., McBride, Y. (2011). Overview of Student Growth Models. Pearson's White Paper. http://researchnetwork.pearson.com/wp-content/uploads/StudentGrowthWP083111.pdf

Paek, P. L., Braun, H., Ponte, E., Trapani, C., & Powers, D. (2010). AP Biology teacher characteristics and practices and their relationship to student achievement. In P. M. Sadler, G. Sonnert, R. H. Tai, K. Klopfenstein (Eds.), AP: A critical examination of the Advanced Placement Program (pp. 63-84). Cambridge, MA: Harvard Education Press.

Peterson, K. (2000). Teacher evaluation: A comprehensive guide to new directions and practices. (2nd ed.). Thousand Oaks, CA: Corwin Press.

Peterson, K. D. (1995). Teacher evaluation: A comprehensive guide to new directions and practices. Thousand Oaks, CA: Corwin.

Peterson, C. (1982). A century's growth in teacher evaluation in the United States. New York: Vantage.

Piche, D. (2007). Basically a Good Model. Educationnext. http://educationnext.org/basically-a-good-model/

Prince, C. D., Schuermann, P. J., Guthrie, J. W., Witham, P. J., Milanowski, A. T., & Thorn, C. A. (n.d.). (2009). The Other 69 Percent : Guide to Implementation : Resources for Applied Practice.

Reardon, S. F., & Raudenbush, S. W. (2009). Assumptions of Value-Added Models for Estimating School Effects, 1–40.

Reckase, M. D. (2004). The real world is more complicated than we would like. Journal of Educational and Behavioral Statistics, 29(1), 117-120.

Rice, J. K. (2003). Executive summary and Introduction. In J. K. Rice, Teacher quality: Understanding the effectiveness of teacher attributes (pp. v-vii and 1-7). Washington, DC: Economic Policy Institute.

Rivers, J. C. (1999). The Impact of Teacher Effect on Student Math Competency Achievement. Doctor of education thesis, University of Tennessee, Knoxville, Knoxville TN, 37996.

Rivkin, E. A., Hanushek, E. A., & Kain, J. F. (2001). Teachers, schools, and academic achievement. Washington, DC: National Bureau of Economic Research.

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

Rockoff, J. E., Jacob, B. A., Kane, T. J., & Staiger, D. O. (2011). Can you recognize an effective teacher when you recruit one?. Education, 6(1), 43-74.

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

Rothstein, Jesse (2011). Review of 'Learning About Teaching: Initial Findings from the Measures of Effective Teaching Project. Boulder, CO: National Education Policy Center.

Rowan B, Correnti R, Miller RJ. (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 2002;104:1525–1567.

Rubin, Donald. 1986. "Which Ifs Have Causal Answers? Discussion of Holland's Statistics and Causal Inference'." Journal of the American Statistical Association 81: 961-62.

Sanders, W. L. (2000). Value-added assessment from student achievement data. Cary, NC: Create National Evaluation Institute.

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

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

Sanders, W. L., & Horn, S. P. (1995). Educational assessment reassessed: The usefulness of standardized and alternative measures of student achievement as indicators for the assessment of educational outcomes. Educational Policy and Analysis Archives, 3

Schacter, J., & Thum. (2003). Paying for high- and low-quality teaching. Economics of Education Review, 23(4), 411-430. doi:10.1016/j.econedurev.2003.08.002

Shinkfield, A. J. and D. L. Stufflebeam. 1995. Teacher Evaluation: Guide to Effective Practice. Kalamazoo, MI: Center for Research on Educational Accountability and Teacher Evaluation.

Solomon, L., White, J. T., Cohen, D. & Woo, D. (2007). The effectiveness of the Teacher Advancement Program. National Institute for Excellence in Teaching, 2007.

Soto, A. C., Sireci, S. G., & Keller, L. A. (2011). Evaluating Teachers Using Value-Added Models: Current Practices and Validity Evidence 1. Educational Assessment, (792), 1-50.

Stiggens, R. J., & Duke, D. (1988) The case for commitment to teacher growth: Research on teacher evaluation. Albany: State University of New York Press.

Stronge, J.H. & Tucker, P.D. (1999). The politics of teacher evaluation: A case study of new system design and implementation. Journal of Personnel Evaluation in Education, 13(4), 339-360.

Sykes, G. (1985). The School and the University. Chapter 10: Teacher Education in the United States. University of California Press.

Taylor, E.S, & Tyler, J.H. (2012). Can Teacher Evaluation Improve Teaching? Education Next, 12(4). - See more at: http://cepa.stanford.edu/content/can-teacher-evaluation-improve-teaching#sthash.nmmBc82i.dpuf

Thomas, E. (1997). Developing a culture-sensitive pedagogy: Tackling a problem of melding 'global culture' with existing cultural contexts. International Journal of Educational Development, 17 (1), 13-26.

Thum, Y. M. (2003). Measuring progress towards a goal: estimating teacher productivity using a multivariate multilevel model for value-added analysis. Sociological Methods and Research, 32(2), 153–207.

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

Webster, W. J., Mendro., R. L., Orsak, T. H., & WeerasLnghe, D. (1998, April). An application of hierarchical linear modeling to the estimation of school and teacher effect. Paper presented at the annual meeting of the American Educational Research Association, San Diego, CA.

Webster, W. J., Mendro, R. L., Orsak, T. H., & Weerasinghe, D. (1996, April). The applicability of selected regression and hierarchical linear models to the estimation of school and teacher effects. Paper presented at the annual meeting of the National Council on Measurement in Education, New York.

Weimer, M. (2013). Learner-centered teaching: Five key changes to practice. John Wiley & Sons.

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.

Wenglinsky, H. (2002). The Link between Teacher Classroom Practices and Student Academic Performance. Education policy analysis archives, 10(12), n12.

Willms, J.D. (2008). Seven key issues for assessing "value-added" in education. Paper prepared for the workshop of the Committee on Value-Added Methodology for Instructional Improvement, Program Evaluation, and Educational Accountability, National Research Council, Washington, DC, November 13-14. Available: http://www7.national academies.org/bota/VAM_Workshop_Agenda.html.

Wilson, S. M., & Floden, R. E. (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.

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

Appendix A. HLM Model Specifications

Table A.1 Results of HLMs for Research Question Two: 2008-Grade 4

	Mathematics			Reading		
Level-1 Fixed Effect[a]	Coef.	s.e.	*p*-value	Coef.	s.e.	*p*-value
Intercept, γ_{00}	346.19	0.37	**<0.001**	340.58	0.37	**<0.001**
2007_Grade3_Score, γ_{10}	0.68	0.03	**<0.001**	0.62	0.02	**<0.001**
Student Race/Ethnicity[b]						
American Indian, γ_{20}	-0.65	3.84	0.866	3.61	3.05	0.236
Asian, γ_{30}	-0.14	0.81	0.863	0.40	0.82	0.630
Hispanic, γ_{40}	1.97	0.36	**<0.001**	1.20	0.49	**0.015**
Multi-race, γ_{50}	1.09	0.67	0.101	0.14	0.93	0.884
White, γ_{60}	0.30	0.78	0.703	-0.77	0.92	0.401
Gifted status, γ_{70}	3.72	0.85	**<0.001**	3.33	0.81	**<0.001**
LEP status, γ_{80}	-1.56	0.52	**0.003**	-2.84	0.58	**<0.001**
SWD status, γ_{90}	-2.49	0.57	**<0.001**	-2.92	0.58	**<0.001**

Random Effect	Variance Component	df	χ^2	*p*-value	Variance Component	df	χ^2	*p*-value
Intercept, u_0	14.77	163	872.28	**<0.001**	12.86	156	667.51	**<0.001**
Level-1 effect, γ_{ij}	24.80				27.05			
Reliability estimate								
Intercept	0.786				0.753			

Note. Bolded values are significant at .05.

[a] Fixed effects at level 2 is not reported here, since they are not of research interests in research question one; [b] Reference group is African American.

Table A.2 Results of HLMs for Research Question Two: 2008-Grade 5

	Mathematics			Reading		
Level-1 Fixed Effect[a]	Coef.	s.e.	*p*-value	Coef.	s.e.	*p*-value
Intercept, γ_{00}	351.48	0.35	**<0.001**	345.87	0.34	**<0.001**
2007_Grade4_Score, γ_{10}	0.70	0.02	**<0.001**	0.61	0.02	**<0.001**
Student Race/Ethnicity[b]						
American Indian, γ_{20}	-0.70	1.83	0.704	-0.64	0.95	0.500
Asian, γ_{30}	2.16	0.82	**0.008**	-0.10	0.59	0.862
Hispanic, γ_{40}	1.47	0.34	**<0.001**	1.48	0.41	**<0.001**
Multi-race, γ_{50}	0.92	0.96	0.335	0.00	1.13	1.000
White, γ_{60}	1.72	0.56	**0.002**	1.67	0.54	**0.002**
Gifted status, γ_{70}	2.34	0.50	**<0.001**	2.10	0.72	**0.004**
LEP status, γ_{80}	--	--	**--**	-2.75	0.48	**<0.001**
SWD status, γ_{90}	-3.18	0.49	**<0.001**	-4.22	0.54	**<0.001**
Random Effect	Variance Component	df	χ^2	*p*-value	Variance Component	df	χ^2	*p*-value
Intercept, u_0	16.13	160	1215.76	**<0.001**	11.37	160	762.05	**<0.001**
Level-1 effect, γ_{ij}	19.04				23.96			
Reliability estimate								
Intercept	0.856				0.770			

Note. Bolded values are significant at .01.

[a] Fixed effects at level 2 is not reported here, since they are not of research interests in research question one; [b] Reference group is African American.

Table A.3 Results of HLMs for Research Question Two: 2008-Grade 6

	Mathematics			Reading		
Level-1 Fixed Effect[a]	Coef.	s.e.	*p*-value	Coef.	s.e.	*p*-value
Intercept, γ_{00}	351.50	0.64	**<0.001**	347.63	0.69	**<0.001**
2007_Grade5_Score, γ_{10}	0.68	0.02	**<0.001**	0.77	0.02	**<0.001**
Gender[b], γ_{20}	-0.77	0.22	**<0.001**	-1.33	0.29	**<0.001**
Student Race/Ethnicity[c]						
American Indian, γ_{30}	-1.95	1.64	0.235	2.24	0.79	**0.004**
Asian, γ_{40}	0.99	0.77	0.200	1.17	0.81	0.148
Hispanic, γ_{50}	0.83	0.40	**0.036**	0.85	0.48	0.075
Multi-race, γ_{60}	1.89	0.78	**0.016**	1.52	0.88	0.084
White, γ_{70}	1.16	0.56	**0.036**	1.90	0.63	**0.003**
Gifted status, γ_{80}	2.62	0.71	**<0.001**	--	--	--
LEP status, γ_{90}	-1.05	0.45	**0.019**	-1.41	0.44	**<0.001**
SWD status, γ_{100}	-2.44	0.72	**<0.001**	-2.91	0.64	**<0.001**

Random Effect	Variance Component	df	χ^2	*p*-value	Variance Component	df	χ^2	*p*-value
Intercept, u_0	15.93	43	810.05	**<0.001**	20.88	47	808.58	**<0.001**
Level-1 effect, γ_{ij}	22.71				24.95			
Reliability estimate								
Intercept	0.922				0.916			

Note. Bolded values are significant at .05.

[a] Fixed effects at level 2 is not reported here, since they are not of research interests in research question one; [b] Reference group is female; [c] Reference group is African American.

Table A.4 Results of HLMs for Research Question Two: 2008-Grade 7

	Mathematics			Reading		
Level-1 Fixed Effect[a]	Coef.	s.e.	*p*-value	Coef.	s.e.	*p*-value
Intercept, γ_{00}	355.02	0.62	**<0.001**	351.52	0.46	**<0.001**
2007_Grade6_Score, γ_{10}	0.73	0.02	**<0.001**	0.67	0.02	**<0.001**
Gender[b], γ_{20}	-0.61	0.22	**0.005**	--	--	**--**
Student Race/Ethnicity[c]						
American Indian, γ_{30}	1.27	1.30	0.328	0.63	1.43	0.660
Asian, γ_{40}	2.79	0.59	**<0.001**	0.92	0.51	0.069
Hispanic, γ_{50}	1.13	0.30	**<0.001**	1.49	0.36	**<0.001**
Multi-race, γ_{60}	0.96	0.66	0.144	0.32	0.73	0.660
White, γ_{70}	0.91	0.35	**0.009**	2.02	0.38	**<0.001**
Gifted status, γ_{80}	0.45	5.39	**<0.001**	1.21	0.34	**<0.001**
LEP status, γ_{90}	--	--	**--**	-1.09	0.45	**0.016**
SWD status, γ_{100}	0.49	-3.01	**0.003**	-3.33	0.58	**<0.001**

Random Effect	Variance Component	df	χ^2	*p*-value	Variance Component	df	χ^2	*p*-value
Intercept, u_0	28.06	80	2228.34	**<0.001**	14.52	45	682.27	**<0.001**
SWD slope, u_{10}	--	--	--	**--**	1.67	45	64.21	**0.031**
Level-1 effect, γ_{ij}	22.47				22.67			
Reliability estimate								
Intercept	0.941				0.874			
SWD	--				0.119			

Note. Bolded values are significant at .05.

[a] Fixed effects at level 2 is not reported here, since they are not of research interests in research question one; [b] Reference group is female; [c] Reference group is African American.

Table A.2 Results of HLMs for Research Question Two: 2008-Grade 8

	Mathematics			Reading		
Level-1 Fixed Effect[a]	Coef.	s.e.	*p*-value	Coef.	s.e.	*p*-value
Intercept, γ_{00}	357.86	0.46	**<0.001**	355.34	0.45	**<0.001**
2007_Grade7_Score, γ_{10}	0.42	0.02	**<0.001**	0.43	0.02	**<0.001**
2006_Grade6_Score, γ_{20}	0.30	0.02	**<0.001**	0.36	0.02	**<0.001**
Student Race/Ethnicity[b]						
American Indian, γ_{30}	-1.59	1.02	0.119	-0.65	0.82	0.424
Asian, γ_{40}	0.73	0.55	0.184	-0.27	0.51	0.591
Hispanic, γ_{50}	1.45	0.33	**<0.001**	0.50	0.29	**0.090**
Multi-race, γ_{60}	0.67	0.80	0.404	2.53	0.84	**0.003**
White, γ_{70}	1.10	0.43	**0.010**	1.18	0.36	**0.001**
LEP status, γ_{80}	-1.82	0.49	**<0.001**	-1.19	0.51	**0.019**

Random Effect	Variance Component	df	χ^2	*p*-value	Variance Component	df	χ^2	*p*-value
Intercept, u_0	16.31	87	1155.95	**<0.001**	15.54	87	1435.05	**<0.001**
Level-1 effect, γ_{ij}	18.47				17.60			
Reliability estimate								
Intercept	0.912				0.913			

Note. Bolded values are significant at .01.

[a] Fixed effects at level 2 is not reported here, since they are not of research interests in research question one; [b] Reference group is African American.

Table A.2 Results of HLMs for Research Question Two: 2009-Grade 4

	Mathematics			Reading		
Level-1 Fixed Effect[a]	Coef.	s.e.	*p*-value	Coef.	s.e.	*p*-value
Intercept, γ_{00}	347.10	0.51	**<0.001**	340.82	0.47	**<0.001**
2008_Grade3_Score, γ_{10}	0.70	0.03	**<0.001**	0.64	0.02	**<0.001**
Student Race/Ethnicity[b]						
American Indian, γ_{20}	3.46	2.05	0.092	4.54	2.12	**0.033**
Asian, γ_{30}	1.68	0.77	**0.030**	2.70	1.00	**0.007**
Hispanic, γ_{40}	2.33	0.58	**<0.001**	1.75	0.62	**0.005**
Multi-race, γ_{50}	2.45	1.15	**0.034**	1.54	1.17	0.190
White, γ_{60}	1.72	0.59	**0.004**	1.64	0.98	**0.094**
Gifted status, γ_{70}	3.50	0.74	**<0.001**	3.11	0.84	**<0.001**
LEP status, γ_{80}	-1.66	0.75	**0.028**	-2.73	0.88	**0.002**
SWD status, γ_{90}	-3.02	0.75	**<0.001**	-1.57	0.64	**0.014**

Random Effect	Variance Component	df	χ^2	*p*-value	Variance Component	df	χ^2	*p*-value
Intercept, u_0	13.28	71	478.60	**<0.001**	10.53	72	352.55	**<0.001**
Level-1 effect, γ_{ij}	24.47				28.18			
Reliability estimate								
Intercept	0.847				0.792			

Note. Bolded values are significant at .01.

[a] Fixed effects at level 2 is not reported here, since they are not of research interests in research question one; [b] Reference group is African American.

Table A.2 Results of HLMs for Research Question Two: 2009-Grade 5

	Mathematics			Reading		
Level-1 Fixed Effect[a]	Coef.	s.e.	*p*-value	Coef.	s.e.	*p*-value
Intercept, γ_{00}	353.02	0.41	**<0.001**	346.32	0.18	**<0.001**
2008_Grade4_Score, γ_{10}	0.51	0.03	**<0.001**	0.42	0.03	**<0.001**
2007_Grade3_Score, γ_{10}	0.27	0.03	**<0.001**	0.30	0.03	**<0.001**
Student Race/Ethnicity[b]						
American Indian, γ_{20}	3.46	2.05	0.092	--	--	--
Asian, γ_{30}	1.68	0.77	**0.030**	--	--	--
Hispanic, γ_{40}	2.33	0.58	**<0.001**	--	--	--
Multi-race, γ_{50}	2.45	1.15	**0.034**	--	--	--
White, γ_{60}	1.72	0.59	**0.004**	--	--	--
Gifted status, γ_{70}	3.50	0.74	**<0.001**	--	--	**--**
LEP status, γ_{80}	-1.66	0.75	**0.028**	--	--	--
SWD status, γ_{90}	-3.02	0.75	**<0.001**	-1.96	0.55	**<0.001**

Random Effect	Variance Component	df	χ^2	*p*-value	Variance Component	df	χ^2	*p*-value
Intercept, u_0	13.28	71	478.60	**<0.001**	1.42	145	221.99	**<0.001**
Level-1 effect, γ_{ij}	24.47				19.95			
Reliability estimate								
Intercept	0.847				0.352			

Note. Bolded values are significant at .01.

[a] Fixed effects at level 2 is not reported here, since they are not of research interests in research question one; [b] Reference group is African American.

Table A.2 Results of HLMs for Research Question Two: 2009-Grade 6

	Mathematics			Reading		
Level-1 Fixed Effect[a]	Coef.	s.e.	*p*-value	Coef.	s.e.	*p*-value
Intercept, γ_{00}	352.84	0.75	**<0.001**	348.27	0.70	**<0.001**
2008_Grade5_Score, γ_{10}	0.48	0.03	**<0.001**	0.44	0.02	**<0.001**
2007_Grade4_Score, γ_{10}	0.35	0.03	**<0.001**	0.37	0.02	**<0.001**
Student Race/Ethnicity[b]						
American Indian, γ_{20}	-1.91	1.27	0.133	-1.60	1.85	0.386
Asian, γ_{30}	0.49	0.61	0.426	1.55	0.63	**0.013**
Hispanic, γ_{40}	-0.26	0.26	0.311	0.43	0.38	0.252
Multi-race, γ_{50}	-0.38	0.63	0.548	-0.71	0.82	0.387
White, γ_{60}	1.21	0.47	**0.010**	1.23	0.64	0.056
Gifted status, γ_{70}	1.34	0.54	**0.013**	0.86	0.43	**0.045**
SWD status, γ_{90}	-1.08	0.48	**0.026**	--	--	--

Random Effect	Variance Component	df	χ^2	*p*-value	Variance Component	df	χ^2	*p*-value
Intercept, u_0	18.46	35	1133.67	**<0.001**	18.26	42	630.10	**<0.001**
Level-1 effect, γ_{ij}	17.74				22.94			
Reliability estimate								
Intercept	0.958				0.903			

Note. Bolded values are significant at .01.

[a] Fixed effects at level 2 is not reported here, since they are not of research interests in research question one; [b] Reference group is African American.

Table A.2 Results of HLMs for Research Question Two: 2009-Grade 7

	Mathematics			Reading		
Level-1 Fixed Effect[a]	Coef.	s.e.	*p*-value	Coef.	s.e.	*p*-value
Intercept, γ_{00}	355.19	0.67	**<0.001**	351.21	0.72	**<0.001**
2008_Grade6_Score, γ_{10}	0.49	0.02	**<0.001**	0.45	0.02	**<0.001**
2007_Grade5_Score, γ_{10}	0.35	0.02	**<0.001**	0.38	0.02	**<0.001**
Student Race/Ethnicity[b]						
American Indian, γ_{20}	2.77	1.63	0.089	0.49	1.41	0.730
Asian, γ_{30}	2.00	0.63	**0.001**	-0.12	0.70	0.860
Hispanic, γ_{40}	1.89	0.35	**<0.001**	0.88	0.33	**0.008**
Multi-race, γ_{50}	1.32	0.74	0.074	0.51	0.60	0.399
White, γ_{60}	0.07	0.48	0.887	0.15	0.66	0.815
Gifted status, γ_{70}	1.59	0.47	**<0.001**	1.32	0.52	**0.011**
LEP status, γ_{80}	-2.27	0.39	**<0.001**	--	--	--
SWD status, γ_{90}	-1.77	0.58	**0.002**	--	--	--

Random Effect	Variance Component	df	χ^2	*p*-value	Variance Component	df	χ^2	*p*-value
Intercept, u_0	14.55	35	808.27	**<0.001**	19.05	37	875.19	**<0.001**
Level-1 effect, γ_{ij}	19.99				19.58			
Reliability estimate								
Intercept	0.943				0.942			

Note. Bolded values are significant at .01.

[a] Fixed effects at level 2 is not reported here, since they are not of research interests in research question one; [b] Reference group is African American.

Table A.4 Results of HLMs for Research Question Two: 2009-Grade 8

	Mathematics			Reading		
Level-1 Fixed Effect[a]	Coef.	s.e.	*p*-value	Coef.	s.e.	*p*-value
Intercept, γ_{00}	359.47	0.41	**<0.001**	355.83	0.42	**<0.001**
2007_Grade7_Score, γ_{10}	0.38	0.02	**<0.001**	0.45	0.04	**<0.001**
2006_Grade6_Score, γ_{20}	0.26	0.02	**<0.001**	0.25	0.05	**<0.001**
Gender[b], γ_{30}	-0.74	0.21	**<0.001**	0.35	0.17	**0.038**
Student Race/Ethnicity[c]						
American Indian, γ_{40}	0.93	0.77	0.226	1.02	1.21	0.401
Asian, γ_{50}	1.91	0.78	**0.017**	1.30	0.46	**0.005**
Hispanic, γ_{60}	1.06	0.29	**<0.001**	0.71	0.27	**0.009**
Multi-race, γ_{70}	1.21	0.71	0.087	1.92	0.61	**0.002**
White, γ_{80}	0.97	0.38	**0.010**	0.99	0.32	**0.002**
Gifted status, γ_{90}	1.66	0.55	**0.003**	-2.45	0.47	**<0.001**
LEP status, γ_{100}	-0.87	0.45	**0.053**	1.26	0.38	**<0.001**
SWD status, γ_{110}	-1.73	0.44	**<0.001**	-1.35	0.43	**0.002**

Random Effect	Variance Component	df	χ^2	*p*-value	Variance Component	df	χ^2	*p*-value
Intercept, u_0	11.07	29	462.47	**<0.001**	12.76	57	832.94	**<0.001**
Asian slope, u_{50}	6.98	29	48.31	**0.014**	--	--	--	--
LEP slope, u_{100}	--	--	--	**--**	3.21	57	85.08	**0.009**
Level-1 effect, γ_{ij}	16.91				16.87			
Reliability estimate								
Intercept	0.930				0.905			
Asian	0.365				0.311			

Note. Bolded values are significant at .05.

[a] Fixed effects at level 2 is not reported here, since they are not of research interests in research question one; [b] Reference group is female; [c] Reference group is African American.

Table A.4 Results of HLMs for Research Question Two: 2010-Grade 4

	Mathematics			Reading		
Level-1 Fixed Effect[a]	Coef.	s.e.	*p*-value	Coef.	s.e.	*p*-value
Intercept, γ_{00}	347.48	0.47	**<0.001**	341.19	0.46	<0.001
2009_Grade3_Score, γ_{20}	0.70	0.02	**<0.001**	0.64	0.02	<0.001
Gender[b], γ_{30}	--	--	**--**	-1.01	0.37	0.007
Student Race/Ethnicity[c]						
American Indian, γ_{40}	4.63	1.84	**0.012**	-4.19	1.64	0.011
Asian, γ_{50}	0.91	0.74	0.223	0.48	0.92	0.598
Hispanic, γ_{60}	2.33	0.38	**<0.001**	0.93	0.44	0.036
Multi-race, γ_{70}	1.67	0.92	0.071	1.48	0.94	0.115
White, γ_{80}	1.13	0.74	0.130	3.12	0.88	<0.001
Gifted status, γ_{90}	4.07	0.85	**<0.001**	3.16	0.90	<0.001
LEP status, γ_{100}	-2.31	0.46	**<0.001**	--	--	**--**
SWD status, γ_{110}	-1.26	0.53	**0.017**	-2.11	0.74	0.004

Random Effect	Variance Component	df	χ^2	*p*-value	Variance Component	df	χ^2	*p*-value
Intercept, u_0	10.05	64	504.97	**<0.001**	7.14	64	320.57	**<0.001**
Level-1 effect, γ_{ij}	23.51				28.58			
Reliability estimate								
Intercept	0.875				0.800			

Note. Bolded values are significant at .05.

[a] Fixed effects at level 2 is not reported here, since they are not of research interests in research question one; [b] Reference group is female; [c] Reference group is African American.

Table A.4 Results of HLMs for Research Question Two: 2010-Grade 5

	Mathematics			Reading		
Level-1 Fixed Effect[a]	Coef.	s.e.	*p*-value	Coef.	s.e.	*p*-value
Intercept, γ_{00}	354.02	0.36	**<0.001**	347.08	0.35	**<0.001**
2009_Grade4_Score, γ_{20}	0.55	0.02	**<0.001**	0.41	0.03	**<0.001**
2008_Grade3_Score, γ_{20}	0.27	0.02	**<0.001**	0.29	0.03	**<0.001**
Student Race/Ethnicity[c]						
American Indian, γ_{40}	--	--	**--**	0.11	1.91	0.954
Asian, γ_{50}	--	--	**--**	0.62	0.82	0.451
Hispanic, γ_{60}	--	--	**--**	0.13	0.41	0.755
Multi-race, γ_{70}	--	--	**--**	1.49	0.82	0.068
White, γ_{80}	--	--	**--**	2.20	0.50	**<0.001**
Gifted status, γ_{90}	1.59	0.71	**0.025**	--	--	**--**

Random Effect	Variance Component	df	χ^2	*p*-value	Variance Component	df	χ^2	*p*-value
Intercept, u_0	6.55	58	430.65	**<0.001**	5.06	59	304.31	**<0.001**
Level-1 effect, γ_{ij}	17.54				20.13			
Reliability estimate								
Intercept	0.853				0.796			

Note. Bolded values are significant at .05.

[a] Fixed effects at level 2 is not reported here, since they are not of research interests in research question one; [b] Reference group is female; [c] Reference group is African American.

Table A.4 Results of HLMs for Research Question Two: 2010-Grade 6

	Mathematics			Reading		
Level-1 Fixed Effect[a]	Coef.	s.e.	*p*-value	Coef.	s.e.	*p*-value
Intercept, γ_{00}	352.75	0.67	**<0.001**	349.46	0.47	**<0.001**
2009_Grade5_Score, γ_{20}	0.43	0.03	**<0.001**	0.37	0.03	**<0.001**
2008_Grade4_Score, γ_{20}	0.27	0.02	**<0.001**	0.29	0.03	**<0.001**
2007_Grade3_Score, γ_{20}	0.15	0.02	**<0.001**	0.13	0.02	**<0.001**
Student Race/Ethnicity[c]						
American Indian, γ_{40}	-2.17	1.79	0.227			
Asian, γ_{50}	1.49	0.46	**0.001**			
Hispanic, γ_{60}	0.68	0.29	**0.017**			
Multi-race, γ_{70}	0.56	0.78	0.475			
White, γ_{80}	-0.23	0.39	0.563			
Gifted status, γ_{90}	1.51	0.48	**0.002**	--	--	**--**
SWD status, γ_{110}	--	--	**--**	-1.97	0.53	**<0.001**

Random Effect	Variance Component	df	χ^2	*p*-value	Variance Component	df	χ^2	*p*-value
Intercept, u_0	16.04	32	1139.78	**<0.001**	7.05	34	539.66	**<0.001**
Level-1 effect, γ_{ij}	18.40				17.68			
Reliability estimate								
Intercept	0.965				0.899			

Note. Bolded values are significant at .05.

[a] Fixed effects at level 2 is not reported here, since they are not of research interests in research question one; [b] Reference group is female; [c] Reference group is African American.

Table A.4 Results of HLMs for Research Question Two: 2010-Grade 7

	Mathematics			Reading		
Level-1 Fixed Effect[a]	Coef.	s.e.	*p*-value	Coef.	s.e.	*p*-value
Intercept, γ_{00}	356.71	0.62	**<0.001**	353.86	0.70	**<0.001**
2009_Grade6_Score, γ_{20}	0.42	0.02	**<0.001**	0.40	0.03	**<0.001**
2008_Grade5_Score, γ_{20}	0.27	0.03	**<0.001**	0.24	0.02	**<0.001**
2007_Grade4_Score, γ_{20}	0.21	0.03	**<0.001**	0.18	0.02	**<0.001**
Gender[b], γ_{30}	-0.72	0.25	**0.004**	-0.62	0.22	**0.005**
Student Race/Ethnicity[c]						
American Indian, γ_{40}	--	--	**--**	-0.06	1.03	0.952
Asian, γ_{50}	--	--	**--**	-0.06	0.49	0.903
Hispanic, γ_{60}	--	--	**--**	0.88	0.29	**0.002**
Multi-race, γ_{70}	--	--	**--**	-0.02	0.70	0.975
White, γ_{80}	--	--	**--**	-0.71	0.55	0.200
LEP status, γ_{110}	--	--	**--**	-0.74	0.36	**0.042**

Random Effect	Variance Component	df	χ^2	*p*-value	Variance Component	df	χ^2	*p*-value
Intercept, u_0	12.83	31	913.78	**<0.001**	14.70	33	837.81	**<0.001**
Level-1 effect, γ_{ij}	18.94				17.38			
Reliability estimate								
Intercept	0.967				0.962			

Note. Bolded values are significant at .05.

[a] Fixed effects at level 2 is not reported here, since they are not of research interests in research question one; [b] Reference group is female; [c] Reference group is African American.

Table A.4 Results of HLMs for Research Question Two: 2010-Grade 8

	Mathematics			Reading		
Level-1 Fixed Effect[a]	Coef.	s.e.	*p*-value	Coef.	s.e.	*p*-value
Intercept, γ_{00}	358.57	0.39	<0.001	356.17	0.55	**<0.001**
2009_Grade7_Score, γ_{20}	0.46	0.02	<0.001	0.35	0.03	**<0.001**
2008_Grade6_Score, γ_{20}	0.20	0.03	<0.001	0.26	0.02	**<0.001**
2007_Grade5_Score, γ_{20}	0.12	0.02	<0.001	0.20	0.02	**<0.001**
Student Race/Ethnicity[c]						
American Indian, γ_{40}	-0.42	1.56	0.789	--	--	--
Asian, γ_{50}	1.77	0.64	0.006	--	--	--
Hispanic, γ_{60}	-0.04	0.36	0.907	--	--	**--**
Multi-race, γ_{70}	0.24	0.58	0.677	--	--	--
White, γ_{80}	0.54	0.47	0.248	--	--	--

Random Effect	Variance Component	df	χ^2	*p*-value	Variance Component	df	χ^2	*p*-value
Intercept, u_0	4.58	31	362.01	**<0.001**	9.47	31	831.13	**<0.001**
Level-1 effect, γ_{ij}	15.90				15.26			
Reliability estimate								
Intercept	0.905				0.960			

Note. Bolded values are significant at .05.

[a] Fixed effects at level 2 is not reported here, since they are not of research interests in research question one; [b] Reference group is female; [c] Reference group is African American.

Appendix B. Full Model Equations

Model 1. Mathematics 2008 Grade 4

Level-1 Model (Student level)

$$Math\_2008\_G4\_Score_{ij} = \beta_{0j} + \beta_{1j}*(Math\_2007\_G3\_Score_{ij}) + \beta_{2j}*(AmericanIndian_{ij}) + \beta_{3j}*(Asian_{ij}) + \beta_{4j}*(Hispanic_{ij}) + \beta_{5j}*(Multi\text{-}race_{ij}) + \beta_{6j}*(White_{ij}) + + \beta_{7j}*(GIFTED_{ij}) + \beta_{8j}*(LEP_{ij}) + \beta_{9j}*(SWD_{ij}) + r_{ij}$$

Level-2 Model (Teacher level)

$$\beta_{0j} = \gamma_{00} + \gamma_{01}*(Class\_size_j) + u_{0j}$$

$$\beta_{1j} = \gamma_{10}$$

$$\beta_{2j} = \gamma_{20}$$

$$\beta_{3j} = \gamma_{30}$$

$$\beta_{4j} = \gamma_{40}$$

$$\beta_{5j} = \gamma_{50}$$

$$\beta_{6j} = \gamma_{60}$$

$$\beta_{7j} = \gamma_{70}$$

$$\beta_{8j} = \gamma_{80}$$

$$\beta_{9j} = \gamma_{90}$$

Model 2. Mathematics 2008 Grade 5

Level-1 Model (Student level)

$$Math\_2008\_G5\_Score_{ij} = \beta_{0j} + \beta_{1j}*(Math\_2007\_G4\_Score_{ij}) + \beta_{2j}*(AmericanIndian_{ij}) + \beta_{3j}*(Asian_{ij}) + \beta_{4j}*(Hispanic_{ij}) + \beta_{5j}*(Multi\text{-}race_{ij}) + \beta_{6j}*(White_{ij}) + \beta_{7j}*(GIFTED_{ij}) + \beta_{8j}*(SWD_{ij}) + r_{ij}$$

Level-2 Model (Teacher level)

$$\beta_{0j} = \gamma_{00} + \gamma_{01}*(Class\_size_j) + u_{0j}$$

$$\beta_{1j} = \gamma_{10}$$

$$\beta_{2j} = \gamma_{20}$$

$$\beta_{3j} = \gamma_{30}$$

$$\beta_{4j} = \gamma_{40}$$

$$\beta_{5j} = \gamma_{50}$$

$$\beta_{6j} = \gamma_{60}$$

$$\beta_{7j} = \gamma_{70}$$

$$\beta_{8j} = \gamma_{80}$$

Model 3. Mathematics 2008 Grade 6

Level-1 Model (Student level)

$$Math\_2008\_G6\_Score_{ij} = \beta_{0j} + \beta_{1j}*(Math\_2007\_G5\_Score_{ij}) + \beta_{2j}*(Gender_{ij}) + \beta_{3j}*(AmericanIndian_{ij}) + \beta_{4j}*(Asian_{ij}) + \beta_{5j}*(Hispanic_{ij}) + \beta_{6j}*(Multi\text{-}race_{ij}) + \beta_{7j}*(White_{ij}) + + \beta_{8j}*(GIFTED_{ij}) + \beta_{9j}*(LEP_{ij}) + \beta_{10j}*(SWD_{ij}) + r_{ij}$$

Level-2 Model (Teacher level)

$$\beta_{0j} = \gamma_{00} + \gamma_{01}*(Class\_size_j) + u_{0j}$$

$$\beta_{1j} = \gamma_{10}$$

$$\beta_{2j} = \gamma_{20}$$

$$\beta_{3j} = \gamma_{30}$$

$$\beta_{4j} = \gamma_{40}$$

$$\beta_{5j} = \gamma_{50}$$

$$\beta_{6j} = \gamma_{60}$$

$$\beta_{7j} = \gamma_{70}$$

$$\beta_{8j} = \gamma_{80}$$

$$\beta_{9j} = \gamma_{90}$$

$$\beta_{10j} = \gamma_{100}$$

Model 4. Mathematics 2008 Grade 7

Level-1 Model (Student level)

$$Math\_2008\_G7\_Score_{ij} = \beta_{0j} + \beta_{1j}*(Math\_2007\_G6\_Score_{ij}) + \beta_{2j}*(Gender_{ij}) + \beta_{3j}*(AmericanIndian_{ij}) + \beta_{4j}*(Asian_{ij}) + \beta_{5j}*(Hispanic_{ij}) + \beta_{6j}*(Multi\text{-}race_{ij}) + \beta_{7j}*(White_{ij}) + \beta_{8j}*(GIFTED_{ij}) + \beta_{9j}*(SWD_{ij}) + r_{ij}$$

Level-2 Model (Teacher level)

$$\beta_{0j} = \gamma_{00} + \gamma_{01}*(Class\_size_j) + u_{0j}$$

$$\beta_{1j} = \gamma_{10}$$

$$\beta_{2j} = \gamma_{20}$$

$$\beta_{3j} = \gamma_{30}$$

$$\beta_{4j} = \gamma_{40}$$

$$\beta_{5j} = \gamma_{50}$$

$$\beta_{6j} = \gamma_{60}$$

$$\beta_{7j} = \gamma_{70}$$

$$\beta_{8j} = \gamma_{80}$$

$$\beta_{9j} = \gamma_{90}$$

Model 5. Mathematics 2008 Grade 8

Level-1 Model (Student level)

$$Math\_2008\_G8\_Score_{ij} = \beta_{0j} + \beta_{1j}*(Math\_2007\_G7\_Score_{ij}) + \beta_{2j}*(Math\_2006\_G6\_Score_{ij}) + \beta_{3j}*(AmericanIndian_{ij}) + \beta_{4j}*(Asian_{ij}) + \beta_{5j}*(Hispanic_{ij}) + \beta_{6j}*(Multi\text{-}race_{ij}) + \beta_{7j}*(White_{ij}) + \beta_{8j}*(LEP_{ij}) + r_{ij}$$

Level-2 Model (Teacher level)

$$\beta_{0j} = \gamma_{00} + \gamma_{01}*(Class\_size_j) + u_{0j}$$

$$\beta_{1j} = \gamma_{10}$$

$$\beta_{2j} = \gamma_{20}$$

$$\beta_{3j} = \gamma_{30}$$

$$\beta_{4j} = \gamma_{40}$$

$$\beta_{5j} = \gamma_{50}$$

$$\beta_{6j} = \gamma_{60}$$

$$\beta_{7j} = \gamma_{70}$$

$$\beta_{8j} = \gamma_{80}$$

Model 6. Mathematics 2009 Grade 4

Level-1 Model (Student level)

$$Math\_2009\_G4\_Score_{ij} = \beta_{0j} + \beta_{1j}*(Math\_2008\_G3\_Score_{ij}) + \beta_{2j}*(AmericanIndian_{ij}) + \beta_{3j}*(Asian_{ij}) + \beta_{4j}*(Hispanic_{ij}) + \beta_{5j}*(Multi\text{-}race_{ij}) + \beta_{6j}*(White_{ij}) + \beta_{7j}*(GIFTED_{ij}) + \beta_{8j}*(SWD_{ij}) + r_{ij}$$

Level-2 Model (Teacher level)

$$\beta_{0j} = \gamma_{00} + \gamma_{01}*(Class\_size_j) + u_{0j}$$

$$\beta_{1j} = \gamma_{10}$$

$$\beta_{2j} = \gamma_{20}$$

$$\beta_{3j} = \gamma_{30}$$

$$\beta_{4j} = \gamma_{40}$$

$$\beta_{5j} = \gamma_{50}$$

$$\beta_{6j} = \gamma_{60}$$

$$\beta_{7j} = \gamma_{70}$$

$$\beta_{8j} = \gamma_{80}$$

Model 7. Mathematics 2009 Grade 5

Level-1 Model (Student level)

$$Math\_2009\_G5\_Score_{ij} = \beta_{0j} + \beta_{1j}*(Math\_2008\_G4\_Score_{ij}) + \beta_{2j}*(Math\_2007\_G3\_Score_{ij}) + \beta_{3j}*(AmericanIndian_{ij}) + \beta_{4j}*(Asian_{ij}) + \beta_{5j}*(Hispanic_{ij}) + \beta_{6j}*(Multi\text{-}race_{ij}) + \beta_{7j}*(White_{ij}) + \beta_{8j}*(SWD_{ij}) + r_{ij}$$

Level-2 Model (Teacher level)

$$\beta_{0j} = \gamma_{00} + \gamma_{01}*(Class\_size_j) + u_{0j}$$

$$\beta_{1j} = \gamma_{10}$$

$$\beta_{2j} = \gamma_{20}$$

$$\beta_{3j} = \gamma_{30}$$

$$\beta_{4j} = \gamma_{40}$$

$$\beta_{5j} = \gamma_{50}$$

$$\beta_{6j} = \gamma_{60}$$

$$\beta_{7j} = \gamma_{70}$$

$$\beta_{8j} = \gamma_{80}$$

Model 8. Mathematics 2009 Grade 6

Level-1 Model (Student level)

$$Math\_2009\_G6\_Score_{ij} = \beta_{0j} + \beta_{1j}*(Math\_2008\_G5\_Score_{ij}) + \beta_{2j}*(Math\_2007\_G4\_Score_{ij}) + \beta_{3j}*(AmericanIndian_{ij}) + \beta_{4j}*(Asian_{ij}) + \beta_{5j}*(Hispanic_{ij}) + \beta_{6j}*(Multi\text{-}race_{ij}) + \beta_{7j}*(White_{ij}) + \beta_{8j}*(GIFTED_{ij}) + \beta_{9j}*(SWD_{ij}) + r_{ij}$$

Level-2 Model (Teacher level)

$$\beta_{0j} = \gamma_{00} + \gamma_{01}*(Class\_size_j) + u_{0j}$$

$$\beta_{1j} = \gamma_{10}$$

$$\beta_{2j} = \gamma_{20}$$

$$\beta_{3j} = \gamma_{30}$$

$$\beta_{4j} = \gamma_{40}$$

$$\beta_{5j} = \gamma_{50}$$

$$\beta_{6j} = \gamma_{60}$$

$$\beta_{7j} = \gamma_{70}$$

$$\beta_{8j} = \gamma_{80}$$

$$\beta_{9j} = \gamma_{90}$$

Model 9. Mathematics 2009 Grade 7

Level-1 Model (Student level)

$$Math\_2009\_G7\_Score_{ij} = \beta_{0j} + \beta_{1j}*(Math\_2008\_G6\_Score_{ij}) + \beta_{2j}*(Math\_2007\_G5\_Score_{ij}) + \beta_{3j}*(AmericanIndian_{ij}) + \beta_{4j}*(Asian_{ij}) + \beta_{5j}*(Hispanic_{ij}) + \beta_{6j}*(Multi\text{-}race_{ij}) + \beta_{7j}*(White_{ij}) + \beta_{8j}*(GIFTED_{ij}) + \beta_{9j}*(LEP_{ij}) + \beta_{10j}*(SWD_{ij}) + r_{ij}$$

Level-2 Model (Teacher level)

$$\beta_{0j} = \gamma_{00} + \gamma_{01}*(Class\_size_j) + u_{0j}$$

$$\beta_{1j} = \gamma_{10}$$

$$\beta_{2j} = \gamma_{20}$$

$$\beta_{3j} = \gamma_{30}$$

$$\beta_{4j} = \gamma_{40}$$

$$\beta_{5j} = \gamma_{50}$$

$$\beta_{6j} = \gamma_{60}$$

$$\beta_{7j} = \gamma_{70}$$

$$\beta_{8j} = \gamma_{80}$$

$$\beta_{9j} = \gamma_{90}$$

$$\beta_{10j} = \gamma_{100}$$

Model 10. Mathematics 2009 Grade 8

Level-1 Model (Student level)

$$Math\_2009\_G8\_Score_{ij} = \beta_{0j} + \beta_{1j}*(Math\_2008\_G7\_Score_{ij}) + \beta_{2j}*(Math\_2007\_G6\_Score_{ij}) + \beta_{3j}*(Gender_{ij}) + \beta_{4j}*(AmericanIndian_{ij}) + \beta_{5j}*(Asian_{ij}) + \beta_{6j}*(Hispanic_{ij}) + \beta_{7j}*(Multi\text{-}race_{ij}) + \beta_{8j}*(White_{ij}) + \beta_{9j}*(GIFTED_{ij}) + \beta_{10j}*(LEP_{ij}) + \beta_{11j}*(SWD_{ij}) + r_{ij}$$

Level-2 Model (Teacher level)

$$\beta_{0j} = \gamma_{00} + \gamma_{01}*(Class\_size_j) + u_{0j}$$

$$\beta_{1j} = \gamma_{10}$$

$$\beta_{2j} = \gamma_{20}$$

$$\beta_{3j} = \gamma_{30}$$

$$\beta_{4j} = \gamma_{40}$$

$$\beta_{5j} = \gamma_{50} + \gamma_{51}*(Class\_size_j) + u_{5j}$$

$$\beta_{6j} = \gamma_{60}$$

$$\beta_{7j} = \gamma_{70}$$

$$\beta_{8j} = \gamma_{80}$$

$$\beta_{9j} = \gamma_{90}$$

$$\beta_{10j} = \gamma_{100}$$

$$\beta_{11j} = \gamma_{110}$$

Model 11. Mathematics 2010 Grade 4

Level-1 Model (Student level)

$$Math\_2010\_G4\_Score_{ij} = \beta_{0j} + \beta_{1j}*(Math\_2009\_G3\_Score_{ij}) + \beta_{2j}*(AmericanIndian_{ij}) + \beta_{3j}*(Asian_{ij}) + \beta_{4j}*(Hispanic_{ij}) + \beta_{5j}*(Multi\text{-}race_{ij}) + \beta_{6j}*(White_{ij}) + \beta_{7j}*(GIFTED_{ij}) + \beta_{8j}*(LEP_{ij}) + \beta_{9j}*(SWD_{ij}) + r_{ij}$$

Level-2 Model (Teacher level)

$$\beta_{0j} = \gamma_{00} + \gamma_{01}*(Class\_size_j) + u_{0j}$$

$$\beta_{1j} = \gamma_{10}$$

$$\beta_{2j} = \gamma_{20}$$

$$\beta_{3j} = \gamma_{30}$$

$$\beta_{4j} = \gamma_{40}$$

$$\beta_{5j} = \gamma_{50}$$

$$\beta_{6j} = \gamma_{60}$$

$$\beta_{7j} = \gamma_{70}$$

$$\beta_{8j} = \gamma_{80}$$

$$\beta_{9j} = \gamma_{90}$$

Model 12. Mathematics 2010 Grade 5

Level-1 Model (Student level)

$$Math\_2010\_G5\_Score_{ij} = \beta_{0j} + \beta_{1j}*(Math\_2009\_G4\_Score_{ij}) + \beta_{2j}*(Math\_2007\_G3\_Score_{ij}) + r_{ij}$$

Level-2 Model (Teacher level)

$$\beta_{0j} = \gamma_{00} + \gamma_{01}*(Class\_size_j) + u_{0j}$$

$$\beta_{1j} = \gamma_{10}$$

$$\beta_{2j} = \gamma_{20}$$

Model 13. Mathematics 2010 Grade 6

Level-1 Model (Student level)

$$Math\_2010\_G6\_Score_{ij} = \beta_{0j} + \beta_{1j}*(Math\_2009\_G5\_Score_{ij}) + \beta_{2j}*(Math\_2008\_G4\_Score_{ij}) + \beta_{3j}*(Math\_2007\_G3\_Score_{ij}) + \beta_{4j}*(AmericanIndian_{ij}) + \beta_{5j}*(Asian_{ij}) + \beta_{6j}*(Hispanic_{ij}) + \beta_{7j}*(Multi\text{-}race_{ij}) + \beta_{8j}*(White_{ij}) + \beta_{9j}*(GIFTED_{ij}) + r_{ij}$$

Level-2 Model (Teacher level)

$$\beta_{0j} = \gamma_{00} + \gamma_{01}*(Class\_size_j) + u_{0j}$$

$$\beta_{1j} = \gamma_{10}$$

$$\beta_{2j} = \gamma_{20}$$

$$\beta_{3j} = \gamma_{30}$$

$$\beta_{4j} = \gamma_{40}$$

$$\beta_{5j} = \gamma_{50}$$

$$\beta_{6j} = \gamma_{60}$$

$$\beta_{7j} = \gamma_{70}$$

$$\beta_{8j} = \gamma_{80}$$

$$\beta_{9j} = \gamma_{90}$$

Model 14. Mathematics 2010 Grade 7

Level-1 Model (Student level)

$$Math\_2010\_G7\_Score_{ij} = \beta_{0j} + \beta_{1j}*(Math\_2009\_G6\_Score_{ij}) + \beta_{2j}*(Math\_2008\_G5\_Score_{ij}) + \beta_{3j}*(Math\_2007\_G4\_Score_{ij}) + \beta_{4j}*(Gender_{ij}) + r_{ij}$$

Level-2 Model (Teacher level)

$$\beta_{0j} = \gamma_{00} + \gamma_{01}*(Class\_size_j) + u_{0j}$$

$$\beta_{1j} = \gamma_{10}$$

$$\beta_{2j} = \gamma_{20}$$

$$\beta_{3j} = \gamma_{30}$$

$$\beta_{4j} = \gamma_{40}$$

Model 15. Mathematics 2010 Grade 8

Level-1 Model (Student level)

$$Math\_2010\_G8\_Score_{ij} = \beta_{0j} + \beta_{1j}*(Math\_2009\_G7\_Score_{ij}) + \beta_{2j}*(Math\_2008\_G6\_Score_{ij}) + \beta_{3j}*(Math\_2007\_G5\_Score_{ij}) + \beta_{4j}*(AmericanIndian_{ij}) + \beta_{5j}*(Asian_{ij}) + \beta_{6j}*(Hispanic_{ij}) + \beta_{7j}*(Multi\text{-}race_{ij}) + \beta_{8j}*(White_{ij}) + r_{ij}$$

Level-2 Model (Teacher level)

$$\beta_{0j} = \gamma_{00} + \gamma_{01}*(Class\_size_j) + u_{0j}$$

$$\beta_{1j} = \gamma_{10}$$

$$\beta_{2j} = \gamma_{20}$$

$$\beta_{3j} = \gamma_{30}$$

$$\beta_{4j} = \gamma_{40}$$

$$\beta_{5j} = \gamma_{50}$$

$$\beta_{6j} = \gamma_{60}$$

$$\beta_{7j} = \gamma_{70}$$

$$\beta_{8j} = \gamma_{80}$$

Model 16. Reading 2008 Grade 4

Level-1 Model (Student level)

$$Reading\_2008\_G4\_Score_{ij} = \beta_{0j} + \beta_{1j}*(Reading\_2007\_G3\_Score_{ij}) + \beta_{2j}*(AmericanIndian_{ij}) + \beta_{3j}*(Asian_{ij}) + \beta_{4j}*(Hispanic_{ij}) + \beta_{5j}*(Multi\text{-}race_{ij}) + \beta_{6j}*(White_{ij}) + \beta_{7j}*(GIFTED_{ij}) + \beta_{8j}*(LEP_{ij}) + \beta_{9j}*(SWD_{ij}) + r_{ij}$$

Level-2 Model (Teacher level)

$$\beta_{0j} = \gamma_{00} + \gamma_{01}*(Class\_size_j) + u_{0j}$$

$$\beta_{1j} = \gamma_{10}$$

$$\beta_{2j} = \gamma_{20}$$

$$\beta_{3j} = \gamma_{30}$$

$$\beta_{4j} = \gamma_{40}$$

$$\beta_{5j} = \gamma_{50}$$

$$\beta_{6j} = \gamma_{60}$$

$$\beta_{7j} = \gamma_{70}$$

$$\beta_{8j} = \gamma_{80}$$

$$\beta_{9j} = \gamma_{90}$$

Model 17. Reading 2008 Grade 5

Level-1 Model (Student level)

$$Reading\_2008\_G5\_Score_{ij} = \beta_{0j} + \beta_{1j}*(Reading\_2007\_G4\_Score_{ij}) + \beta_{2j}*(AmericanIndian_{ij}) + \beta_{3j}*(Asian_{ij}) + \beta_{4j}*(Hispanic_{ij}) + \beta_{5j}*(Multi\text{-}race_{ij}) + \beta_{6j}*(White_{ij}) + \beta_{7j}*(LEP_{ij}) + \beta_{8j}*(SWD_{ij}) + r_{ij}$$

Level-2 Model (Teacher level)

$$\beta_{0j} = \gamma_{00} + \gamma_{01}*(Class\_size_j) + u_{0j}$$

$$\beta_{1j} = \gamma_{10}$$

$$\beta_{2j} = \gamma_{20}$$

$$\beta_{3j} = \gamma_{30}$$

$$\beta_{4j} = \gamma_{40}$$

$$\beta_{5j} = \gamma_{50}$$

$$\beta_{6j} = \gamma_{60}$$

$$\beta_{7j} = \gamma_{70}$$

$$\beta_{8j} = \gamma_{80}$$

Model 18. Reading 2008 Grade 6

Level-1 Model (Student level)

$$Reading\_2008\_G6\_Score_{ij} = \beta_{0j} + \beta_{1j}*(Reading\_2007\_G5\_Score_{ij}) + \beta_{2j}*(Gender_{ij}) + \beta_{3j}*(AmericanIndian_{ij}) + \beta_{4j}*(Asian_{ij}) + \beta_{5j}*(Hispanic_{ij}) + \beta_{6j}*(Multi\text{-}race_{ij}) + \beta_{7j}*(White_{ij}) + \beta_{8j}*(LEP_{ij}) + \beta_{9j}*(SWD_{ij}) + r_{ij}$$

Level-2 Model (Teacher level)

$$\beta_{0j} = \gamma_{00} + \gamma_{01}*(Class\_size_j) + u_{0j}$$

$$\beta_{1j} = \gamma_{10}$$

$$\beta_{2j} = \gamma_{20}$$

$$\beta_{3j} = \gamma_{30}$$

$$\beta_{4j} = \gamma_{40}$$

$$\beta_{5j} = \gamma_{50}$$

$$\beta_{6j} = \gamma_{60}$$

$$\beta_{7j} = \gamma_{70}$$

$$\beta_{8j} = \gamma_{80}$$

$$\beta_{9j} = \gamma_{90}$$

Model 19. Reading 2008 Grade 7

Level-1 Model (Student level)

$$Reading\_2008\_G7\_Score_{ij} = \beta_{0j} + \beta_{1j}*(Reading\_2007\_G6\_Score_{ij}) + \beta_{2j}*(AmericanIndian_{ij}) + \beta_{3j}*(Asian_{ij}) + \beta_{4j}*(Hispanic_{ij}) + \beta_{5j}*(Multi\text{-}race_{ij}) + \beta_{6j}*(White_{ij}) + \beta_{7j}*(GIFTED_{ij}) + \beta_{8j}*(LEP_{ij}) + \beta_{9j}*(SWD_{ij}) + r_{ij}$$

Level-2 Model (Teacher level)

$$\beta_{0j} = \gamma_{00} + \gamma_{01}*(Class\_size_j) + u_{0j}$$

$$\beta_{1j} = \gamma_{10}$$

$$\beta_{2j} = \gamma_{20}$$

$$\beta_{3j} = \gamma_{30}$$

$$\beta_{4j} = \gamma_{40}$$

$$\beta_{5j} = \gamma_{50}$$

$$\beta_{6j} = \gamma_{60}$$

$$\beta_{7j} = \gamma_{70}$$

$$\beta_{8j} = \gamma_{80}$$

$$\beta_{9j} = \gamma_{90} + \gamma_{91}*(Class\_size_j) + u_{9j}$$

Model 20. Reading 2008 Grade 8

Level-1 Model (Student level)

$$Reading\_2008\_G8\_Score_{ij} = \beta_{0j} + \beta_{1j}*(Reading\_2007\_G7\_Score_{ij}) + \beta_{2j}*(Reading\_2006\_G6\_Score_{ij}) + \beta_{3j}*(AmericanIndian_{ij}) + \beta_{4j}*(Asian_{ij}) + \beta_{5j}*(Hispanic_{ij}) + \beta_{6j}*(Multi\text{-}race_{ij}) + \beta_{7j}*(White_{ij}) + \beta_{8j}*(LEP_{ij}) + r_{ij}$$

Level-2 Model (Teacher level)

$$\beta_{0j} = \gamma_{00} + \gamma_{01}*(Class\_size_j) + u_{0j}$$

$$\beta_{1j} = \gamma_{10}$$

$$\beta_{2j} = \gamma_{20}$$

$$\beta_{3j} = \gamma_{30}$$

$$\beta_{4j} = \gamma_{40}$$

$$\beta_{5j} = \gamma_{50}$$

$$\beta_{6j} = \gamma_{60}$$

$$\beta_{7j} = \gamma_{70}$$

$$\beta_{8j} = \gamma_{80}$$

Model 21. Reading 2009 Grade 4

Level-1 Model (Student level)

$$Reading\_2009\_G4\_Score_{ij} = \beta_{0j} + \beta_{1j}*(Reading\_2008\_G3\_Score_{ij}) + \beta_{2j}*(AmericanIndian_{ij}) + \beta_{3j}*(Asian_{ij}) + \beta_{4j}*(Hispanic_{ij}) + \beta_{5j}*(Multi\text{-}race_{ij}) + \beta_{6j}*(White_{ij}) + \beta_{7j}*(LEP_{ij}) + \beta_{8j}*(GIFTED_{ij}) + \beta_{9j}*(SWD_{ij}) + r_{ij}$$

Level-2 Model (Teacher level)

$$\beta_{0j} = \gamma_{00} + \gamma_{01}*(Class\_size_j) + u_{0j}$$

$$\beta_{1j} = \gamma_{10}$$

$$\beta_{2j} = \gamma_{20}$$

$$\beta_{3j} = \gamma_{30}$$

$$\beta_{4j} = \gamma_{40}$$

$$\beta_{5j} = \gamma_{50}$$

$$\beta_{6j} = \gamma_{60}$$

$$\beta_{7j} = \gamma_{70}$$

$$\beta_{8j} = \gamma_{80}$$

$$\beta_{9j} = \gamma_{90}$$

Model 22. Reading 2009 Grade 5

Level-1 Model (Student level)

$$Reading\_2009\_G5\_Score_{ij} = \beta_{0j} + \beta_{1j}*(Reading\_2008\_G4\_Score_{ij}) + \beta_{2j}*(Reading\_2007\_G3\_Score_{ij}) + \beta_{3j}*(SWD_{ij}) + r_{ij}$$

Level-2 Model (Teacher level)

$$\beta_{0j} = \gamma_{00} + \gamma_{01}*(Class\_size_j) + u_{0j}$$

$$\beta_{1j} = \gamma_{10}$$

$$\beta_{2j} = \gamma_{20}$$

$$\beta_{3j} = \gamma_{30}$$

Model 23. Reading 2009 Grade 6

Level-1 Model (Student level)

$$Reading\_2009\_G6\_Score_{ij} = \beta_{0j} + \beta_{1j}*(Reading\_2008\_G5\_Score_{ij}) + \beta_{2j}*(Reading\_2007\_G4\_Score_{ij}) + \beta_{3j}*(AmericanIndian_{ij}) + \beta_{4j}*(Asian_{ij}) + \beta_{5j}*(Hispanic_{ij}) + \beta_{6j}*(Multi\text{-}race_{ij}) + \beta_{7j}*(White_{ij}) + r_{ij}$$

Level-2 Model (Teacher level)

$$\beta_{0j} = \gamma_{00} + \gamma_{01}*(Class\_size_j) + u_{0j}$$

$$\beta_{1j} = \gamma_{10}$$

$$\beta_{2j} = \gamma_{20}$$

$$\beta_{3j} = \gamma_{30}$$

$$\beta_{4j} = \gamma_{40}$$

$$\beta_{5j} = \gamma_{50}$$

$$\beta_{6j} = \gamma_{60}$$

$$\beta_{7j} = \gamma_{70}$$

Model 24. Reading 2009 Grade 7

Level-1 Model (Student level)

$$Reading\_2009\_G7\_Score_{ij} = \beta_{0j} + \beta_{1j}*(Reading\_2008\_G6\_Score_{ij}) + \beta_{2j}*(Reading\_2007\_G5\_Score_{ij}) + \beta_{3j}*(AmericanIndian_{ij}) + \beta_{4j}*(Asian_{ij}) + \beta_{5j}*(Hispanic_{ij}) + \beta_{6j}*(Multi\text{-}race_{ij}) + \beta_{7j}*(White_{ij}) + \beta_{8j}*(GIFTED_{ij}) + r_{ij}$$

Level-2 Model (Teacher level)

$$\beta_{0j} = \gamma_{00} + \gamma_{01}*(Class\_size_j) + u_{0j}$$

$$\beta_{1j} = \gamma_{10}$$

$$\beta_{2j} = \gamma_{20}$$

$$\beta_{3j} = \gamma_{30}$$

$$\beta_{4j} = \gamma_{40}$$

$$\beta_{5j} = \gamma_{50}$$

$$\beta_{6j} = \gamma_{60}$$

$$\beta_{7j} = \gamma_{70}$$

$$\beta_{8j} = \gamma_{80}$$

Model 25. Reading 2009 Grade 8

Level-1 Model (Student level)

$$Reading\_2009\_G8\_Score_{ij} = \beta_{0j} + \beta_{1j}*(Reading\_2008\_G7\_Score_{ij}) + \beta_{2j}*(Reading\_2007\_G6\_Score_{ij}) + \beta_{3j}*(Gender_{ij}) + \beta_{4j}*(AmericanIndian_{ij}) + \beta_{5j}*(Asian_{ij}) + \beta_{6j}*(Hispanic_{ij}) + \beta_{7j}*(Multi\text{-}race_{ij}) + \beta_{8j}*(White_{ij}) + \beta_{9j}*(LEP_{ij}) + \beta_{10j}*(GIFTED_{ij}) + r_{ij}$$

Level-2 Model (Teacher level)

$$\beta_{0j} = \gamma_{00} + \gamma_{01}*(Class\_size_j) + u_{0j}$$

$$\beta_{1j} = \gamma_{10}$$

$$\beta_{2j} = \gamma_{20}$$

$$\beta_{3j} = \gamma_{30}$$

$$\beta_{4j} = \gamma_{40}$$

$$\beta_{5j} = \gamma_{50}$$

$$\beta_{6j} = \gamma_{60}$$

$$\beta_{7j} = \gamma_{70}$$

$$\beta_{8j} = \gamma_{80}$$

$$\beta_{9j} = \gamma_{90} + \gamma_{91}*(Class\_size_j) + u_{9j}$$

$$\beta_{10j} = \gamma_{100}$$

Model 26. Reading 2010 Grade 4

Level-1 Model (Student level)

$$Reading\_2010\_G4\_Score_{ij} = \beta_{0j} + \beta_{1j}*(Reading\_2009\_G3\_Score_{ij}) + \beta_{2j}*(Gender_{ij}) + \beta_{3j}*(AmericanIndian_{ij}) + \beta_{4j}*(Asian_{ij}) + \beta_{5j}*(Hispanic_{ij}) + \beta_{6j}*(Multi\text{-}race_{ij}) + \beta_{7j}*(White_{ij}) + \beta_{8j}*(GIFTED_{ij}) + \beta_{9j}*(SWD_{ij}) + r_{ij}$$

Level-2 Model (Teacher level)

$$\beta_{0j} = \gamma_{00} + \gamma_{01}*(Class\_size_j) + u_{0j}$$

$$\beta_{1j} = \gamma_{10}$$

$$\beta_{2j} = \gamma_{20}$$

$$\beta_{3j} = \gamma_{30}$$

$$\beta_{4j} = \gamma_{40}$$

$$\beta_{5j} = \gamma_{50}$$

$$\beta_{6j} = \gamma_{60}$$

$$\beta_{7j} = \gamma_{70}$$

$$\beta_{8j} = \gamma_{80}$$

$$\beta_{9j} = \gamma_{90}$$

Model 27. Reading 2010 Grade 5

Level-1 Model (Student level)

$$Reading\_2010\_G5\_Score_{ij} = \beta_{0j} + \beta_{1j}*(Reading\_2009\_G4\_Score_{ij}) + \beta_{2j}*(Reading\_2007\_G3\_Score_{ij}) + \beta_{3j}*(AmericanIndian_{ij}) + \beta_{4j}*(Asian_{ij}) + \beta_{5j}*(Hispanic_{ij}) + \beta_{6j}*(Multi\text{-}race_{ij}) + \beta_{7j}*(White_{ij}) + r_{ij}$$

Level-2 Model (Teacher level)

$$\beta_{0j} = \gamma_{00} + \gamma_{01}*(Class\_size_j) + u_{0j}$$

$$\beta_{1j} = \gamma_{10}$$

$$\beta_{2j} = \gamma_{20}$$

$$\beta_{3j} = \gamma_{30}$$

$$\beta_{4j} = \gamma_{40}$$

$$\beta_{5j} = \gamma_{50}$$

$$\beta_{6j} = \gamma_{60}$$

$$\beta_{7j} = \gamma_{70}$$

Model 28. Reading 2010 Grade 6

Level-1 Model (Student level)

$$Reading\_2010\_G6\_Score_{ij} = \beta_{0j} + \beta_{1j}*(Reading\_2009\_G5\_Score_{ij}) + \beta_{2j}*(Reading\_2008\_G4\_Score_{ij}) + \beta_{3j}*(Reading\_2007\_G3\_Score_{ij}) + \beta_{4j}*(SWD_{ij}) + r_{ij}$$

Level-2 Model (Teacher level)

$$\beta_{0j} = \gamma_{00} + \gamma_{01}*(Class\_size_j) + u_{0j}$$

$$\beta_{1j} = \gamma_{10}$$

$$\beta_{2j} = \gamma_{20}$$

$$\beta_{3j} = \gamma_{30}$$

$$\beta_{4j} = \gamma_{40}$$

Model 29. Reading 2010 Grade 7

Level-1 Model (Student level)

$$Reading\_2010\_G7\_Score_{ij} = \beta_{0j} + \beta_{1j}*(Reading\_2009\_G6\_Score_{ij}) + \beta_{2j}*(Reading\_2008\_G5\_Score_{ij}) + \beta_{3j}*(Reading\_2007\_G4\_Score_{ij}) + \beta_{4j}*(Gender_{ij}) + r_{ij}$$

Level-2 Model (Teacher level)

$$\beta_{0j} = \gamma_{00} + \gamma_{01}*(Class\_size_j) + u_{0j}$$

$$\beta_{1j} = \gamma_{10}$$

$$\beta_{2j} = \gamma_{20}$$

$$\beta_{3j} = \gamma_{30}$$

$$\beta_{4j} = \gamma_{40}$$

Model 30. Reading 2010 Grade 8

Level-1 Model (Student level)

$$Reading\_2010\_G8\_Score_{ij} = \beta_{0j} + \beta_{1j}*(Reading\_2009\_G7\_Score_{ij}) + \beta_{2j}*(Reading\_2008\_G6\_Score_{ij}) + \beta_{3j}*(Reading\_2007\_G5\_Score_{ij}) + r_{ij}$$

Level-2 Model (Teacher level)

$$\beta_{0j} = \gamma_{00} + \gamma_{01}*(Class\_size_j) + u_{0j}$$

$$\beta_{1j} = \gamma_{10}$$

$$\beta_{2j} = \gamma_{20}$$

$$\beta_{3j} = \gamma_{30}$$