2-1-2014

How Much Value is Added by Value Added Models? An Analysis of Teachers’ Performance Over Time Using New York State Assessment Data

Mariana Ristea
Graduate Center, City University of New York

How does access to this work benefit you? Let us know!
Follow this and additional works at: http://academicworks.cuny.edu/gc_etds

Part of the Educational Psychology Commons

Recommended Citation
Ristea, Mariana, "How Much Value is Added by Value Added Models? An Analysis of Teachers’ Performance Over Time Using New York State Assessment Data" (2014). CUNY Academic Works.
http://academicworks.cuny.edu/gc_etds/1422

This Dissertation is brought to you by CUNY Academic Works. It has been accepted for inclusion in All Graduate Works by Year: Dissertations, Theses, and Capstone Projects by an authorized administrator of CUNY Academic Works. For more information, please contact deposit@gc.cuny.edu.
Dissertation

HOW MUCH VALUE IS ADDED BY VALUE-ADDED MODELS?
An Analysis of Teachers’ Performance Over Time Using New York State Assessment Data

by

Mariana Ristea

A dissertation submitted to the Graduate Faculty in Educational Psychology in partial fulfillment of the requirements for the degree of Doctor of Philosophy, The City University of New York

2014
This manuscript has been read and accepted for the
Graduate Faculty in Educational Psychology in satisfaction of the
dissertation requirement for the degree of Doctor of Philosophy.

Dr. David Rindskopf

Date: ____________________________

Chair of Examining Committee

Alpana Bhattacharya

Date: ____________________________

Executive Officer

Dr. David Rindskopf
Dr. Sophia Catsambis
Dr. Howard Everson

Supervisory Committee

THE CITY UNIVERSITY OF NEW YORK
Abstract

How Much Value is Added by Value Added Models

by

Mariana Ristea

Adviser: Professor David Rindskopf

There is a strong movement to evaluate teachers on the basis of students’ performance. To compare teachers fairly, as each may have a mixture of students with different abilities in a given subject area, one should account for variables reflective of students’ subject knowledge and background when entering a course. Most methods of control consist of highly sophisticated statistical models mostly difficult to explain to educators who are being evaluated using such methods. This research presents two value-added methods that could be replicated by using in-house resources and standardized student assessment data which are either continuous or ordinal. One method is simpler to implement if one’s goal is to evaluate teachers’ performance based on students’ assessments scores reported as ordinal measures. The second method is similar to a more typical value-added approach and uses hierarchical linear structures to determine a classification of teachers’ performance based on their students’ assessment scores reported as continuous measures. Teachers’ “value-added” in a given academic year is typically calculated using students’ longitudinal New York State assessment data, reported in both ordinal and continuous forms. Comparison of results obtained from both methods, along with their interpretations, are used to examine trade-offs between accuracy of methods and their ease of use and transparency. The code used is included for practitioners who may wish to replicate this value-added methodology. Suggestions related to educational policy and feasibility of implementation of methods are also discussed.
Table of Contents

I. Introduction ........................................................................................................... 1

II. Background
   II.A. What is value-added? ....................................................................................... 4
   II.B. Value-added in education .................................................................................. 5

III. Instruments, Measurement and Issues in Measurement ........................................... 9

IV. Growth and Value Added Models .......................................................................... 15
    Limitations of Value-Added Models ........................................................................ 43

V. Mathematics Assessment Data: Overview ................................................................. 45

VI. Questions ............................................................................................................... 49

VII. Methodology ......................................................................................................... 50

VIII. Software for Growth Analysis ............................................................................. 71

IX. Results ................................................................................................................. 72

X. Findings and Limitations ........................................................................................ 89

XI. Practical Significance of the Study and Further Research ...................................... 91
Appendix 1 – SPSS Sorting Feature and Format Change ................................. 95
Appendix 2 – Excel Procedure to Generate Counts for Cross Tables .................. 96
Appendix 3 - NYS Math Weights used in Method 1 Procedure ........................... 96
Appendix 4 - Bootstrap Procedure for 2- and 3- score estimates ............................ 98
References ........................................................................................................ 105
**List of Tables**

1. Quick Summary of Status and Growth Models ........................................ 16
2. Cohort, instruments and rounds for data collection for longitudinal analyses ... 45
3. NYS Performance indicators for standardized yearly math assessments ........ 47
4. Variables for Multi-level Analyses ......................................................... 48
5. Relation between Year 0 and Year 1 Scores for a 25-student Class .............. 54
6. Observed Student Score Counts by Year ................................................. 54
7. Weighted Averages by Year ................................................................. 55
8. Calculated Teacher Performance at the End of Year 1 .............................. 56
9. Adjusted Weighted Teacher Performance at the End of Year 1 .................... 57
10. Descriptive Statistics for Continuous Student Pre and Post Scores ............. 60
11. Correlation table among student level variable considered for HLM analyses ... 63
12. Summary of Multilevel Models under Consideration ............................... 65
13. 3-Score Category Means for Teachers with 95% and 84% Confidence Intervals in Ascending Order ................................................................. 73
14. 2-Score Category Means for Teachers with 95% and 84% Confidence

    Intervals in Ascending Order ............................................................ 76
15. Model 1 - Final estimation of fixed effects ............................................ 78
16. Model 1 - Final estimation of variance components .................................. 78
17. Model 2 - Final estimation of fixed effects ............................................ 79
18. Model 2 - Final estimation of variance components .................................. 79
19. Results of Pre-score only model with 84% confidence intervals ................. 81
20. Teachers’ 2- and 3- score Average Correlations ........................................ 83

21. Side-by-side Mean Comparisons for Teachers with Students in 2- and 3-score
categories ........................................................................................................... 84

22. Side-by-side Mean Comparisons for teachers with students in 2- and 3-score categories
with unadjusted end of year average included ................................................. 85
List of Diagrams

1. Evaluating New York Teachers, Perhaps the Numbers Do Lie .................................. 3
2. Chart of Confidence Intervals with Included Teachers’ Class Averages at the End of the Year and Overlapping 84% and 95% Confidence Intervals for Continuous Scores ................................................................................................................................. 74
3. Q-Q Plot to Check Normality for Students Continuous Scores .............................. 77
4. Teacher Ranking based on Pre-Score Model only for Continuous Scores .......... 82
5. Correlation between the adjusted and the unadjusted 3-score teacher averages ........ 86
6. Correlations between 3-scores, 2-scores and unadjusted averages for teachers’ scores ...... 87
7. Raw Means Diagram for Teacher Rankings with Scores in Two Categories .................. 88
8. Bootstrapped Means Diagram for Teacher Rankings with Scores in Two Categories ........ 88
Introduction

Various initiatives, such as The Race to the Top, together with our country’s decline in international rankings in students’ math and English performance, have brought our country’s educational system to the center of the nation’s attention. From the citizen reading the news section of New York Times to the eager politician running for election, most began questioning whether the way we do things in education is appropriate for preparing students efficiently for the demands of a highly technological and competitive global economy.

This created the context for state policy makers to rapidly adopt appropriate strategies to measure value-added for teachers and for administrators leading public schools. Local schools are researching ways to adopt local value-added models based on local measures in order to obtain feedback about how efficiently their own teachers are at teaching students. Since state policy makers have given districts the option to include local growth measures in teachers’ annual evaluations they created a need for adoption of local value-added models (New York State Education Department [NYSED], 2012). This initiative requires the use of statistical models simple enough to be transparent to the average educator and to be implemented with local educational expertise and technology, but not too simple to significantly affect the accuracy of the results obtained.

Over time, value-added models in education generated many local and nationwide controversies (Sanders, 2000) as they attempt to analyze one’s teaching by summarizing the important variables related to student achievement in one or in more complex equations. Some believe that any such attempt is an impossible task in education, while others see it as an
opportunity to obtain some concrete data about a process not easy to encapsulate in a mathematical model.

George Box said that “All models are wrong but some models are useful” to exemplify situations where statistical models are not perfect but may be approximately right. Others who attempt to portray the essence of this process in newspapers show very complex equations, scary to look at and difficult to decipher by non-statisticians, in an attempt to point out how opaque these models can be. In a fairly recent New York Times article, “Evaluating New York Teachers, Perhaps the Numbers Do Lie” a value-added equation for a given subject, grade and school year shown below is described as being “like one of those equations that in “Good Will Hunting” that only Matt Damon was capable of solving. The process appears transparent, but it is clear as mud, even for smart lay people like teachers, principals and — I hesitate to say this — journalists.” (Winerip, 2011). New York City teachers are given yearly evaluation scores based on the equation shown in diagram 1 below which most if not all perceive as being very difficult to understand by practitioners.
Therefore the need for simpler methods that would produce similar results to more complex methods such as the one depicted above becomes more necessary in the current educational arena where teachers and administrators are being evaluated based on what experts refer to as value-added.
II.A. **What is value-added?**

Braun, Chudowsky, and Koenig (2010) state that “value-added” was first used in manufacturing to calculate “the difference between the value of the output and the cost of the raw materials”. Thus in economics, value-added is defined as the “difference between the total sales revenue of an industry and the total cost of components, materials, and services purchased from other firms within a reporting period (usually one year).” (WebFinance, Inc., 2013). Consequently this concept was adopted and used in education to refer to the changes in test scores due to specific factors such as teacher quality. Most recently, many states in the United States are looking to adopt or are in the process of adopting new systems of evaluating teachers’ performance based on their value-added to their students’ education over one year.

Not long ago, schools began to become accountable for the amount of improvement as reflected in students’ test scores. Schools that failed to make adequate yearly progress suffered loss of federal funding through Title I or in some cases were closed down (Raudenbush, 2004). The need for improved (though complex) growth models, such as value-added models, became apparent in order to make more informed decisions about schools’ contribution to individual students’ progress (Hull, 2007).

Some argue that true growth for a certain individual during a predetermined time period is difficult to measure since students’ academic growth may vary in intensity over time intervals just as their biological growth does (Goldstein, 1999). Others stress the measurement issues associated with longitudinal growth (Martineau, 2006) to be at least as challenging as the biological ones are.
In actuality most or all growth models are not really growth models in that they cannot calculate one’s growth in the same fashion that we calculate height or weight growth on a standard scale. That is, scores in Grade 3, for example, are not on the same scale as scores from Grade 4. Yet, we refer to them as “growth models” since this is how they are sometimes referenced in the literature. More recently in New York we move more in the direction of similar scaled scores for adjacent grade levels.

II.B. Value-added in education

In recent years applications of multilevel models to the field of education have spread widely. Many education analysts have realized the difficulties of conducting true experimental studies to establish cause and effect as well as the limitations of repeated measure studies in a school setting due to structures inherent in the systems, such as students’ assignment to classes, teachers’ assignments, missing data from certain assessments or flexible timing and more relaxed curricula used to test students with special needs or the change to a more rigorous Common Core curriculum.

What is the best way to define growth? And once we define growth how do we ensure that we have measured it properly? Can this be used to calculate a measure of “value-added”?

One way to attempt to fairly assess individual growth is to compare current students’ performance in a subject to that of students with similar backgrounds and scores from prior years in the same subject area. Projecting where they should perform in comparison to where their peers with similar performance and background end up performing constitutes the basis of calculating a numerical form of value-added (Doran, 2003).
The methods used to estimate value added in education refer to “the relative contributions of specific teachers, schools, or programs to student test performance.” (Braun, Chudowsky, & Koenig, 2010) while accounting for “differences in prior achievement and (perhaps) other measured characteristics that students bring with them to school.” Thus the interpretation of value-added estimates according to Martineau (2006) may refer to either “the value units add to student gains on a simple construct”, “the value units add to student gains on a mix of constructs” or “on a grade-specific mix of constructs where the mix is defined by the representation of the various constructs in grade-specific assessments”. He recommends this interpretation for models where teachers are held accountable for “student growth on constructs defined by the curriculum and mirrored by the assessments” (Martineau, 2006) which also represent a good match for the methodologies proposed by our study.

In New York State, just as in other states, parents and school districts receive results of assessments measuring students’ performance at the end of a school year in a specific subject. School districts in states where measures of local growth or “value-added” components are part of the teachers’ or administrators’ end-of-year evaluations may need or want to adopt systems that produce “value-added” measures based on either summative or formative local assessment or based on state assessment data (NYSED, 2011). Measuring teachers’ change in performance or the “value” they add for a given class in a given year is a complex task typically addressed through multilevel modeling and advanced statistical procedures to account for individual variability around group averages (Goldstein, 2011). According to Raudenbush and Bryk (2002, p.161), “The development of hierarchical linear models has created a powerful set of techniques for research on individual change. When applied with valid measurements from a multiple-time-point design, these models afford an integrated approach for studying the structure and predictors
of individual growth”. However, most methods presented in the literature are fairly advanced and difficult to implement by school district personnel who are not highly specialized in the area of statistics and measurement or who lack access to advanced statistical software.

Depending on their available expertise, local entities can employ simpler or more complex statistical analyses to attempt to quantify how much value is “added” to individual students’ performance over time by classroom teachers, schools, or districts. The complexity of various statistical models and steep budgetary cuts make it difficult for schools to hire experts in value-added accountability systems or to purchase existing services from third parties. If sufficient staff development or simpler models should become available some districts may move in the direction of creating their own “value-added” or local growth systems. Our research seeks to provide schools with the tools necessary to develop their own capacity to create value-added models.

This study researches two methodologies for determining teachers’ value-added for a given academic year given students’ performance from the prior academic year. Further, it explains the necessary steps to follow for anyone interested in designing a local school or district “value-added” system consisting in ranking teachers’ performance based on students’ assessment results. The first method uses ordinal data from students’ state assessment scores to calculate conditional weighted averages for each teacher within an organization. The second method uses continuous measures as reflected by students’ yearly state scores together with a linear mixed level approach to rank teachers based on the error of their class intercept also known as their class average after accounting for student related factors. The findings are followed by a brief analysis of consistency of results from both methods and suggestions about their application to
educational policy. Throughout the course of this study we will make reference to instruments, measurements and to other statistics of interest pertaining to either growth or to value-added.

A review of several growth and value-added models is incorporated for the purpose of giving the reader a brief introduction to such models.
III. Instruments, Measurements and Issues in Measurement

Instruments

Existing value-added methods used to determine teachers’ value-added and in turn their rating for a set time period are typically based on students’ results on state standardized exams historically shown to be valid and reliable. Standardization in this context refers to the procedures followed to prepare and administer the examinations. Current NYS grades 3-8 state assessments were adopted in 2005-2006, when public schools administered them to students. The tests were since administered to all students enrolled in grades 3 through 8 once a year, in March of each grade level until 2008-2009 when the month of the administration changed to April of each year, and the content indicators tested increased in number. All such standardized exams include items reflecting the New York State curriculum adopted in 2005. During the time period of our study all students took the assessment exam during the same testing period with assessment guidelines regulated by NYS and distributed to all proctors prior to the administration of each administration. The period of time selected for our study was also determined based on NYSED consistency of the structure of their testing instrument and of fairly consistent scoring guidelines of students’ state assessments. In this study we are using the students’ NYS math assessment scores from two consecutive years as the basis for our research.

The tests consist of multiple-choice and open-ended response items administered to students over a period of two or three-days depending on the student’s grade level. More information about the purpose, design and development, validity, administration and scoring of the test and data collection, data analysis, IRT scaling and equating of items as well as test reliability and standard error of measurement can be found in the Technical Report of the New
York State Testing Program (NYSED, 2008) or in literature related to measurement (Crocker & Algina, 1986).

For the purpose of a value-added analysis we note that, while there is an underlying conceptual and instructional flow among all New York State mathematics content indicators for grades 3 to 8, the scale scores are not vertically aligned across grade levels. Yet they can be compared for each grade level across cohorts. Kolen’s study (as cited in Braun, 2004) states “..The process by which scores on such a sequence of tests are placed on a common scale is called vertical scaling.” However, adjacent grade levels have similar process curriculum indicators and fairly similar content indicators which would make the measurement of growth between such time periods more plausible (NYSED, 2005). In our study, we selected to use the NYS math scores since there appears to be a very close relationship between the mathematical content described by the NYS math content standards presented at adjacent elementary grades. Thus this would lead to a better accuracy of growth.

Measurements

One of the greatest strengths of the students’ New York State assessment scores used in our study is that they are reported as both ordinal and continuous measures allowing for design of methodologies using both types of measurements. The ordinal measures also known as performance levels were derived from the continuous ones following a standard-setting procedure explained in the technical literature published following the administration of the assessments (http://www.p12.nysed.gov/assessment/pub/2007/math-sstr-07.pdf).

In modeling growth or value-added we encounter the terms of fixed and random effects. Technically speaking, in education research fixed effects could be generalized to the conditions
found in the experiment or observational study while random effects are those which can be
generalized to other settings if there are significant similarities between our students and the
students from those settings. It is also important to note the difference between fixed and
random effects as well as the difference between fixed and random variables or between fixed
and random coefficients. Fixed variables are those that remain unchanged over time while
random variables may change with time. Fixed and random coefficients are the equivalent of
fixed or random slopes which may or may not need to be further explained through additional
regressions in multilevel models.

Issues in Measurement

Measuring value-added is a fairly complex task due both to its conceptualization to a
given context and also to the need of selecting appropriate scaling and measurements with the
fewest limiting factors to statistical analyses. Perhaps an appropriate amount of value-added for a
specific individual can never be established as it is impossible to determine what would represent
an appropriate rate of growth at a given age for a set period of time or to account for all possible
variables. This could become a limitation for appropriately determining teachers’ value added
based on their students’ data.

The value-added approach was preceded by the use of students’ standardized assessment
scores used in accountability systems to track Adequate Yearly Progress (AYP) and also to
determine achievement gaps at the state level (U.S. Department of Education [USDOE], 2002;
Doran, 2003). This approach did not control for students’ backgrounds or for classroom-level
variables when tracking longitudinal growth. As practitioners began to focus more on adopting
value-added models for measuring students’ and teachers’ progress over time so did the
importance for more accurate longitudinal measurements grew in significance. “…VAM [Value-Added Model] estimates are sensitive to the way in which achievement is measured, including the content of the tests and the methods used to put the results from successive grades onto a common scale.” (McCaffrey, Lockwood, Koretz, & Hamilton, 2003). Aside from some of the shortcomings already mentioned in the measurement and instruments sections, Braun (2004) explains that “Typically, in a given subject, tests administered in successive grades vary in content and emphasis. The process by which scores on such a sequence of tests are placed on a common scale is called vertical scaling (Kolen, 2003)”. Yet, even if vertical scaling is implemented we should be aware that the different interval scales may lead to different estimated teachers’ effects (Braun, 2004). Unfortunately, investigation and discussion of the issue raised by the use of VAM in education has not been properly addressed in the methodological research where “much of the discussion remains unpublished, and the practical import of these concerns when VAM is applied to student achievement remains highly unclarified.” (McCaffrey et al., 2003). More recently, however, teams of experts have made a concerted effort to review and discuss a number of methodological challenges of VAMs, such as, measurement issues addressing test alignment to subject area standard, measurement error and vertical linking of tests, scaling (Doran & Fleischman, 2005) and models of learning along with analytic issues addressing complexity versus transparency of VAMs the quality of data used where missing data may impact on VA calculations and bias and precision of stability (Braun et al., 2010). Experts in VA warned us that “Small sample sizes are a particular problem when estimating teachers’ effects, because teachers often have only a relatively small number of students in a given year” (Braun et al., 2010) which may be compounded by other sources of variation such as “school leadership, peer effects, and student mobility” (Braun et al., 2010) among others. For certain
models, if the test scores between different grades are not vertically linked or measured on a “common scale so that students’ scores from different grades can be compared directly” (Braun et al., 2010), and if they are not measured using interval scales where a 1-point gain has the same significance in any two grades in the analysis then the value-added produced may not be valid. (Braun et al., 2010, Lewis, 2001; Linn 2001; Sanders & Horn, 1994) Thus, measurement instruments of similar difficulty levels and measuring the same construct at different points on the continuum of the subject knowledge would be ideal. Further, such test scores are subject to measurement errors similarly to any other measurement (Ladd & Walsh, 2002).

At times, when tests assess the same construct through items that have different content or difficulty levels and are used to measure students at different developmental points or grade levels it becomes more difficult to establish vertical scaling (Linn, 1993). Martineau (2006) claims that “psychometricians tend to agree that scales spanning wide grade/developmental ranges also span wide content ranges and that scores cannot be considered exchangeable along the various portions of the scale”. Assessment measurements pertaining to two adjacent grade levels, closer in content, are better measures of student gains than those where grade levels are further apart. “With current technology, there are no vertical scales that can be validly used in high-stakes analyses for estimating value added to student growth in either grade-specific or student-tailored construct mixes – the two most desirable interpretations of value-added to student growth” (Martineau, 2006). Braun (2004) concurs with these results by saying that “Strictly speaking, the assumption of an interval scale for test scores cannot be justified. It is probably a workable approximation for one or, perhaps, two grades but it is difficult to defend for multiple grades.” Despite all the warnings against employing vertical scales across grades for more than a few academic years at the time, various researchers such as Hauser (2003)
developed vertical scales across subject areas and across grade levels, each subject being on a single scale ranging from grade 2 to grade 10 and then attempted to calibrate the difficulty of test items across all these grades by using various statistical models (Lord, 1980).

Value-added calculations offer only as much information as the learning standards that are incorporated, as reflected by the tests whose scores they are based on. Some may feel that grade-level learning standards are not well aligned to the “developmental pathways of learning” (Braun et al., 2010) across grades, and that improved standards are needed.
IV. Growth and Value-Added Models

Overview

Teachers’ measures of value-added have been and will continue to be a priority of schools, states and federal governments when considering allocation of funds, end of year reviews and other policies related to curriculum and instruction. The scope of growth measurement and further of value-added of teachers and of schools evolved from statistics pertaining to average performance of groups of students to measures concerning the individual and its progress over time. This was a result of a shift in educational policy from the earlier ESEA (Elementary and Secondary Education Act) in 1994, to the NCLB act in 2002 with the goal that all students be proficient in their state standards by 2013-2014 (Braun, 2004) and to the more recent accountability measures imposed by the Race to the Top in 2010-2011. This has already been adopted by many states looking for federal funds including by NYS.

Historically, growth models measuring students’ progress over time evolved from simpler to more complex models (Counsel of Chief State School Officers [CCSSO], 2005). A value-added model could be modeled as a specific type of growth model. Some of the more popular growth models are summarized in Table 1 below and then further discussed.
## Table 1

Quick Summary of Status and Growth Models

<table>
<thead>
<tr>
<th>Characteristics of Models</th>
<th>Status Models (e.g., Improvement Models)</th>
<th>Growth Models (e.g., Simple Growth)</th>
<th>Value-Added Models</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Objective</strong></td>
<td>To evaluate schools based on their performance at one point in time</td>
<td>To evaluate schools based on difference in performance between two points in time</td>
<td>To evaluate schools based on how their students’ average performance differ from their expected performance</td>
</tr>
<tr>
<td><strong>Data requirement</strong></td>
<td>One year</td>
<td>Two years or two data points at possible equal time intervals</td>
<td>At least two years’ worth of data but preferably three</td>
</tr>
<tr>
<td><strong>Assesses individual students’ growth over time</strong></td>
<td>No</td>
<td>Yes, computed by taking the difference, ( \text{Year}_{T+1} - \text{Year}_T ) or by growth regressions</td>
<td>Statistically modeled through hierarchies or layers</td>
</tr>
<tr>
<td><strong>Type of scaled scores required</strong></td>
<td>Aligned to the state requirements and appropriate to compare different cohorts performance for same grade level</td>
<td>May require vertically aligned scaled scores from adjacent grade levels</td>
<td>Very likely requires vertical alignment among scaled scores from consecutive grade levels</td>
</tr>
<tr>
<td><strong>Pluses</strong></td>
<td>Fairly simple to calculate and to implement</td>
<td>Accounts for students’ change in scores over time and gives predictions, if necessary</td>
<td>Attempts to separate home from school effects</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Fairly simple to implement;</td>
<td>Accounts for students’ change in scores over time while controlling for school and non-school related factors</td>
</tr>
<tr>
<td><strong>Limitations</strong></td>
<td>Favors high performing schools;</td>
<td>Less likely to confound school and students’ effects on students’ performance but may still confound them</td>
<td>Tries to separate school and students’ effects on students’ performance</td>
</tr>
<tr>
<td></td>
<td>Results may confound school and students’ effects on students’ performance Focuses on students who are on the edge or around the cutoff scores between proficient and not</td>
<td>Favors school growth regardless of the subgroups students are part of;</td>
<td>Produces more accurate results than predicted growth;</td>
</tr>
</tbody>
</table>
Status Models (CCSSO, 2005) compare students “proficiency level by group or by subgroup, as it appears at one point in time, with an external target established by the state, usually expressed as a percentage of students meeting that target goal.” For instance, according to the NCLB act, the AMO (the annual measurable objective) consisting of a state-established formula to assess schools based on students’ test scores is the target to be accomplished in order to make AYP (Annual Yearly Progress). Improvement models are examples of status models aiming to calculate the difference between two consecutive cohorts’ student average proficiency performance enrolled in the same grade to determine whether on average a cohort does better than the prior one. Note that this model does not focus on individual growth nor does it compare the average performance of the same cohort in consecutive years. It only helps to identify if, on average, a larger percentage of students become proficient in a certain subject area of a given grade level in a year as it compares to the percentage of students proficient in the same subject area, of the same grade level of the prior year’s cohort also known as “Safe Harbor” in NCLB (U.S. Department of Education [USDOE], 2006; USDOE, 2008). Status models could be unconditional when using an unadjusted school performance, or conditional when the school performance is adjusted for out-of-school factors (CCSSO, 2005)

Performance Index Models are also used to measure how much “growth” is made by a certain cohort of students in a certain year when measuring achievement in students’ proficiency on a set assessment. For instance, let us assume that students can fall under any one of the four score categories: 1, 2, 3, or 4 based on their proficiency level on a math achievement exam, where 4 is highly proficient, 3 is proficient, 2 is somewhat proficient and 1 is not proficient. If we award 100 points for students scoring at level 4, 66 points for those achieving at level 3, 33 points for those achieving at level 1 and 0 points for those achieving at level 1, we can calculate
a performance index for the school for that cohort by using a weighted average. If there is an increase in indices between two consecutive years, schools are said to have made progress by moving more students toward becoming proficient or highly proficient. So, even if Johnny did not achieve proficiency in math this year or performed at or above grade level, (http://www.p12.nysed.gov/assessment/pub/2007/math-sstr-07.pdf) the fact that he moved from a level 1 (below standard or proficiency) score to a level 2 (meets basic proficiency) score is taken into account when computing the performance index model for a group of students.

As mentioned in the Guide for Informed Decision Making, published by the Center for Public Education (Hull, 2007), the computation for a Performance Index Model is similar to the computation of a GPA (Grade Point Average). As of 2006, 12 states including New York State adopted this model.

Simple Growth Models are the first to rely on individual growth when calculating student and school growth from year to year. Mathematically, these models measure by how many points a student’s scaled score has changed between two consecutive years to determine whether there is a positive, negative or zero difference or growth. Therefore we need at least two data points, but more data points will result in a better indication of students’ average growth over time. Yet, while this model is measuring some aspects of change in the individual’s student’s score it does not truly estimate appropriate growth for each child from year to year. This model should be accompanied by appropriate growth guidelines for each grade level or category of performance and also should be in line with the policy goals put forth by the respective state or district. In the case of vertical scales used for consecutive years’ assessments we need to question if a growth of 10 points for instance, on any given scale, means the same thing for a third grader as it does for a six grader. Yet, determining what an appropriate amount of growth would be, whether the
assessments are vertically aligned as well as other considerations about correlating content on assessments to state standards, should always be considered before fully implementing such models.

These models rely on calculating a school’s average performance from students’ average performance as measured by their scaled scores on standardized exams in a given year and then comparing them to the same students’ average performance from the prior school year to determine whether the school achieved growth. Defining the meaning and value of annual growth further plays an important role in evaluating students’ performance over time, and this is left up to current educational laws and state policy makers. NYSED currently uses several such growth models to evaluate yearly growth and performance for students, teachers and administrators attending or working in public school settings (NYSED, 2013).

Growth to Proficiency Models, also known as Growth to Standards models, analyze students’ growth in the context of becoming proficient over time at the established state standards for a given state. They are more popular and widely used at the state level than the models presented above due to the proficiency requirements outlined in the No Child Left Behind law. The time frame for achieving proficiency varies from state to state from three to four years to the end of high school (CCSSO, 2005).

Beginning in the 1990s, and more recently in 2002, the reauthorization of the ESEA and later the authorization of the No Child Left Behind Law (NCLB) led to the federal government requiring that students in all states meet the states’ proficiency standards by 2013-2014. NCLB held states and school districts accountable for students’ attainment of each state’s standards as reflected through AYP.
AYP is estimated by calculating the proportion of all students in a given grade meeting standards and then by comparing this result with prior year’s performance of the percentage of students on the same standards (Educational Testing Service, 2005). For example, if a third grader is on track to becoming proficient by the end of 5th grade, where the grade proficiency score is 500, and he is currently at a score of 400, he has two years to grow at least 100 points to reach proficiency. This model is currently used in the NCLB law to determine whether a school has made adequate growth from year to year or AYP. In NCLB, schools are given credit not only for students who reach proficiency but also for students who make the expected annual growth, both categories helping to determine if the overall percentage of students who either made individual growth or who became proficient exceeded the designated goal for the school, also known as AMO, Annual Measurable Objective.

The same type of AYP calculations are conducted for specific subgroups (e.g., students with disabilities, students with limited English proficiency, etc.). The quantitative component introduced by this system of accountability has some limitations such as, not accounting for the level of subject proficiency at which students enter a grade level, failing to distinguish between the amounts of instruction that students receive in school versus what they receive at home, the neighborhood effect as well as for other teacher and school effects that may vary from year to year and cannot be accounted for (Strand, 1997). These methods also fell short in accounting for students’ incoming level of knowledge when entering a certain grade, which can easily become a confounding variable with the quality of instruction students’ receive during that school year (Tekwe, Carter, Ma, Algina, Lucas, Roth, et al., 2004)
Given the evolution of new technologies, availability of more data and more sophisticated statistical methodologies, school systems and policy makers began steering away from group accountability and focusing on individual accountability. Achieving proficiency through AYP at the cohort level did not prove to be enough as some students, despite meeting standards, did not appear to grow in performance from year to year in subjects such as math, English or science. Thus, statisticians proposed value-added models where individual students’ progress can be measured through longitudinal gain scores and adjusted for specific variables and for the initial level of knowledge with which students’ enter a school year.

Statistically, simpler growth models are based on ANCOVA analyses, where the outcome variable is represented by students’ post-test scores, while the explanatory variables include students’ scores on the pre-test, or at the baseline along with variables related to students’ family background or to their educational setting. Using the notation from Bryk and Weisberg (1976), an ANCOVA model would consist of the following two equations:

\[
Y_{2ij} = \mu + \alpha_j + \beta_0 Y_{1ij} + \sum_{k=1}^{k} \beta_k M_{kj} + E_{2ij}
\]  

(1)

where,

\(Y_{2ij}\) = the student’s score on the posttest;

\(\mu\) = the overall mean, school or class on posttest;

\(\alpha_j\) = the effect of a student being in a certain class or program;

\(Y_{1ij}\) = the score on pretest of student i enrolled in class j;

\(\beta_0\) = the effect of pretest score on outcome;
\( \beta_k \) = the effects of covariates on outcome;

\( M_{kij} \) = explanatory variable or covariate “k” for student i enrolled in class or attending program j;

\( E_{2ij} \) = the amount of random error, or effects unaccounted for which are also uncorrelated with the other variables in the equation;

Using rules of expected values we can then express the class or program effect as

\[
\alpha_j = \bar{Y}_{2*} - (\mu + \beta_0 \bar{Y}_{1*} + \sum_{k=1}^{k} \beta_k \bar{M}_{kij})
\]

(2)

where,

\( \alpha_j \) = the effect for being in class or in program j after accounting for the mean of various covariates included in the model;

\( \bar{Y}_{2*} \) = the mean of all subjects in a certain class or program,

\( \bar{Y}_{1*} \) = the mean of all pretest scores;

\( \bar{M}_{kij} \) = the mean of covariates included in the analysis

The model assumes that the relationship between the pre- and post-tests and the rest of the covariates is linear, the error term is independent from the covariates, pre-tests and post-tests are reliable and valid, the variables are normally distributed around their means and they represent what a student should know at that particular point in time.

Value-Added models are among the most complex growth models. They focus primarily on individual students’ growth over time typically measured by standardized assessments while trying to separate the contribution that various factors (family, school, instructional programs, etc.) have on student’s growth over a set period of time (Raudenbush & Bryk, 1986). Thus, value-added models predict individual students’ expected growth based on their prior
performance while controlling for certain characteristics related to category of performance or other factors (family income, gender, the neighborhood effect or others). Mathematically, one can look at value-added as the second derivative of students’ performance over time, or “the rate of change of the rate of changes” (Raudenbush, 2004, p.7) Given the layers of modeling in value-added, we can calculate the value added by teachers, by schools (Meyer, 1997) or by school districts based on students’ growth in their yearly achievement as measured by standardized assessments. Estimates of value-added are calculated as the difference between actual growth and predicted growth for a certain unit (e.g., students, school). Other studies define value-added for a group as the mean difference between that group’s observed post-test mean and the predicted mean outcome “on the basis of natural maturation” (Bryk & Weisberg, 1976). Therefore one may attempt to calculate the amount by which an individual or a group of individuals exceeded or met the predicted expected growth.

By controlling for variables that characterize a student as s/he enters a teacher’s class, these models help us come closer to measuring the effects that a certain grouping taught by a certain individual has over time on the student’s performance gain. Unlike the criterion-referenced indicator where AYP measures whether cohorts of students within their respective subgroups (ESL, special-education, ethnic groups, etc.) performed better on an outcome variables (e.g., math scores) than the same cohorts from the prior year, value-added allows for customization of growth at the student level. So the student is not compared on the measure of interest to an external indicator but rather to his or her own performance from prior years.

Further one may conclude that value-added can be modeled in more than one form. One methodology used to model “value-added” is through hierarchical linear models. According to Raudenbush and Bryk (2002), “The development of hierarchical linear models has created a
powerful set of techniques for research on individual change. When applied with valid measurements from a multiple-time-point design, these models afford an integrated approach for studying the structure and predictors of individual growth.” Given its powerful characteristics, value-added has gained an important status among growth models.

In order to discuss how different value-added models are being implemented it is important to understand the fundamental theory behind them since most are modeled through hierarchical linear models (Bryk & Weisberg, 1976; Philips & Adcock, 1996; Raudenbush & Bryk, 1986).

In education, we have information about students’ performance on standardized exams as they are typically administered every year. We also have information about students’ characteristics such as family income, gender, classification as a second language learner, and about teachers and schools’ characteristics. These can help explain the contribution that such factors may have on students’ growth over time. We assume that the effect of good teaching is to improve students’ growth rates between two points in time as compared to the average growth rate in the school or in the district, the state or the nation.

In 1976, Bryk and Weisberg used the set of equations shown below to attempt to calculate the value-added for a student between two testing periods before and after the implementation of a new program.

\[ Y_{1i} = \beta a_{1i} + D_{1i} + E_{1i} \]  
\[ Y_{2i} = \beta a_{2i} + D_{2i} + E_{2i} \]  

\[ Y_{1i} = \text{the score on pretest of student } i; \]
\[ Y_{2i} = \text{the score on posttest of student } i; \]
$a_i =$ the age of student $i$ at the time of testing;

$D_i =$ covariates, constant over time representing fixed effects of individuals;

$E_i =$ random component, varying across individuals and independent from both, $D_i$ and $a_i$;

$E(E_i) = 0$, and $\text{Var}(E_i) = \sigma_E^2$;

If, after the implementation of a new program, the student’s posttest is expected to change by the “value-added”, $\nu$, then the new posttest score equation becomes:

$$Y_{2i} = \beta a_{2i} + D_{2i} + E_{2i} + \nu; \quad (5)$$

An estimate of the value-added for an individual student, then, is:

$$Y_{2i} - Y_{1i} - \beta(a_{2i} - a_{1i}) = \nu + E_{2i} - E_{1i}; \quad (6)$$

And for a group of students, it would be:

$$V = \overline{Y}_{2*} - \overline{Y}_{1*} - \beta(\overline{a}_{2*} - \overline{a}_{1*}) = \nu + \overline{E}_{2*} - \overline{E}_{1*}, \quad (7)$$

where $E(V) = \nu$, and the terms with bars over them represent sample means;

This model, together with other earlier versions of value-added models, were developed to account for the lack of randomization in schools and also to try to adjust for pre-existing mean differences among groups to calculate the change in students’ scores after an intervention is put in place. This constitutes the basis for determining teacher or school value-added.

In a more recent paper, Raudenbush and Bryk (2002) developed an HLM model by using a set of multi-level equations as shown in our methods section, where the notations are the same as those used in their text. More recent models are based on similar logic and typically are built up from simpler unconditional equations to equations including covariates.
Different states have adopted value-added models, most of which are widely based on the one developed by William Sanders and implemented in Tennessee. Schools, districts and states may decide to use variations of the models mentioned above.

Other Value-Added and Growth Models in Use

LMEM (Layered Mixed Effects Model) is the basic model behind what was later introduced as the Tennessee Value-Added Assessment System (TVAAS) by Sanders and Horn (1994) and presented below. Among current LMEMs we note:

- Multivariate LMEMs are used to analyze students’ growth in more than one subject area at a time thus allowing for the use of multiple student outcomes versus a single outcome as in the model below.

- Univariate LMEMs are used to analyze students’ individual growth in one subject area at a time (McCaffrey, Lockwood, Koretz, & Hamilton, 2003).

SFEM (Simple Fixed Effects Models) are employed due to their simpler mathematical structure. They involve calculating the difference between school mean change score and the mean of means change for all schools in the district. In SEM, the population of schools under consideration is assumed to be fixed where in HLMM, the schools studied are assumed to be a sample from a larger population (Littell, Milliken, Stroup, & Wolfinger, 1996).

REACH (Rate of Expected Academic Change), used in California, is a growth to proficiency model where a statistical model determines the true growth rate that each student needs to achieve in a set period of time, or by the end of a school year, based on prior information and measurements about students’ performance.
The basic growth model for this system consists of mixed effects statistical model (Pinheiro & Bates, 2000; Searle, 1971; Raudenbush and Bryk, 2002) expressed as:

\[ Y_i = X_i \beta + Z_i \theta + \varepsilon_i, \tag{8} \]

where

\( Y_i \) = the response vector of students math or reading scores (with n x n dimension);

\( X_i \) = the design matrix with n x p dimension;

\( \beta \) =vector of fixed effects (p x n);

\( Z_i \) = the design matrix for random effects (dimension, n x q)

\( \theta \) =the vector of random effects (q x n);

\( \varepsilon_i \) = the within group error term (n x n);

Students’ performance in each subject (math and reading) is measured at various time points to obtain the data necessary for the multilayers used in the model (observations within students and students within schools)

The most basic form of this model is similar to the unconditional model presented by Raudenbush and Bryk (2002). The combined level equation is:

\[ Y_{ij} = \mu + \beta_0 t + \theta_{0j} + \theta_{ij} t + \delta_{0ij} + \delta_{ij} t + \varepsilon_{ij}, \tag{9} \]

where

\( Y_{ij} \) is the \( i^{th} \) student’s score in a subject test, enrolled in school \( j \) at time \( t \);

\( \mu \) = the grand mean for all students’ scores considered at time \( t \) in a certain subject and at a certain school;

\( \beta_0 \) = the main effect for time;

\( \theta_{0j} \), and \( \theta_{ij} \) = school level effects;
\( \delta_{0ij} \) and \( \delta_{ij} \) = student level effects;

\( \varepsilon_{ij} \) = within group residual.

By using this model, we can find the estimated true growth rate for any student \( i \) in school \( j \):

\[
ETGR_{ij} = \beta_0 + \theta_{ij} + \delta_{ij}
\]

(10)

In order to further define what the appropriate growth is for each student and to ensure that all students achieve proficiency within a certain amount of time, a REACH score for each child is calculated as:

\[
R \text{EACH}_i = \frac{\lambda_{ps} - y_{ij}}{T - \alpha} = \frac{(prof\_cutscore\_subj\_p\_school\_g) - (stud\_score)}{(highest\_grade\_school) - (grade\_stud\_enrolled\_in\_at\_time\_t)}
\]

If \( R \text{EACH}_ij \geq 1 \) then student \( i \) is likely to achieve proficiency in school \( j \) by the time he or she completes the highest grade in that school, assuming that the student grows according to the estimated rate of change every year.

If \( R \text{EACH}_ij < 1 \) then the student may not achieve proficiency by the end of the last grade in school \( j \) unless instruction is differentiated for that child.

The schools are further classified as making or not making appropriate growth based on the percentage of students who scored at or above the proficiency cut-off point (PAC) across all grades and all tests offered in the school, in conjunction with the estimated rate of growth for that school.

\[
ETGR_j = \beta_0 + \theta_{ij}, \text{ where } \beta \text{ and } \theta \text{ have the same meaning as shown in the student level equations discussed above.}
\]

Depending on the percentage of students who meet proficiency every year, schools are classified as outstanding, sustaining, improving or underperforming.

Some of the advantages of this model consist of vertical alignment of test scores to students’ growth measured on a continuous developmental scale, the use of REACH score for
other school related projects in order to better plan for units of study or the differentiation of
instruction and the contribution of the model to identifying students in need of intervention.

The Rowan, Correnti, and Miller (RCM) growth model (McCaffrey et al., 2003) consists
of four different models ranging from simpler to more complex. The purpose of these models is
to account for at least 4,000 students’ growth as they are nested within at least 300 classrooms
further nested within 120 schools. The data used come from two cohorts of students from the
Prospects study. The first cohort had students tested in grades 1, 2, and 3 while in the second
cohort the students were tested in grades 3, 4, 5, and 6. Each of the four models is fit to this data
by subject, cohort and grade to attempt to analyze the various sources of variance in students’
test scores.

The first model is a three-level nested ANOVA, with students nested within classes, and
classes nested within schools. The second model includes students’ prior year scores as a
covariate in addition to classroom and school effects or to other students’ covariates. Similarly to
the ANCOVA model presented by Bryk and Weisberg (1976) the RCM covariate adjustment
model for students’ prior year’s data as presented by McCaffrey et al. (2003) is:

\[ Y_{ijt} = \alpha + \beta Y_{ijt-1} + \gamma x_{ijt} + \eta_{it} + \theta_{ij} + E_{2ij}, \]

where

\( Y_{ijt} = \) the \( j^{th} \) student’s score on the math or reading tests enrolled in school \( i \) at time \( t \);

\( Y_{ijt-1} = \) the prior year’s score (at “t-1”) of student \( j \), enrolled in school \( i \);

\( x_{ijt} = \) explanatory variable or covariate “k” for student \( j \) enrolled in school \( i \), to account for
student’s characteristics;
\( E_{ijt} \) = the amount of random error, or effects unaccounted for which are also uncorrelated with the other variables in the equation;

\( \eta_{it} \) = the effect of school i when administering the test at time t;

\( \theta_{ij} \) = the teacher’s or the classroom effect for the student taking the test at time i;

The third model is a one-year gain score model, similar to the second model shown above. The student’s gain score is a linear function of his prior year’s score, student covariates and random error, and with random class and school effects included in the model.

The fourth model is similar to the Raudenbush and Bryk (2002) cross-classified model with the exception that is modeled through a quadratic function.

\[
Y_{ijt} = \alpha + \beta t + \delta t^2 + \alpha_i + \beta_i t + \alpha_{ij} + \beta_{ij} t + \gamma X_{ijt} + \theta_{ij} + \ldots + \theta_{ij} + E_{ijt}
\]  (12)

In this model, \( \alpha \) and \( \beta \) are random intercepts or slopes for classes when the index is “i” or for students when the index is “ij”, \( X_{ijt} \) accounts for student’s characteristics not necessarily constant over time, and \( \theta_{ij} \) are the effects of the class or teachers’ on students’ scores at various times. Note how these effects are additive where the effects of prior teachers on students’ achievement are kept in the model.

Some disadvantages of these models include: sampling of units of analysis not discussed in the model, handling analysis weights also not discussed, using ANOVA with missing data (73% of the students had at least one missing score) without explaining how missing data were handled and without making reference to using imputations to account for missing data, and others. Among some advantages of these models is the use of all cases, including those who have
missing data to produce better parameter estimates (less biased) through the cross-classified models.

CPSP (Chicago Public School Productivity) Model is a multilayer model capable of accounting for students’ initial status as well as for gains in between tests while controlling for students’ and teachers’ related variables. It is based on a testing system with vertically aligned test scores (Byrk, Thum, Easton, & Luppescu, 1998).

CRESST (Center for Research on Evaluation, Standards and Student Testing) predicts students’ future growth based on their original status (Choi, Seltzer, Herman, & Yamashiro, 2004) in an attempt to estimate both a distribution of students’ growth in a school as well as the school average growth over time. The model is complex, it can work with multiple test-scores for various cohorts and it can control for students’ and school level variables, but it requires very large data sets.

Value-Added Models for measuring growth also vary along with their limitations that experts in the field have discussed since the 1980s. Before discussing a few widely employed value-added models let us underline some of their advantages and disadvantages.

As advantages, HLM (hierarchical linear models) allow for nesting of data and for correlation between students in the same classroom: individuals in the same group are more similar than individuals in different groups, thus data cannot be independent due to nesting (which is an assumption of regular regression). Additionally, HLM allows for variability between clusters or classrooms in intercepts and slopes as a function of group characteristics (e.g., classroom size or a teacher educational philosophy) and it allows for unbalanced group sizes: classrooms can vary in size, with missing observations so that units with incomplete observations can still be included in analyses. For missing data we can use an EM (expectation-
maximization) algorithm of likelihood estimation function for finding parameter estimates. Multilevel modeling has the advantage of handling missing data in a better way than repeated measure models or other models do (McCaffrey et al., 2003). “All the common methods for salvaging information from cases with missing data typically make things worse: They introduce substantial bias, make the analysis more sensitive to departures from MCAR (missing completely at random), or yield standard error estimates that are incorrect (usually too low).” (Allison, 2001) In addition, multilevel modeling also has the advantage of handling more complex student-level error structures (Raudenbush & Bryk, 2002). HLM allows for testing of main effects in interaction within and between levels and for a flexible error structure between variables (especially important for longitudinal data analysis).

TVAAS (Tennessee Value Added Assessment System) developed by Sanders and his team entered the educational arena through the Tennessee Educational Improvement Act in 1992 (Kupermintz, 2003) as a result of a lawsuit brought by a small group of rural schools against the state about funding inequities among different schools. This model was used in Tennessee in 1993 (Braun, 2004) to produce value-added reports, yet research shows that the first TVAAS reports with respect to school effectiveness on students’ instruction was developed in late 1992 (Sanders, 1998; Sanders & Horn, 1998). Since then various forms of it have been adopted by other states in an effort to provide evidence about improvement in education at various levels (classroom, teachers, others) through the lenses of students’ achievement scores on standardized assessments over time.

In Tennessee, the scores used were a result of state standardized student testing using the TCAP (Tennessee Comprehensive Assessment Program). Its main output consists of learning gains but the model does not take into consideration SES variables for students or their incoming
knowledge level in a subject area, as a covariate, a decision still under debate in the literature. Certain studies claim that students’ performance in a given school is very much influenced by their prior experiences, family factors and other programs they may attend out of school (Raudenbush, 2004). To account for these variations, Sanders calculates the covariance between students’ test scores at different time points to produce a measure of value-added effects over time (Sanders & Horn, 1994).

This model contains proprietary computations not easily accessible to researchers (CCSSO, 2005) and it is fairly costly to implement. It is mainly based on students’ scores on annual assessments, administered in multiple subjects. The model uses a mixed model or multilayer approach in order to project or to predict a multivariate, longitudinal analysis for students’ data linked to state assessments (Sanders & Horn, 1998). The multivariate version of it is also available. The TVAAS model assumes that accounting for students’ background variables is not necessary given that they do not correlate with students’ gain scores and that in his model each student acts as its own control (Sanders & Horn, 1998). Aside from student-level variables, this model also does not account for the interaction between a school’s initial status and its growth over time. Furthermore, Sanders argues that the best predictor for student achievement is the quality of teaching that students experienced during each academic year (Sanders & Horn, 1998; Wright, Sanders, & Rivers, 2006; Wright, White, Sanders, & Rivers, 2010).

A brief mathematical description of Sanders’ model consisting of a multilevel mixed model with equations customizable to different subjects in different years (Ballou, Sanders, & Wright, 2004):

\[ y_{i}^{k} = b_{i}^{k} + u_{i}^{k} + e_{i}^{k} \]  
\[ y_{r+1}^{k+1} = b_{r+1}^{k+1} + u_{r+1}^{k+1} + e_{r+1}^{k+1} \]  

(13)  
(14)
where

\( y_{kt}^k \) = the student’s score on subject assessment administered in grade k, and in year t;

\( b_{kt}^k \) = the district overall average score, in grade k, year t, considered a fixed effect;

\( u_{kt}^k \) = the teacher effect to the student’s score in grade K, year t, considered a random effect;

\( e_{kt}^k \) = error term or unexplained variation in students score, from grade k, year t.

The next equation has the same variables with the same meanings but adapted for the next grade level, grade k+1, year t+1.

The goal of EVAAS is to predict or project students’ future or current scores based on their prior assessment data. The projection equation used to project individual students’ scores is (Wright, Sanders, & Rivers, 2006; Sanders et al., 1997):

\[
\text{Projected}_\text{Score} = M_y + b_1(X_1 - M_1) + b_2(X_2 - M_2) + \ldots = M_y + X_i^T b, \quad (15)
\]

Where, \( M_i \) = estimated mean scores for the outcome variable, or for the explanatory variables, \( X_i \);

\( b_i = C_{XX}^{-1}C_{XY} \), where C’s are covariance matrices that, when used with each student’s scores, can help project that child’s outcome, \( Y \);

\( C \) = the covariance matrix is obtained by using an in-school pooled covariance matrix, with a maximum likelihood estimation calculated with an EM algorithm applied to data centered around the school mean. This is useful when dealing with missing data.

\( M_y \) = the mean for an average school obtained by averaging the means of all schools by using the data from the most recent school year;

In Projected_Scores, the system uses the \( Y \)’s from the most recent years and the \( X \)’s from prior years to project \( Y \) values for students who do not have a \( Y \) in current year but have \( X \)’s
from prior years (Wright, Sanders, & Rivers, 2006). Therefore, the projection parameters consist of vectors of estimated means and of estimated covariances.

Using Raudenbush and Bryk (2002) notation, the individual students’ projections denoted by $Y_{it}$ are:

$$Y_{it} = (\beta_{oo} + r_{0i}) + (\beta_{oi} + r_{1i})t + \varepsilon_{it} = (\beta_{oo} + r_{0i}) + (r_{oi} + r_{1i}t + \varepsilon_{it}) = \mu_{i} + \delta_{it},$$  \hspace{1cm} (16)

where

$\mu_{i}$ reflects the means discussed above, and

$\delta_{it}$ reflects all errors, $r$’s and $\varepsilon$ from the different level equations, whose variances, $C_{i}$ for any student $i$, reflective of all the term variances in the equations above are given by,

$$C_{i} = \text{var}(\delta_{it}) = Z_{i}TZ_{i}^{T} + I\sigma^{2}$$

where,

$\sigma^{2} = \text{var}(\varepsilon_{it})$ is assumed the same for all times “t” and students “i”;

$T = \text{var}(\{ r_{0i}, r_{1i} \})$, is the same for all students, $i$, and

$Z_{i}$ is a column matrix consisting of a column of 1’s (fixed as the y-intercept) and a column of t’s;

The literature references that currently there are a number of EVAAS models in use and that “the results of these analyses, along with additional diagnostic information and querying capabilities, are made available via a secure web application” (Wright, White, Sanders, & Rivers, 2010).

The univariate response model (URM) proposed by Sanders consists of an ANCOVA, where the outcome $Y_{i}$ is the student’s math assessment score, and the independent continuous variables are the students’ scores from prior years, and the categorical predictors treated as random are class, teacher, school, or district level variables. This model requires at least three
prior data points for each student in order to predict more accurately students’ performance
during the current academic year in a subject area.

“..the first step in the URM [Univariate Response Model], is to obtain a projection for each student using whatever set of predictors that student has available. However, rather than projecting a future score using a student’s present and past scores, in the URM one is
“projecting” a student’s present score using their past scores.” (Wright, White, Sanders, &
Rivers, 2010). The studies further show that a projection score is a weighted-composite of students’ prior scores representing a summary of student’s prior achievements. The difference between two such composites or projection scores from prior and current years may be interpreted as value-added. Also, other variations of EVAAS use prior cohorts similar in performance to the one being evaluated to produce a prediction equation, which is then applied to the current year’s data to determine a difference further interpreted as value-added.

The metric used in the model is represented by the normal curve equivalents (NCE), which are made available by states or are calculated by the model. NCE are obtained by first calculating the students’ frequency distribution scores at the population level, then finding the associated cumulative frequency (number of students scoring at or below that score), and its respective percentage. Convert these percentages into percentile ranks, determining their z-scores. Sanders and his team suggest that “...NCEs are scaled so that they exactly match the percentile ranks at 1, 50, and 99. This is accomplished by multiplying each z-score by approximately 21.063 (the standard deviation off the NCE scale) and adding 50 (the mean on the NCE scale)” (Wright, White, Sanders, & Rivers, 2010). Unlike in a regular value-added model, individual students’ equations may have different sets of predictors (X’s) to accommodate for missing data.
Some of the differences and similarities between this model and the classic value-added model are represented in the ideas to follow (Wright, Sanders, & Rivers, 2006, p.5). As in value-added models using hierarchical structures, in EVAAS as well students are nested within classrooms, and within schools, districts or larger educational systems. EVAAS makes projections for individual students assuming for each student’s yearly growth the average rate of growth for the overall school and does not require vertically linked data, or data from the same subjects. Furthermore, EVAAS uses unstructured covariance, does not rely on or assume linear growth from time of measurement to time of measurement and the dependent and independent variables are not required to have the same scale as one another. Unlike other models, EVAAS estimated means are not linked functionally to follow a certain path over time.

Each student included in EVAAS analyses needs to have at least two data points but preferably three data points. A certain year’s cohort’s scores are being utilized to produce parameter estimates to also be used in projections for an incoming cohort of students who have not yet been tested in that same grade level.

Unlike EVAAS, an HLM model uses the same cohort’s scores to produce the parameter estimates and projections for that same cohort under analysis. TVAAS is used to determine students’ academic gains over a set period of time instead of discussing students’ absolute growth based on a target score (Sanders & Horn, 1998).

Strong points and other findings of TVAAS or EVAAS

The advantages of the model is that it uses all available scores or other such linear variables for each student for as many years as the districts can provide, and it handles missing data well. It also takes into consideration, students’ aptitude toward instruction and other external
environmental influences on students’ learning as an important factor in the teaching and learning progress (Corno et al., 2002; Sanders & Horn, 1998) by controlling for students’ assessment scores from prior years. This serves as a measure of their prior knowledge and readiness for instruction (Kupermintz, 2003). Simply put, TVAAS uses each child as its own control by taking into account his or her prior test scores when predicting a student’s likely future performance.

TVAAS uses random assignment of students to teachers with teachers’ teaching balanced classes. Yet, there are studies challenging the random distribution of students to classes since after a reanalysis of data from Sanders and Rivers (1996) the distribution of teachers’ performance appear to be correlated to students’ prior performance. (Kupermintz, 2003). It does not require that test scores be vertically aligned or that they originate from the same subject, yet it requires that explanatory variables be good predictors of the outcome. Depending on the state which implements it, the EVAAS models typically use students’ scores in math, English and in Science. They do not require a certain shape of growth curves over time but can accommodate projections for long periods into the future (Wright, Sanders, & Rivers, 2006).

These models handle well missing data, large data sets and when various models’ assumptions are violated, the EVAAS model seems to produce more robust projections than other linear models do. Some EVAAS models take into account students’ nonlinear growth patterns, unlike regular linear growth models that do not (Sanders, 1997).

The effectiveness of teachers is the main factor in explaining students’ academic growth, where class size and other SES-related explanatory variables were found as having a small to negligible effect on achievement outcomes (Wright et al., 1997; Sanders & Rivers, 1996). The teachers’ effects on students’ academic performance are additive and summative over time
However, when interpreting results from TVAAS one should be cautious about their accuracy and about their implementation in educational policy as there are teachers with less data whose effects would revert toward the system’s (school, district) mean or average, thus making it more difficult to differentiate well their effects from the average teacher’s performance in the system (Sanders, Saxton, & Horn, 1997; Kupermintz, 2003).

Similarly, the effects of teachers whose classes have a more transient population may also be pulled toward the system’s mean as compared to teachers’ effects for those teaching less transient students. Different teachers’ effects will be pulled toward their respective system’s average (Kupermintz, 2003). Not including students’ background variables (e.g., SES, race, other) may introduce bias in calculating the teachers’ effects on students’ scores. Studies have shown that the inclusion of such variables may contribute to a better attribution of students’ growth to such background variables, especially for students ranking at the top of their classes or for students attending high income schools (Bryk & Weisberg, 1976; Berk, 1988; Thum & Bryk, 1997).

Some of the more significant TVAAS findings show that less than a third of the teachers differ from their average colleague when looking at the differences in estimated teachers’ effects. Through simulations (Wright, Sanders, & Rivers, 2006) in most cases students’ growth appears to follow a non-linear trend rather than a linear trend. Therefore, when used on the same data set as a hierarchical linear growth model, if the assumptions of each model are not violated TVAAS produces projections for students’ scores that are fairly robust (Wright, Sanders, & Rivers, 2006).
Rivkin, Hanushek, and Kain (RHK) is a value-added model geared toward trying to separate the effects of teachers on students’ growth from the effect of other potential growth factors. The sample consists of about half a million students enrolled in 2,156 elementary schools with multi-year data. The method encompasses multi-step differential gain scores between consecutive years of student data in order to first separate such gains from class, school or other factors’ effects on students’ achievement but not on their growth over time.

The set of equations used in the model are (McCaffrey, et al., 2003, pp. 31-35):

\[ d_{ijg} = y_{ijg} - y_{ijg-1}, \quad (17) \]

where \( y_{ijg} \) is the student’s \( j \) score, enrolled in cohort \( i \) and in grade \( g \), and “\( d \)” represents the differences in scores between any two consecutive years;

\[ a_{ij} = d_{ijg+1} - d_{ijg}, \quad (18) \]

where “\( a \)” represents the difference of differences, as shown above.

The average of all \( a \)’s will produce, \( A_i \), or \( A_i = T_{ig+1} - T_{ig} + e_i \), where \( T_{ig} \) represents the average effect of teachers teaching students enrolled in grade \( g \).

\[ D = (A_{i+1} - A_i)^2, \quad (19) \]

which appears to be smaller when teachers overlap between grades and larger when they do not overlap as much.

In order to estimate teachers’ effects, \( D \) is then modeled as an outcome of a linear regression function where the teachers’ turnover rate is an explanatory variable along with other school related factors.

Among some of the disadvantages found in this method we note the inclusion of those students’ scores who remain within the analyzed schools for at least three consecutive years, the
exclusion of certain student-level covariates from the model which may lead to bias when explaining the outcome, and the fact that its results hold better on larger samples but not as well on smaller ones, due to factors related to teachers’ mobility or assignments over different grades.

The RAND Model, based on TVAAS, is a multivariate longitudinal mixed model which incorporates students’ and teachers’ related variables into the multilevel equations. It was developed by McCaffrey and his team, and unlike Sanders’ model it does not assume that teachers’ effects are time invariant. It is very similar to the RCM 4th model described above.

DVAAS (the Dallas Value-Added Accountability System) used in Dallas is based on a longitudinal methodology analysis developed by Webster and Mendro (1997). Its goal is to estimate the school effects on students’ scores and on their growth. This model unlike Sander’s model includes certain student level covariates (SES, gender, free and reduced-lunch, language-proficiency) in addition to using students as their own control (Doran & Izumi, 2004). It consists of a two-step process. First, students’ test scores are modeled through an OLS regression as functions of students’ level covariates, considered fairness factors (e.g., the amount of time that a student was enrolled in a school to be one academic year minus at most six weeks of school, others) which cannot be accounted for through the HLM stage-two model. The second equation in the model uses the residuals from the first equation as outcomes explained again in term of covariates in order to ensure comparability between members of a certain socio-economic group (gender, income, etc.) and the average performance of that group.

First stage: \( Y_i = \beta_0 + \beta_1 X_{i1} + \ldots + \beta_j X_{ij} + r_i \), where \( r_i \sim N(0, \sigma^2) \) \hspace{1cm} (20)

The \( Y_i \) represents an outcome, or student achievement or attendance regressed on the fairness variables, \( X \) (ethnicity, Limited-English proficiency, gender, free lunch status, census income, census poverty, and census college attendance) and on the first three variables’ first-
second-level interactions (Webster & Mendro, 1997, pp.4-7). The residuals are then standardized to a mean of zero and a standard deviation of 1, and after considering the interactions, the residuals for the 16 resulting subgroups are entered in the second HLM stage for analysis.

Second stage, hierarchical linear model equations (Webster & Mendro, 1997, p. 6):

Level 1: \[ Y_{ij} = \beta_0 + \beta_1 X_{ij} + \ldots + \beta_{kj} X_{kij} + r_{ij} \] (21)

Level 2: \[ \beta_{0j} = \gamma_{00} + \gamma_{01} W_j + u_{0j} \] (22)
\[ \beta_{1j} = \gamma_{10} + \gamma_{11} W_j + u_{1j} \] (23)

\[ \ldots \ldots \ldots \]
\[ \beta_{kj} = \gamma_{k0} + \gamma_{k1} W_j + u_{kj} \] (24)

Where, \( r_{ij} \sim N(0, \sigma^2) \), and \( u_{kj} \sim N(0, \tau_{ko}) \), for all \( k \).

In stage two, the first level equation is a student equation regressing residuals of student outcome variables (such as achievement) on residuals of prior achievement or other such considered variables. Thus, \( Y_{ij} \) represents residuals of student outcome variables calculated in stage one, \( X_{kij} \) represents residuals of prior achievement or others, and \( \gamma \) stands for school level variables (students’ mobility, percent minority, or percent on free lunch, others). In order to solve the HLM equations the models use a Bayesian estimation to obtain an empirical Bayes estimate for each school, used as a measure of school effectiveness and a student residual, used to estimate teacher effectiveness.

One of the main disadvantages of this system is that DVAAS uses an OLS regression as an entry level equation, which does not necessarily account for the nesting of students in the district and therefore violates some of the OLS statistical assumptions.
HLMM, an unconditional model with an intercept only, or HLMM with covariates based on demographics and on the level of knowledge that students enter the year with is especially useful in studies attempting to design value-added models based on data coming from standardized assessments that are not necessarily vertically aligned (Noell, Gansle, Patt, & Schafer, 2009). It is best described through the set of multi-level equations used by Raudenbush and Bryk (2002) and presented at the beginning of this section.

Limitations of Value-Added Models

The limitations of various value-added models occur as a result of assumptions that researchers need to make when completing analyses.

Given that value-added models work with multi-year data, vertical scaling or equating of scores is highly desirable, as it is related to the construct validity of the students’ scores. Yet this is not always an easy task to accomplish; using interval-scaling would not work either unless data from only two years is used (Braun, 2004).

Validity and reliability of test scores are usually accomplished when standardized assessments are administered. As noted in the literature they vary by type of assessment and by state (Braun, 2004).

Among the disadvantages of value-added models we note the lack of randomization in assigning students to classes (Braun, 2004) or depending on the growth model considered, the exclusion of certain covariates from the student level equations, which can further lead to biased estimators. Given that it is difficult to organize a randomized experiment in school settings, as it is difficult to randomly assign students or teachers to classes, causal inferences about teachers’ effectiveness based on value-added models should be made with caution. These models do not
produce unbiased results if data are not missing at random or if more than 10% of the data are missing.

Linear mixed models combine fixed with random effects. In certain studies, classroom-level effects are considered fixed while in others they are considered random. Some expressed reservations about classifying class-level effects as random (Raudenbush, 2004) while others consider that approach optimal in order to get the Best Linear Unbiased Predictors through empirical Bayes estimates (Braun, 2004).

Teachers’ effects vary over time. Studies have led to different results based on how this assumption was incorporated into the model. Causality between teachers’ effectiveness and students’ growth over time is hard to establish, given the fact that most value-added or growth studies are observational and not experimental.

The mathematical model requires that error terms in equations be normally distributed and not related to the effects of other covariates.

In addition to analytical and methodological assumptions that lead to limitations of value-added models we shall mention that there are limitations related to sociological factors or to neighborhood effects on students’ achievement consequently reflected in teachers’ effects. As discussed at large in education, it is unclear the extent to which parental level of education, or poverty level are measured and accounted for properly in value-added models.
V. Mathematics Assessment Data: Overview

The longitudinal student math assessment data were selected to measure students’ growth over time through a multilevel approach with continuous outcomes as well as through a simpler approach consisting of mean comparisons with non-continuous outcomes. As growth models derive strength from longitudinal assessment scores complemented by background data specific to each unit of analysis we chose to use the cohort shown in Table 2 which has the most consistent data points available, since New York State first introduced the mathematics assessments in 2005.

Among the strengths of this data set are the inclusion of valid and reliable performance scores, score alignment to New York State mathematics indicators for each grade level, and the attribution of ordinal as well as continuous scores to students’ performance in each academic year. Among some of the weaknesses we enumerate the lack of measurement of students’ performance in subject area only once a year.

The rounds and instruments for data collection are also shown in table 2.

Table 2: Cohort, instruments and rounds for data collection for longitudinal analyses

<table>
<thead>
<tr>
<th>School Year</th>
<th>Grade level</th>
<th>Rounds of Data Collection</th>
<th>Instruments of Data Collection</th>
</tr>
</thead>
<tbody>
<tr>
<td>2005 - 2006</td>
<td>3rd grade</td>
<td>March 2006</td>
<td>3rd grade NYS math assessment</td>
</tr>
<tr>
<td>2006 - 2007</td>
<td>4th grade</td>
<td>March 2007</td>
<td>4th grade NYS math assessment</td>
</tr>
<tr>
<td>2007 - 2008</td>
<td>5th grade</td>
<td>March 2008</td>
<td>5th grade NYS math assessment</td>
</tr>
<tr>
<td>2008 - 2009</td>
<td>6th grade</td>
<td>March 2009</td>
<td>6th grade NYS math assessment</td>
</tr>
</tbody>
</table>
The data set contains approximately 481 3rd grade students from two different New York State school districts whose math assessment data are monitored over a period of four years as they advance from grade 3 to grade 6. The students come from seven elementary and middle schools which are part of two suburban school districts with somewhat different demographics. One district has three elementary schools with grades K-5, one middle school with grades 6-8, one high school with grades 9-12 and one alternative program. Another district consists of four elementary schools, two with grades K-2 and two with grades 3-5, one middle school with grades 6-8, one high school with grades 9-12 and one alternative program including high school students identified as not being able to succeed in a traditional academic setting. On average there are about 20 teachers for each grade level per year, each with an average of about 25 students in their classes. Union rules typically prevent significant differences in class sizes or allocation of students among teachers in order to ensure equity of resources per student per class. The cohorts under consideration were assessed in mathematics in March of each year and data were collected and reported by school districts to the NYSED. The NYSED reported assessment scores back to the districts within several months of each assessment administration.

When preparing the data sample for our analyses we further eliminated teachers whose class sizes consisted of eight or fewer students as they are likely to be those who may assist students classified as special needs or ELL. We also eliminated students who did not have scores from two consecutive years of study in that same school district given that we use their performance from each end-of-year year as their entry score or performance for the following year. After the data cleaning process our sample consisted of 400 students entering grade 4 who also had grade 3 math assessments scores and who attended one of the 20 teachers’ classrooms retained in the analysis. The measurable outcome was a continuous variable consisting of the
students’ end-of-year NYS math assessment score and reflecting students’ proficiency in topics
tested in their respective grade-level math assessments.

A secondary outcome was the proficiency level recorded for each student who took the
math assessment. Following a standardized assessment, students’ performance fell into one of
the four performance categories (1, 2, 3, or 4) where 4 represents the best possible performance
level, and 1 is the weakest performance level. In 2010-2011, The NYS Education Department
renamed students’ performance levels in order to align them with higher expectations of students
passing the first high school regents’ exam necessary for graduation (NYSED, 2010), as shown
in Table 3. All students’ math performance was measured at the same time during the testing
period and all assessments administered were standardized and aligned to the New York State
math standards for each grade level.

Table 3: NYS Performance indicators for standardized yearly math assessments

<table>
<thead>
<tr>
<th>Performance Level</th>
<th>Label</th>
<th>Grade 3</th>
<th>Grade 4</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>% Students at Each Level</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Level 1</td>
<td>Below Standard</td>
<td>6.35%</td>
<td>7.41%</td>
</tr>
<tr>
<td>Level 2</td>
<td>Meets Basic Proficiency Standard</td>
<td>13.13%</td>
<td>14.59%</td>
</tr>
<tr>
<td>Level 3</td>
<td>Meets Proficiency Standard</td>
<td>55.42%</td>
<td>52.12%</td>
</tr>
<tr>
<td>Level 4</td>
<td>Exceeds Proficiency Standard</td>
<td>24.11%</td>
<td>25.88%</td>
</tr>
</tbody>
</table>


Copyright 2006 by the New York State Education Department.
Variables

The student-, class- and school-level variables that were collected from the school districts are summarized in Table 4. The student-level variables were recorded by schools or by district personnel through a survey administered to parents or guardians when they first enroll their children in school. The classroom- and school-level variables were compiled from surveys administered to teachers by schools and reported to the state while the district-level data were compiled from information reported to NYSED by individual schools. The type of certification held by each teacher is categorized by New York State as permanent, professional or initial. Teachers may sometime hold more than one certification.

Table 4: Variables for Multi-level Analyses

<table>
<thead>
<tr>
<th>Linkage</th>
<th>Student-Related Variables</th>
<th>Teacher-Related Variables</th>
</tr>
</thead>
<tbody>
<tr>
<td>Student ID</td>
<td>Yearly Scaled Math Score</td>
<td>Teacher ID</td>
</tr>
<tr>
<td></td>
<td>Proficiency Level/ Score</td>
<td>Percentage of male students</td>
</tr>
<tr>
<td></td>
<td>Half Numerical level</td>
<td>Percentage of Students on Free Lunch or on Reduced Lunch</td>
</tr>
<tr>
<td></td>
<td>Student Gender</td>
<td>Percentage of minority students</td>
</tr>
<tr>
<td></td>
<td>SES, Free Lunch; Reduced Price Lunch</td>
<td>Percentage of LEP students</td>
</tr>
<tr>
<td></td>
<td>Ethnicity, African American, Asian, Native American, Hispanic, White</td>
<td>Percentage of Classified Students</td>
</tr>
<tr>
<td></td>
<td>Limited English Proficiency, based on performance on NYSESLAT</td>
<td>Mean Class Math Achievement for each Year</td>
</tr>
<tr>
<td></td>
<td>Learning Disability</td>
<td>Class Size</td>
</tr>
<tr>
<td></td>
<td>Multiple-Choice and Constructive Response Scores;</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Total Number of Assessment Measures</td>
<td></td>
</tr>
</tbody>
</table>
VI. Questions

New changes in current national and local educational policy together with the objectives of this study lead us to the following research questions:

1. Are the VA results obtained from using ordinal measures of students’ scores significantly different from the VA results obtained with continuous measures on the same data points? If so, which method may be the most economical one to implement in order to measure value-added?

2. How well do these models inform actionable policy? If effective, which of these methods might be practical for districts to adopt without having to pay for services consisting of more complex statistical modeling?
VII. Methodology

In various states, students were and continue to be tested yearly in subject areas such as Mathematics, English Language Arts, Science or Social Studies. After testing in New York State, parents and schools receive assessment results reported as ordinal and continuous measures of students’ performance reflecting students’ performance for one academic year. This study uses both assessment measures to investigate two methods that analyze data collected in different formats to produce value-added for teachers. We also investigate how the outcomes of the methodologies applied to the same dataset consisting of students’ math assessment scores compare, and then discuss some qualities and shortcomings of each method.

The goals of this study are to:

1. Propose and evaluate a new method to measure value-added in students’ math performance over time, which allows for the use of ordinal scales or intake of discrete data;
2. Use mixed models with fixed and random effects to measure students’ growth over time allowing for the use of scores measured on a continuous scale.
3. Compare the outcomes of these two methods, their ease of use, and their ease of interpretation to decide whether the implementation of a more complex model is justifiable.

Description of Method 1 Using Ordinal Measures to Assess Value Added

In order to evaluate value-added from year to year or between any two data points for a specific teacher it is necessary to determine students’ proficiency in a subject area at the beginning and at the end of the instructional period under consideration. This helps to control for students’ status from prior year as well as for the make-up or structure of each teacher’s class in terms of students’ performance for the school year under analysis. When the outcome variable, y, in our case representing students’ math performance on the end-of-year tests is not continuous
but rather ordinal it is more difficult to embed it into a hierarchical linear approach that assumes normality at level 1 (Raudenbush, & Bryk, 2002). Some have explored solutions to this problem such as Goldstein’s (1984) and Longford’s (1993) software approach to multilevel models with discrete (including ordinal) outcomes using somewhat sophisticated statistical models perhaps not easy to follow or implement by school districts with light expertise in statistics.

Due to the level of complexity of such methods we propose and further explore a simple method that can be implemented by the non-sophisticated statistician who uses any type of database processing software. This approach is consistent with the findings of Martineau (2006) who claims that, “With current technology, there are no vertical scales that can be validly used in high-stakes analyses for estimating value-added to student growth in either grade-specific or student-tailored construct mixes – the two most desirable interpretations of value-added to student growth. At this point, this leaves only one satisfactory approach to high-stakes VAA using current technology; the measurement of a given grade-level’s content in both the grade below and the appropriate grade level to obtain an estimate of value added to a static mix of constructs specific to each grade.” Given that at the elementary level, the math content and process state indicators are fairly close in difficulty level and also well aligned between grades, the approach of using two consecutive years for comparisons in students’ performance is appropriate.

We evaluated teachers’ performance over the course of the school year using their students’ ordinal scores. To do this fairly, we adjusted for the scores of students in the previous year by conditioning on previous year’s scores, and then computed a weighted average using the same weights for each teacher. The weights could be derived in several plausible ways, including
equal weights, by the distribution of weights for an average class in a school, or by the average New York State class. We used weights based on New York State averages.

The procedure consists of the following main steps:

1. For each teacher, we created a cross-tab table showing the distribution of their students’ ordinal scores (1, 2, 3 or 4) at the beginning of year 1 and at the end of year 1, where students enter year 1 with the scores they received on their math state assessments at the end of year 0. The data was cleaned to ensure its calculability (e.g., when finding zero frequency adding .025 to produce a non-zero quantity and to avoid division by zero)

2. Determined the marginal frequency distributions for each score level by finding subtotals for each row and column in the table.

3. Adjusted the score distributions for each level score in year 1 by students’ performance from year 0. This was accomplished by dividing each row of distribution scores by the frequency score distributions for that score level from year 0, if non-zero.

4. Calculated a weighted average score at each level/score from year 0 to determine the teacher’s performance for that score level at the end of Year 1, using frequencies as weights. Two versions of weighting are discussed in our example though we are modeling and using in our comparison of results section the method reflecting state weights (NYS overall mathematics test performance level distributions in mathematics for grades 3-8, Appendix 3).

5. Then used weighted averages at each level to calculate an overall score for the teacher, by either averaging the numbers determined at each level (the unweighted average) and by accounting for the number of students who scored at each level at the end of Year 1 (class weighted average) or by using the school or the overall New York State students’
performance weights at each level to determine a weighted average, which models an average state class.

6. Value-added was determined by ranking the teachers in the order of their class mean scores thus finding who added more value to their students as compared to the rest of the teachers in their school or group.

Note that teachers in grades 3-6 generally teach one math class along with other subjects, so data are sparse. Therefore if one of the score level/category is zero, meaning that there are no students falling in that specific category, the procedure described could be followed by considering only three instead of the initial four categories and by adjusting the weights accordingly.

We begin by using artificial data along with necessary explanations to model this procedure in order to give the interested practitioner an opportunity to understand and to possibly replicate this method.

From the existing longitudinal database, consider students’ performance at the end and at the beginning of year 1. Recall that for a given student, we consider the math performance level at the end of year 0 to be the equivalent of his performance score at the beginning of year 1, although this assumption has its own limitations, given that students may forget content or may increase their knowledge of mathematics over the summer break. Similarly, a class performance structure of ordinal scores is found by aggregating the students’ performance at the beginning of that academic year and at the end of the same academic year.

Let us assume that teacher A has 25 students each with scores ranging from 1 to 4 in Year 0 and in Year 1, with a score distribution as shown in Table 5. Let us further assume that state
weights are those shown in Table 8, as we will use them to exemplify computations of further weighting for local and statewide comparisons.

Table 5: Relation between Year 0 and Year 1 Scores for a 25-student Class

<table>
<thead>
<tr>
<th>Year 0 Score</th>
<th>Year 1 Score</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>1</td>
</tr>
<tr>
<td>4</td>
<td>0</td>
</tr>
<tr>
<td>3</td>
<td>1</td>
</tr>
<tr>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>1</td>
<td>2</td>
</tr>
</tbody>
</table>

The proposed algorithm is shown below:

1) From the table 5, add across each row to determine the number of students who began the year in each of the four score categories as shown in Table 6.

Table 6: Observed Student Score Counts by Year

<table>
<thead>
<tr>
<th>Scores</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>Year 0 Observed</th>
</tr>
</thead>
<tbody>
<tr>
<td>4</td>
<td>0</td>
<td>1</td>
<td>2</td>
<td>7</td>
<td>10</td>
</tr>
<tr>
<td>3</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>0</td>
<td>6</td>
</tr>
<tr>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>0</td>
<td>6</td>
</tr>
<tr>
<td>1</td>
<td>2</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>3</td>
</tr>
<tr>
<td>Year 1 Observed</td>
<td>5</td>
<td>6</td>
<td>7</td>
<td>7</td>
<td>25</td>
</tr>
</tbody>
</table>

This teacher began the year with 10 students who scored a “4”, 6 students who scored a “3”, 6 students who scored a “2”, and 3 students who scored a “1”.

2) To determine the distribution of students in each of the subscore categories, divide each row by its row total, if non-zero, and round to the nearest hundredth.
3) Calculate a weighted average at each score level from Year 0 to determine each teacher’s performance as measured by how well he or she did for students at each level at the end of Year 1.

Table 7: Weighted Averages by Year

<table>
<thead>
<tr>
<th>Student Scores in Year 0</th>
<th>Year 1</th>
<th>Weighted avg. in Year 1 adjusted for Year 0 performance</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>4</td>
<td>0</td>
<td>.10</td>
</tr>
<tr>
<td>3</td>
<td>.17</td>
<td>.33</td>
</tr>
<tr>
<td>2</td>
<td>.33</td>
<td>.33</td>
</tr>
<tr>
<td>1</td>
<td>.67</td>
<td>.33</td>
</tr>
</tbody>
</table>

To calculate how students who received a “4” in year 0 performed in year 1, we calculate a weighted average for each score category as:

Weighted Average Score (score level 4) = (0)1 + (.10)2 + (.20)3 + (.70)4 = 3.60

(25)

Similarly,

Weighted Average (level 3) = (.17)1 + (.33)2 + (.50)3 + (0)4 = 2.33

(26)

Weighted Average (level 2) = (.33)1 + (.33)2 + (.33)3 + (0)4 = 1.98

(27)

Weighted Average (level 1) = (.67)1 + (.33)2 + (0)3 + (0)4 = 1.33

(28)

At this point in the analysis, the results from Table 7 suggest that, on average, students who entered the year with scores of “4”, “3”, and “2” have decreased in their math performance at the end of that academic year while those who entered the year with the lowest score have gone slightly up. Note that there will almost always be regression to the mean; unless all students who score 4 in Year 0 do so again in Year 1, the weighted average must be less than 4 (and similarly for students who score 1 in Year 0, the weighted average must be greater than 1.)
We can weight each category by equal weights, school or district weights, or by New York State weights, in order to compare teachers, since we are controlling for the marginal differences and including the conditional means.

For equal weights, we would just average the conditional averages for each Year 0 score. In order to assess how teachers compare to other teachers within their own school or within their own state we need to multiply the weight by category by the school weight or by the state weight. When using the school weight we determine the same teacher performance if he would teach an average class in that school. When using the state weight we would determine this teacher’s performance if he would be teaching an average New York State class.

4) Then aggregate the data at the school level to see how teacher performance compares over the course of the year. For each column, average the numbers (sum and divide by 4) in order to determine, the standardized adjusted proportion in each score category. Note, that if any row total is zero we recommend omitting that row and dividing the sum by the number of remaining categories, in our example by 3. Then, report the categories/rows on which it is based.

Table 8: Calculated Teacher Performance at the End of Year 1

<table>
<thead>
<tr>
<th>Teachers</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>Unweighted Avg.</th>
<th>Weighted Avg. by class using state weights</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1.33</td>
<td>1.98</td>
<td>2.33</td>
<td>3.60</td>
<td>2.31</td>
<td>2.56</td>
</tr>
<tr>
<td>2</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>n</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Unweighted Avg\textsubscript{Class 1} = \sum_{i} \frac{1}{4} x_{i} = (1.33 + 1.98 + 2.33 + 3.60) / 4 = 2.31 \hspace{1cm} (29)
Suppose the state weights for each categorical score are as shown below.

Table 9: Adjusted Weighted Teacher Performance at the End of Year 1

<table>
<thead>
<tr>
<th>Score Level</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
</tr>
</thead>
<tbody>
<tr>
<td>State</td>
<td>.06</td>
<td>.14</td>
<td>.53</td>
<td>.27</td>
</tr>
</tbody>
</table>

Weighted Avg.\(_{\text{Class 1}}\) = \(\sum w_i \bar{x}_i = 1.33(.06) + 1.98(.14) + 2.33(.53) + 3.60(.27) = 2.56\) \(\text{(30)}\)

Where, \(\bar{x}_i\) = the mean for class \(i\);
\(w_i\) = the state weights attributed to each level mean.

\[\sum_{i=1}^{4} w_i = 1 \text{ and } w_i \geq 0.\] \(\text{(31)}\)

If, in a specific district, students are being tracked resulting in students’ scores spread over three instead of the four categories, the same calculations will be employed, but the results will only be compared with teachers teaching similar classes in each respective school, school district or in New York State, if data are available. It may be useful to analyze the distribution of scores by teachers at the beginning of the analysis and to group teachers first into specific categories, those who have all four categories represented in their class, only the top three categories (2, 3, and 4) or only the bottom three categories (1, 2, and 3). Unless there is a strict tracking system employed in the school, we estimate that students’ scores will fall into one of the groups just enumerated. If teachers are missing students in more than one category a different methodology or an extension of this one may be developed.

Interpretation of VA for Method 1

Value added is reflected in the weighted average of a given teacher’s class and the results produced if he or she had taught the average school class or the average state class. These
calculations cannot be interpreted standing alone. They must be compared with other teachers from their school, district or state to help us determine whether teachers added or did not add value to their students’ education in one year above what they would have contributed to the average class in their school, district or state.

We also assess the standard error of the estimate in order to produce interval estimates for teachers’ value-added by using the bootstrapping procedure (Hesterberg, Monaghan, Moore, Clipson, & Epstein, 2003). Given that class sizes are fairly small for each teacher taken individually, the bootstrapping generates a large number of samples of students’ scores similar to the ones that make up each class scores by sampling from each teacher’s scores with replacement as if each teacher had taught thousands of such classes. “The original sample represents the population from which it was drawn. Thus, resamples from the original sample represent what we would get if we took many samples from the population. The bootstrap distribution of a statistic based on the resamples, represents the sampling distribution of the statistic.” (Hesterberg, Monaghan, Moore, Clipson, & Epstein, 2003). Using R we then produce confidence intervals for each teacher’s average.

Some of the advantages of this method consist of using the same cohort of students, thus controlling for student level variables, predicting growth between two consecutive years which would address the limitations of vertically aligned scores, using fairly simple calculations to enable practitioners to implement this method with in-house resources, and making use of ordinal scores to calculate VA in a fairly straightforward manner, which could be replicated at different levels within the educational system.
Description of Method 2

A hierarchical linear model with fixed and random effects to measure students’ growth over time allowing for use of scores measured on a continuous scale.

The purpose of this method is to evaluate teachers’ performance over the course of the school year using their students’ continuous assessment scores from that year while controlling for students’ scores from the end of the prior year. In this model, the entry scores were centered around the New York State averages for that specific school year to make the interpretation of our results more meaningful. In the proposed hierarchical linear models the outcome variable is the student’s assessment score at the end of Year 1. This outcome was modeled by adding meaningful covariates to the level 1, such as the student’s score at the beginning of year 1, or variables related to students’ school classification such as LEP (Limited English Proficiency), SES (socio-economic status) or to their ethnicity. Class size, a variable at the teacher level was used to model the level 2 equations to further help in explaining the intercept for the unconditional model and the intercept and slopes for the conditional models. Thus, we compared teachers’ performance over one year with that of their co-workers by comparing the error terms found in the level 2 equations for intercepts. If the errors of the second level intercept equations are positive than we can argue that the teacher’s performance is above average for that specific year as compared to the rest of the teachers from the pool (school, district or state). Similarly, if any teachers’ averages are negative we can argue that their performance is below average.
Determining value-added

We interpreted that a teacher added “value” to his students’ education over the course of a year if his results showed large intercepts and small slopes or consequently large error terms ($u_0$) in the 2nd level equation for intercept and small error terms ($u_i$) in the second level equations explaining slopes. We organized teachers’ errors of the second level intercept equations in ascending order and interpreted value-added similarly to the first method. The teachers who ranked more to the right of the chart were said to have added more value to their students’ education during that school year as compared to their sample counterparts with lower scores and non-overlapping confidence intervals lower than theirs.

Metric Descriptive Statistics

The metric used reflected students’ scaled scores as reported by the New York State Education Department to the school districts. Typically scaled scores from consecutive grades have very similar means and standard deviations during the school years under consideration.

Table 10. Descriptive Statistics for Continuous Student Pre and Post Scores

<table>
<thead>
<tr>
<th></th>
<th>N</th>
<th>Minimum</th>
<th>Maximum</th>
<th>Mean</th>
<th>Std. Deviation</th>
</tr>
</thead>
<tbody>
<tr>
<td>St_Score.0: Scaled Score</td>
<td>459</td>
<td>599</td>
<td>770</td>
<td>689.82</td>
<td>34.938</td>
</tr>
<tr>
<td>St_Score.1: Scaled Score</td>
<td>476</td>
<td>615</td>
<td>800</td>
<td>700.41</td>
<td>35.827</td>
</tr>
<tr>
<td>Valid N (listwise)</td>
<td>426</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Note: St_Score.0 represent the students’ assessments scores at the end of year zero which we equate with their scores on the pre-test as they enter year 1.
St_Score.1: represent the students’ assessments scores at the end of year one which we equate with their scores on the post-test as they exit year 1.

By considering assessment tests from two consecutive years at the elementary level we look at very similar curricular contents which make comparability more meaningful.

Hierarchical Models

In a multilevel linear model we have fixed and random components to model the students’ variability about their own averages as well as about the average performance of the grouping of which they are a part at the school, district or state levels.

In order to keep our analysis to a two-level model we will consider the nesting as students nested within teachers or classes and classes nested within the data pool consisting of the two combined school districts since this study uses data from two school districts and from seven schools.

Based on a preliminary analysis of data, specific variables will be used as controls in our mixed models. The variables included were selected based on the correlations shown in Appendix 5 and also on latest New York State guidance document. The NYS Education Department indicated in their latest APPR guidance document the following possible student characteristics to be considered in their value-added models they aim to adopt in the near future: student state assessment history, poverty indicators, disability indicators, ethnicity/race, gender, percent daily student attendance, student suspension data, retention, summer school participation, student new to school in a non-articulating year, and student age. The plot showing the student state assessment score history shows a big floor and ceiling effects. If this is in fact the case then a linear model would not be the correct one to use. Other non-student characteristics considered
by NYSED in fitting the hierarchical models would include classroom characteristics such as class size and percent with each demographic characteristics in a class, school characteristics such as percentage with each demographic characteristic, class average and grade configuration and/or, educator experience level in his/her role (NYSED, 2012).

As previously discussed, in Sander’s models variables related to student ethnicity, student limited English proficiency status, and student disability classification have effects already built into both the pretest score from a prior year, a covariate accounted for in Level 1 and the current year score and thus are not explicitly included in our models. Some, however, may consider including them as covariates in order to test if they still play a statistically significant role in a multilevel model.

Due to the complex nature of multilevel modeling and to the structure of our data we will restrict our multilevel models to only two level equations.

After running correlations between variables to be considered as shown in Table 11 included below, it is customary to begin the analysis with a visual inspection of the equations, check assumptions for normal distribution of variables (skewness and kurtosis), run a baseline unconditional or very simple conditional model in order to provide useful baseline statistics to further evaluate individual growth (Raudenbush & Bryk, 2002). The first model begins with the student-level equation consisting of the mean class score on the math posttest (the intercept), the students pre-scores from prior years and the error. The next conditional model was obtained by adding covariates as controls into the unconditional model first-level equations mentioned before or to second-level equations to try to improve the fit of the model. Continuous variables, such as students’ scores from prior year and class size for teachers were centered in order to make the interpretation of results more meaningful.
Table 11: Correlation table among student level variable considered for HLM analyses

**Correlations on 400 student data** (continued on the next page)

<table>
<thead>
<tr>
<th></th>
<th>Score1</th>
<th>Score0</th>
<th>St_SES</th>
<th>St_LEP</th>
<th>St_Classified</th>
<th>White</th>
<th>Asian</th>
<th>Hispanic</th>
<th>Black</th>
</tr>
</thead>
<tbody>
<tr>
<td>Score1 Pearson Correlation</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Sig. (2-tailed)</td>
<td>.000</td>
<td>.000</td>
<td>.022</td>
<td>.195</td>
<td>.298</td>
<td>.000</td>
<td>.000</td>
<td>.000</td>
<td>.000</td>
</tr>
<tr>
<td>N</td>
<td>400</td>
<td>400</td>
<td>400</td>
<td>400</td>
<td>400</td>
<td>400</td>
<td>400</td>
<td>400</td>
<td>400</td>
</tr>
<tr>
<td>Score0 Pearson Correlation</td>
<td>.627**</td>
<td>1</td>
<td>-.387**</td>
<td>-.240**</td>
<td>-.076</td>
<td>.102*</td>
<td>.304**</td>
<td>-.287**</td>
<td>-.297**</td>
</tr>
<tr>
<td>Sig. (2-tailed)</td>
<td>.000</td>
<td>.000</td>
<td>.000</td>
<td>.129</td>
<td>.041</td>
<td>.000</td>
<td>.000</td>
<td>.000</td>
<td>.000</td>
</tr>
<tr>
<td>N</td>
<td>400</td>
<td>400</td>
<td>400</td>
<td>400</td>
<td>400</td>
<td>400</td>
<td>400</td>
<td>400</td>
<td>400</td>
</tr>
<tr>
<td>St_SES Pearson Correlation</td>
<td>-.281**</td>
<td>-.387**</td>
<td>1</td>
<td>.293**</td>
<td>-.010</td>
<td>-.328**</td>
<td>-.241**</td>
<td>.485**</td>
<td>.301**</td>
</tr>
<tr>
<td>Sig. (2-tailed)</td>
<td>.000</td>
<td>.000</td>
<td>.000</td>
<td>.846</td>
<td>.000</td>
<td>.000</td>
<td>.000</td>
<td>.000</td>
<td>.000</td>
</tr>
<tr>
<td>N</td>
<td>400</td>
<td>400</td>
<td>400</td>
<td>400</td>
<td>400</td>
<td>400</td>
<td>400</td>
<td>400</td>
<td>400</td>
</tr>
<tr>
<td>St_LEP Pearson Correlation</td>
<td>-.115*</td>
<td>-.240**</td>
<td>.293**</td>
<td>1</td>
<td>.018</td>
<td>-.268**</td>
<td>.037</td>
<td>.348**</td>
<td>-.080</td>
</tr>
<tr>
<td>Sig. (2-tailed)</td>
<td>.022</td>
<td>.000</td>
<td>.000</td>
<td>.722</td>
<td>.000</td>
<td>.464</td>
<td>.000</td>
<td>.109</td>
<td></td>
</tr>
<tr>
<td>N</td>
<td>400</td>
<td>400</td>
<td>400</td>
<td>400</td>
<td>400</td>
<td>400</td>
<td>400</td>
<td>400</td>
<td>400</td>
</tr>
<tr>
<td>St_Classified Pearson Correlation</td>
<td>-.065</td>
<td>-.076</td>
<td>-.010</td>
<td>.018</td>
<td>1</td>
<td>-.039</td>
<td>.029</td>
<td>-.002</td>
<td>.030</td>
</tr>
<tr>
<td>Sig. (2-tailed)</td>
<td>.195</td>
<td>.129</td>
<td>.846</td>
<td>.722</td>
<td>.441</td>
<td>.557</td>
<td>.963</td>
<td>.546</td>
<td></td>
</tr>
<tr>
<td>N</td>
<td>400</td>
<td>400</td>
<td>400</td>
<td>400</td>
<td>400</td>
<td>400</td>
<td>400</td>
<td>400</td>
<td>400</td>
</tr>
<tr>
<td>White Pearson Correlation</td>
<td>.052</td>
<td>.102*</td>
<td>-.328**</td>
<td>-.268**</td>
<td>-.039</td>
<td>1</td>
<td>-.570**</td>
<td>-.461**</td>
<td>-.247**</td>
</tr>
<tr>
<td>Sig. (2-tailed)</td>
<td>.298</td>
<td>.041</td>
<td>.000</td>
<td>.000</td>
<td>.441</td>
<td>.000</td>
<td>.000</td>
<td>.000</td>
<td>.000</td>
</tr>
<tr>
<td>N</td>
<td>400</td>
<td>400</td>
<td>400</td>
<td>400</td>
<td>400</td>
<td>400</td>
<td>400</td>
<td>400</td>
<td>400</td>
</tr>
<tr>
<td></td>
<td>Pearson Correlation</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>-------</td>
<td>---------------------</td>
<td>-------</td>
<td>-------</td>
<td>-------</td>
<td>-------</td>
<td>-------</td>
<td>-------</td>
<td>-------</td>
<td>-------</td>
</tr>
<tr>
<td>Asian</td>
<td>Sig. (2-tailed)</td>
<td>.000</td>
<td>.000</td>
<td>.000</td>
<td>.464</td>
<td>.557</td>
<td>.000</td>
<td>.000</td>
<td>.001</td>
</tr>
<tr>
<td>N</td>
<td></td>
<td>400</td>
<td>400</td>
<td>400</td>
<td>400</td>
<td>400</td>
<td>400</td>
<td>400</td>
<td>400</td>
</tr>
<tr>
<td>Hispanic</td>
<td>Sig. (2-tailed)</td>
<td>.000</td>
<td>.000</td>
<td>.000</td>
<td>.000</td>
<td>.963</td>
<td>.000</td>
<td>.000</td>
<td>.009</td>
</tr>
<tr>
<td>N</td>
<td></td>
<td>400</td>
<td>400</td>
<td>400</td>
<td>400</td>
<td>400</td>
<td>400</td>
<td>400</td>
<td>400</td>
</tr>
<tr>
<td>Black</td>
<td>Sig. (2-tailed)</td>
<td>.000</td>
<td>.000</td>
<td>.000</td>
<td>.109</td>
<td>.546</td>
<td>.000</td>
<td>.001</td>
<td>.009</td>
</tr>
<tr>
<td>N</td>
<td></td>
<td>400</td>
<td>400</td>
<td>400</td>
<td>400</td>
<td>400</td>
<td>400</td>
<td>400</td>
<td>400</td>
</tr>
</tbody>
</table>

**. Correlation is significant at the 0.01 level (2-tailed).
*. Correlation is significant at the 0.05 level (2-tailed).
Table 12: Summary of Multilevel Models under Consideration

<table>
<thead>
<tr>
<th>Model</th>
<th>Model Type</th>
<th>Variables Included</th>
<th>Interactions</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Conditional</td>
<td>Slope - intercept</td>
<td>Level 1: Pretest</td>
<td>None</td>
</tr>
<tr>
<td>2. Conditional</td>
<td>Slope - intercept</td>
<td>Level 1: Pretest, Score, StSES, StLEP, White, Asian, Hispanic, Black</td>
<td>None</td>
</tr>
</tbody>
</table>

1st Model: **A Model with Conditional Pre-Score First-Level and Unconditional Second-Level Equations**

This first model assesses how variation in students’ end-of-year scores is allocated between students and teachers. The outcome is student post-assessment math score at the end of year 1, and the only predictor is student pre-score in mathematics or the score each student enters the school year with from their previous year’s math state assessment.

Level 2: has the intercept and slope as outcomes with no predictors.

The intercept, $\beta_0$ is specified as random and the slope, $\beta_1$ is also specified as random.

Level-1: $Y_{1ij} = \beta_0 + \beta_1 (Pre_{ij} - 680) + r_{ij}$  \hspace{1cm} (32) \hspace{1cm} (student-level equation)

Level 2: $\beta_{0j} = \gamma_{00} + u_{0j}$  \hspace{1cm} (33) \hspace{1cm} (teacher/class-level equations)

$\beta_{1j} = \gamma_{10} + u_{1j}$  \hspace{1cm} (34)

Mixed Model: $Y_{ij} = (\gamma_{00} + u_{0j}) + (\gamma_{10} + u_{1j})(Pre_{ij} - 680) + r_{ij}$  \hspace{1cm} (35)
Where, \( i = \) individual student and \( j = \) teacher or class, \( E(r_{ij}) = 0 \), and

- \( Y_{ij} = \) math assessment score at the end of Year 1 for student \((i)\) in class \((j)\) for that subject during that year;
- \( \beta_{0j} = \) math mean assessment score of class \((j)\) at the end of Year 1 for students who had an average score on the pre-test;
- \( r_{ij} = \) random error at level-1 with \( \text{var}(r_{ij}) = \sigma^2 \) being the amount of variance remaining unexplained after accounting for the effect of pretest on end of Year 1 test. This is also known as a random “student effect” showing the deviation of student \((i)\) in class \(j\) from the math assessment classroom mean. We assume these effects to be normally distributed with a mean of zero and variance, \( \sigma^2 \).
- \( \gamma_{00} = \) the mean performance of students’ teachers from both districts combined at the end of Year 1 when adjusted for students pre-score from deviation from the average intercept;
- \( u_{0j} = \) random error at level-2 showing the deviation of mean performance of teacher \(j\) from the grand mean. We will further assume these effects normally distributed with a mean of zero and a variance, \( \tau_{00} \);
  - \( \text{var}(u_{0j}) = \tau_{00} \) – is the population variance among the teachers’ intercepts;
- \( \sigma^2 = \text{var}(e_{ij}) = \) variance within classes or within group residual variance;
- \( \tau_{00} = \text{var}(u_{0j}) = \) variance among classes in the pool;
- \( \tau_{11} = \text{var}(u_{1j}) = \) variance in the teachers’ slopes;
The covariate pretest will be centered around the state average from the year when the state exam was administered and rounded to the nearest ten. This will then become the centered grand mean of the math pretest scores from the baseline year.

2nd Model: A Conditional Model with Conditional First and Second-Level Equations

Level-1: \( Y_{ij} = \beta_{0j} + \beta_{1j}(\text{Pre}_{ij} - 680) + \beta_{2j}(\text{St SES}_{ij}) + \beta_{3j}(\text{White}_{ij}) + \beta_{4j}(\text{Hispanic}_{ij}) + \beta_{5j}(\text{Asian}_{ij}) + \beta_{6j}(\text{Black}_{ij}) + \beta_{7j}(\text{ST LEP}_{ij}) + r_{ij} \) (36) (student level equation)

Level 2: \( \beta_{0j} = \gamma_{00} + \gamma_{01}(\text{Class SI}_j) + u_{0j} \) (37) (teacher/class-level equation)

\[
\begin{align*}
\beta_{1j} &= \gamma_{10} + u_{1j} \\
\beta_{2j} &= \gamma_{20} \\
\beta_{3j} &= \gamma_{30} \\
\beta_{4j} &= \gamma_{40} \\
\beta_{5j} &= \gamma_{50} \\
\beta_{6j} &= \gamma_{60} \\
\beta_{7j} &= \gamma_{70}
\end{align*}
\]

Where,

- \( \text{Pre}_{ij} \) = preassessment math score for student “i” in class “j” at the beginning of the year;
- \( \text{Pre}_{ij} - 680 \) = preassessment math score for student “i” in class “j” at the beginning of the year centered around the state average math score;
- \( \text{CS}ize \) = class size by teacher;
\( \gamma_{01} \) = the effect of class size on class intercept ; we will find the average intercept for a class while controlling for teacher’s class size;

\( \gamma_{p0} (p=2 \text{ to } 7) \) represents the fixed effects of students socio-economic status, ethnicity and limited English proficiency on the end of year score, when we model while controlling for the others.

\( \gamma_{10} \) is the effect of math pre-test score on the math post-test score;

We are not including random effects for each level 1 covariates because we do not have enough students per teacher to obtain accurate estimates of these effects.

Given that there are very few classified students and that students’ classification status was not significantly correlated with students outcome at the end of year 1 we excluded this variable from our analyses. Class size was included in our analyses since research states that class size has a significant impact on student class instruction and thus possibly on their end of year achievement. For example, teachers who teach larger classes might be at a disadvantage that needs adjustment.

When running value-added models we check if the following assumptions are met (Raudenbush, & Bryk, 2002):

- \( e_{ij} \) errors are normally distributed with homogeneous variances across groups or \( e_{ij} \sim N(0, \sigma^2) \);

- The subscript \( j \) used for intercept and slope indicates that each group has its unique intercept and slope which they have a bivariate normal distribution across the population of classes;
Method 1 involves arranging the data in the wide format, using SPSS or Excel to calculate cross-tabs (see Appendix 1 for code) in order to determine the allocation of students’ math performance level by teacher. 95% confidence intervals around each teacher’s math class average scores were calculated by using the bootstrapping technique. These confidence intervals and the teachers’ mean math scores were further used to determine teachers’ rankings. We considered the state true weights when calculating teachers’ averages and the corresponding confidence intervals for each average. For interested users, similar code can be used to produce teachers’ averages based on equal weights or distribution scores for each of the four categories. Research shows that while 95% confidence intervals are widely reported with statistics they are not completely appropriate when comparing averages of teachers’ scores. Goldstein and Healy (1993) recommend the use of 84% confidence intervals which would represent the equivalent of 1.41 standard deviations on each side of the means. Thus we calculated the new confidence intervals in a similar fashion as the 95% confidence intervals and reported both.

Appendix 2 shows the procedure necessary to generate the Excel code for cross-tabs to organize students’ assessment data by teacher, while Appendix 3 displays the code necessary to calculate each teacher’s final year score based on their students’ performance scores as the baseline. Appendix 4 shows the bootstrapping code used to obtain teachers’ confidence intervals associated with their end of year or end of course scores.

For Method 2, this study exemplifies several conditions considered as necessary prerequisites for hierarchical linear analyses. First we obtained the correlation coefficients among the variables included in the model with findings shown in Table 11.
We model the intercept and the slope of the first level equations as having randomly varying residuals. That means that the intercept and slope vary not only as a function of the pre-score as predictors but also as a function of a unique teacher effect. These residuals for intercept and for slopes are assumed sampled from a bivariate normal distribution.

In the outcome, the level-1 residuals are useful to check normality where the level 2 residuals are being used to rank teachers and to further interpret value-added. Running two conditional models assist in modeling the variability that remains unexplained in the intercepts and slopes across teachers. Therefore we can investigate what some of the teachers’ or some of the classes’ characteristics that may help explain such remaining variability in each model are.

As previously mentioned, it is important to note that teachers with large intercepts and small slopes will have the largest contributions to their students’ end of year performance scores conditioned on the scores that their students had when they entered the year. This implies that obtaining large class residuals (u_{0j}) and small residuals for other covariates included in the student-level equation (u_{1j}, u_{2j}, ..) will generate larger VA measures for those teachers. This is determined by ordering the teachers on their class residuals, (u_{0j}) to obtain a ranking scale system similar to the one we created when using the students’ ordinal scores.
VIII. Software for Growth Analysis

For the less sophisticated statistical models where we work with averages, Excel, R or SPSS packages were used to conduct exploratory analyses and to calculate the marginal frequencies or the expected growth rate and teacher ranking between time points.

However for the more complex models using multilevel linear equations we used SPSS, R and HLM (Bryk, Raudenbush, & Congdon, 1996), since model analyses required a more complex hierarchical structure. For practitioners who wish to explore other statistical software we should also mention MLwiN (Prosser, Rasbash, & Goldstein, 1996) as a useful package.
IX. Results

After running cross-tabs between students pre-test and post-test scores at the teacher level we used the first method to calculate the teachers’ averages with their respective 95\% and 84\% confidence intervals and rankings. In order to make our comparisons more meaningful teachers with students in three categorical scores (2, 3 and 4) were grouped together and separated from those with students’ scores falling into only two categories as well (3, and 4). The first group includes 12 of the 20 teachers since 8 of them have only students in the score categories 3 and 4 and had to be eliminated from the first VA analysis.

The second analysis includes teachers with students scoring in two of the four categories showing the corresponding counts for the two respective categories (3 and 4). In order to enlarge our sample and to also determine how they compare with their own performance when all three categories are being considered we included the 3-category score teachers with their counterparts with students’ scores in two-score categories only.

Results for teachers with students’ scores in three score categories (2, 3, and 4) from first analysis
Table 13: 3-Score Category Means for Teachers with 95% and 84% Confidence Intervals in Ascending Order

<table>
<thead>
<tr>
<th>Teacher ID</th>
<th>Teacher Code</th>
<th>Teacher Mean with State Weights</th>
<th>95% Confidence Interval for Means with State Weights for 3 score categories</th>
<th>84% Confidence Interval for Means with State Weights for 3 score categories</th>
</tr>
</thead>
<tbody>
<tr>
<td>13</td>
<td>07_I1</td>
<td>2.92 (2.87)</td>
<td>2.57</td>
<td>3.1</td>
</tr>
<tr>
<td>10</td>
<td>07_G2</td>
<td>3.03 (2.98)</td>
<td>2.82</td>
<td>3.14</td>
</tr>
<tr>
<td>11</td>
<td>07_H1</td>
<td>3.07 (3.02)</td>
<td>2.84</td>
<td>3.22</td>
</tr>
<tr>
<td>8</td>
<td>07_F2</td>
<td>3.24 (3.17)</td>
<td>3</td>
<td>3.34</td>
</tr>
<tr>
<td>4</td>
<td>07_C1</td>
<td>3.28 (3.21)</td>
<td>2.99</td>
<td>3.41</td>
</tr>
<tr>
<td>14</td>
<td>07_J1</td>
<td>3.34 (3.27)</td>
<td>3.04</td>
<td>3.5</td>
</tr>
<tr>
<td>16</td>
<td>07_K1</td>
<td>3.37 (3.30)</td>
<td>3.19</td>
<td>3.45</td>
</tr>
<tr>
<td>15</td>
<td>07_J2</td>
<td>3.42 (3.37)</td>
<td>3.15</td>
<td>3.52</td>
</tr>
<tr>
<td>2</td>
<td>07_B1</td>
<td>3.43 (3.35)</td>
<td>3.2</td>
<td>3.54</td>
</tr>
<tr>
<td>9</td>
<td>07_G1</td>
<td>3.48 (3.40)</td>
<td>3.21</td>
<td>3.56</td>
</tr>
<tr>
<td>7</td>
<td>07_E1</td>
<td>3.48 (3.40)</td>
<td>3.24</td>
<td>3.55</td>
</tr>
<tr>
<td>20</td>
<td>08_A1</td>
<td>3.57 (3.48)</td>
<td>3.12</td>
<td>3.74</td>
</tr>
</tbody>
</table>

* the bootstrapped means are shown in parentheses

Diagram 2 below shows a graphic representation of the teachers’ averages accompanied by their confidence intervals created in Excel to indicate overlapping 84% and 95% confidence intervals around the teachers’ averages for teachers with students’ scores in three categories only.
Diagram 2: Chart of Confidence Intervals with Included Teachers’ Class Averages at the End of
the Year and Overlapping 84% and 95% Confidence Intervals for Ordinal Scores (for 3-score
category teachers)

Based on the results from Diagram 2 we can say that teachers 9 (07_G1), 7 (07_E1) and 20
(08_A1) appear to have added more value to their students’ math education during the academic
year under analysis than teachers 13 (07_I1), 10 (07_G2), and 11 (07_H1). Not only these
teachers’ class averages are higher than their counterparts just mentioned but their confidence
intervals are not overlapping and appear to be greater than the corresponding confidence
intervals of their counterparts, teachers 13, 10 and 11.

Since the mean of the 1000 bootstrap samples will generally not equal but be very close to the
actual teachers’ class means we can claim that either the actual or the bootstrapped means could
be equally considered for VA analyses in our demonstrations. As shown in the results from this
section the numerical difference between the two means, the actual mean and the bootstrap mean is very small which leads to a small biased correction (Hesterberg, et al., 2003). While in bootstrapping the difference between the actual and the bootstrapped means is used to produce biased corrected averages or standard errors for various statistics this statistical adjustment is not performed here in order to keep the procedure simple and appealing to the field practitioners.

Results for teachers with students’ scores in two score categories (3 and 4)
A similar analysis was conducted for all teachers in the pool having students score in only two of the score categories, 3 and 4. In order to enlarge our data pool we included the teachers with students’ scores falling into score categories 2, 3 and 4 by eliminating their population who scored a 2 and keeping only those students who scored a 3 or a 4 for comparability reasons.
Table 14: 2-Score Category Means for Teachers with 95% and 84% Confidence Intervals in Ascending Order

<table>
<thead>
<tr>
<th>Teacher ID</th>
<th>Teacher Code</th>
<th>Teacher Mean with State Weights</th>
<th>84% Confidence Interval for Means with State Weights for 2 score categories</th>
<th>95% Confidence Interval for Means with State Weights for 2 score categories</th>
</tr>
</thead>
<tbody>
<tr>
<td>13</td>
<td>07_I1</td>
<td>3 (3.07)</td>
<td>(2.80,3.34)</td>
<td>(2.67,3.40)</td>
</tr>
<tr>
<td>10</td>
<td>07_G2</td>
<td>3.21 (3.24)</td>
<td>(3.04,3.42)</td>
<td>(2.90,3.50)</td>
</tr>
<tr>
<td>6</td>
<td>07_D2</td>
<td>3.23 (3.25)</td>
<td>(3.03,3.44)</td>
<td>(2.93,3.50)</td>
</tr>
<tr>
<td>1</td>
<td>07_A2</td>
<td>3.24 (3.26)</td>
<td>(3.03,3.48)</td>
<td>(2.90,3.58)</td>
</tr>
<tr>
<td>11</td>
<td>07_H1</td>
<td>3.26 (3.26)</td>
<td>(3.06,3.45)</td>
<td>(2.93,3.55)</td>
</tr>
<tr>
<td>8</td>
<td>07_F2</td>
<td>3.23 (3.26)</td>
<td>(3.04,3.47)</td>
<td>(2.86,3.53)</td>
</tr>
<tr>
<td>4</td>
<td>07_C1</td>
<td>3.33 (3.33)</td>
<td>(3.09,3.54)</td>
<td>(2.98,3.59)</td>
</tr>
<tr>
<td>18</td>
<td>07_L2</td>
<td>3.42 (3.40)</td>
<td>(3.19,3.60)</td>
<td>(3.07,3.68)</td>
</tr>
<tr>
<td>16</td>
<td>07_K1</td>
<td>3.44 (3.42)</td>
<td>(3.23,3.59)</td>
<td>(3.09,3.63)</td>
</tr>
<tr>
<td>3</td>
<td>07_B2</td>
<td>3.47 (3.44)</td>
<td>(3.24,3.68)</td>
<td>(3.12,3.70)</td>
</tr>
<tr>
<td>14</td>
<td>07_J1</td>
<td>3.48 (3.45)</td>
<td>(3.23,3.66)</td>
<td>(3.11,3.74)</td>
</tr>
<tr>
<td>2</td>
<td>07_B1</td>
<td>3.54 (3.50)</td>
<td>(3.31,3.68)</td>
<td>(3.2,3.74)</td>
</tr>
<tr>
<td>12</td>
<td>07_H2</td>
<td>3.55 (3.51)</td>
<td>(3.3,3.69)</td>
<td>(3.2,3.74)</td>
</tr>
<tr>
<td>17</td>
<td>07_K2</td>
<td>3.56 (3.51)</td>
<td>(3.28,3.71)</td>
<td>(3.19,3.76)</td>
</tr>
<tr>
<td>19</td>
<td>07_M2</td>
<td>3.61 (3.55)</td>
<td>(3.37,3.74)</td>
<td>(3.25,3.77)</td>
</tr>
<tr>
<td>7</td>
<td>07_E1</td>
<td>3.62 (3.57)</td>
<td>(3.38,3.74)</td>
<td>(3.26,3.77)</td>
</tr>
<tr>
<td>5</td>
<td>07_C2</td>
<td>3.62 (3.56)</td>
<td>(3.38,3.74)</td>
<td>(3.19,3.81)</td>
</tr>
<tr>
<td>9</td>
<td>07_G1</td>
<td>3.65 (3.59)</td>
<td>(3.40,3.75)</td>
<td>(3.27,3.79)</td>
</tr>
<tr>
<td>15</td>
<td>07_J2</td>
<td>3.66 (3.60)</td>
<td>(3.39,3.78)</td>
<td>(3.28,3.84)</td>
</tr>
<tr>
<td>20</td>
<td>08_A1</td>
<td>3.7 (3.62)</td>
<td>(3.3,3.86)</td>
<td>(3.16,4.0)</td>
</tr>
</tbody>
</table>

* the bootstrapped means are shown in parentheses

The results of 2-score categories show similar teacher rankings as the results obtained from teachers with 3-score categories. Teachers 13, 10 and 11 continue to rank at the lower end of the group while teachers 9, 7 and 20 continue to maintain high ranking with mean scores at the upper end of the spectrum. Despite the fairly small sample we are manipulating we can claim that teacher 20 added more value to his students’ math education during that year than teacher 13 did given that their 84% confidence intervals do not overlap.
2nd Method Results

Teachers who scored higher on the ranking produced by this method will be considered among those who added more unit values to their students' education during that year while those who ranked toward the bottom would be considered as having added fewer or even negative unit-values as compared to the average teacher. The remaining residuals related to the slopes will provide additional information about other factors included in the models.

When running the first and then the second conditional models in HLM, sigma squared slightly improved from 709.80538 to 703.53636 thus indicating a small improvement in the fit of the second model as compared to the first one.

When using the level 1 residuals to check the normality of the data through a Q-Q plot, the graph in Diagram 3 indicates overall normality of data with a slightly non-normal behavior for values located at the lower and at the higher ends of the spectrum. This behavior is perhaps due to floor and ceiling effects which could be further explored.

Diagram 3: Q-Q Plot to Check Normality for Students Continuous Scores
Reported below are also the final estimations of fixed effects and of variance components for both models from method 2, the model with pre-score as covariates and the second one including additional covariates.

It is interesting to notice that when adding further covariates to the level 1 or level 2 equations their effects on the post-test are not statistically significant (p-value>.05). For instance the p-values for covariates such as class size (.288), student SES (.250), white, Asian, Hispanic or black are all greater than .05 and thus not statistically significant in our population. Therefore, when calculating confidence intervals to rank teachers we will base our calculations on the first method estimates for teacher effects. These results also show that including the pre-score as a covariate accounts for the inherent effects that the other covariates such as ethnicity and SES have on the post-score.

<table>
<thead>
<tr>
<th>Fixed Effect</th>
<th>Coefficient</th>
<th>Standard Error</th>
<th>T-ratio</th>
<th>Approx. d.f.</th>
<th>P-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>INTRCPT1, $\beta_0$</td>
<td>691.268124</td>
<td>2.137946</td>
<td>323.333</td>
<td>19</td>
<td>0.000</td>
</tr>
<tr>
<td>INTRCPT2, $\gamma_{10}$</td>
<td>0.723340</td>
<td>0.047386</td>
<td>15.265</td>
<td>19</td>
<td>0.000</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Random Effect</th>
<th>Standard Deviation</th>
<th>Variance Component</th>
<th>df</th>
<th>Chi-square</th>
<th>P-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>INTRCPT1, $u_0$</td>
<td>6.99961</td>
<td>48.99459</td>
<td>19</td>
<td>46.29222</td>
<td>0.001</td>
</tr>
<tr>
<td>CENTERED slope, $u_1$</td>
<td>0.09120</td>
<td>0.00832</td>
<td>19</td>
<td>25.06507</td>
<td>0.158</td>
</tr>
<tr>
<td>level-1, r</td>
<td>26.64217</td>
<td>709.80538</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table 15: Model 1- Final estimation of fixed effects

Table 16: Model 1 - Final estimation of variance components
Table 17: Model 2 - Final estimation of fixed effects

<table>
<thead>
<tr>
<th>Fixed Effect</th>
<th>Coefficient</th>
<th>Standard Error</th>
<th>T-ratio</th>
<th>Approx. d.f.</th>
<th>P-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>For INTRCPT1, $\beta_0$</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>INTRCPT2, $\gamma_{00}$</td>
<td>710.243216</td>
<td>35.028143</td>
<td>20.276</td>
<td>18</td>
<td>0.000</td>
</tr>
<tr>
<td>CLASS_SI, $\gamma_{01}$</td>
<td>-1.065213</td>
<td>0.973075</td>
<td>-1.095</td>
<td>18</td>
<td>0.288</td>
</tr>
<tr>
<td>For CENTERED slope, $\beta_1$</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>INTRCPT2, $\gamma_{10}$</td>
<td>0.662332</td>
<td>0.047684</td>
<td>13.890</td>
<td>373</td>
<td>0.000</td>
</tr>
<tr>
<td>For ST_SES slope, $\beta_2$</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>INTRCPT2, $\gamma_{20}$</td>
<td>-4.976540</td>
<td>4.321452</td>
<td>-1.152</td>
<td>373</td>
<td>0.250</td>
</tr>
<tr>
<td>For WHITE slope, $\beta_3$</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>INTRCPT2, $\gamma_{30}$</td>
<td>5.714446</td>
<td>27.313657</td>
<td>0.209</td>
<td>373</td>
<td>0.834</td>
</tr>
<tr>
<td>For ASIAN slope, $\beta_4$</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>INTRCPT2, $\gamma_{40}$</td>
<td>12.830441</td>
<td>27.445216</td>
<td>0.467</td>
<td>373</td>
<td>0.640</td>
</tr>
<tr>
<td>For HISPANIC slope, $\beta_5$</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>INTRCPT2, $\gamma_{50}$</td>
<td>4.022553</td>
<td>27.297893</td>
<td>0.147</td>
<td>373</td>
<td>0.883</td>
</tr>
<tr>
<td>For BLACK slope, $\beta_6$</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>INTRCPT2, $\gamma_{60}$</td>
<td>-4.532511</td>
<td>27.656986</td>
<td>-0.164</td>
<td>373</td>
<td>0.870</td>
</tr>
<tr>
<td>For ST_LEP slope, $\beta_7$</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>INTRCPT2, $\gamma_{70}$</td>
<td>4.056866</td>
<td>5.454743</td>
<td>0.744</td>
<td>373</td>
<td>0.458</td>
</tr>
</tbody>
</table>

Table 18: Model 2 - Final estimation of variance components

<table>
<thead>
<tr>
<th>Random Effect</th>
<th>Standard Deviation</th>
<th>Variance Component</th>
<th>df</th>
<th>Chi-square</th>
<th>P-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>INTRCPT1, $u_0$</td>
<td>9.04976</td>
<td>81.89820</td>
<td>18</td>
<td>59.11059</td>
<td>0.000</td>
</tr>
<tr>
<td>level-1, $r$</td>
<td>26.52426</td>
<td>703.53636</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
In order to rank teachers and to ensure comparability with results from our first method we will use 84% confidence intervals. Thus from the second level residual file we extract the variance of the intercept (PV0_0) and take its square root in order to determine the appropriate standard error for each teacher. We then add and subtract approximately 1.4 such standard errors to and from each teacher intercept (EB_{intercept} or \( \mu_0 \)) to find the teachers’ confidence intervals around the means.

Model 1 estimates of teachers’ scores

The teachers’ intercepts are either positive or negative since they are all centered around the students’ math assessment state average for year 1 (2006-2007 academic year).

Teachers whose confidence intervals are overall greater than others are classified as those who added more value to their students’ math education during that academic year as compared to their counterparts. In our example teacher 7 (coded as 07_E1) had superior results when compared to teachers 11 and 13 (coded as 07_A2, or 07_B2) or others with non-overlapping lower confidence intervals.
Table 19: Results of Pre-score only model with 84% confidence intervals

<table>
<thead>
<tr>
<th>Teacher Recoded</th>
<th>Teacher</th>
<th>Number of Students</th>
<th>EB Intercept</th>
<th>Lower Limit of CI for EB Intercept</th>
<th>Upper Limit of CI for EB Intercept</th>
</tr>
</thead>
<tbody>
<tr>
<td>6</td>
<td>07_D2</td>
<td>22</td>
<td>-7.557</td>
<td>-12.8361</td>
<td>-2.27787</td>
</tr>
<tr>
<td>11</td>
<td>07_H1</td>
<td>21</td>
<td>-6.286</td>
<td>-12.35</td>
<td>-0.22202</td>
</tr>
<tr>
<td>12</td>
<td>07_H2</td>
<td>22</td>
<td>-5.6</td>
<td>-10.9153</td>
<td>-0.28466</td>
</tr>
<tr>
<td>13</td>
<td>07_I1</td>
<td>16</td>
<td>-5.38</td>
<td>-12.3985</td>
<td>1.638487</td>
</tr>
<tr>
<td>1</td>
<td>07_A2</td>
<td>17</td>
<td>-5.304</td>
<td>-11.1166</td>
<td>0.508553</td>
</tr>
<tr>
<td>20</td>
<td>07_G2</td>
<td>22</td>
<td>-5.291</td>
<td>-10.5649</td>
<td>-0.01715</td>
</tr>
<tr>
<td>5</td>
<td>07_C2</td>
<td>24</td>
<td>-2.378</td>
<td>-7.6009</td>
<td>2.844903</td>
</tr>
<tr>
<td>4</td>
<td>07_C1</td>
<td>19</td>
<td>-1.562</td>
<td>-7.98021</td>
<td>4.856211</td>
</tr>
<tr>
<td>15</td>
<td>07_J2</td>
<td>25</td>
<td>-1.55</td>
<td>-6.84699</td>
<td>3.746985</td>
</tr>
<tr>
<td>8</td>
<td>07_F2</td>
<td>22</td>
<td>-1.457</td>
<td>-7.17333</td>
<td>4.259329</td>
</tr>
<tr>
<td>18</td>
<td>07_L2</td>
<td>16</td>
<td>-0.821</td>
<td>-6.91099</td>
<td>5.268992</td>
</tr>
<tr>
<td>3</td>
<td>07_B2</td>
<td>20</td>
<td>-0.196</td>
<td>-5.87412</td>
<td>5.482117</td>
</tr>
<tr>
<td>19</td>
<td>07_M2</td>
<td>18</td>
<td>0.687</td>
<td>-5.41669</td>
<td>6.790688</td>
</tr>
<tr>
<td>2</td>
<td>07_B1</td>
<td>17</td>
<td>0.932</td>
<td>-6.13827</td>
<td>8.002275</td>
</tr>
<tr>
<td>17</td>
<td>07_K2</td>
<td>18</td>
<td>2.63</td>
<td>-3.1436</td>
<td>8.403601</td>
</tr>
<tr>
<td>16</td>
<td>07_K1</td>
<td>19</td>
<td>3.345</td>
<td>-3.05367</td>
<td>9.743666</td>
</tr>
<tr>
<td>9</td>
<td>07_G1</td>
<td>21</td>
<td>4.639</td>
<td>-1.52546</td>
<td>10.80346</td>
</tr>
<tr>
<td>14</td>
<td>07_J1</td>
<td>23</td>
<td>6.532</td>
<td>0.330083</td>
<td>12.73392</td>
</tr>
<tr>
<td>20</td>
<td>08_A1</td>
<td>16</td>
<td>11.994</td>
<td>4.837087</td>
<td>19.15091</td>
</tr>
<tr>
<td>7</td>
<td>07_E1</td>
<td>22</td>
<td>12.622</td>
<td>6.493445</td>
<td>18.75055</td>
</tr>
</tbody>
</table>

A graphical representation of the teachers’ ranking based on their error variance of the intercept in the level 2 HLM equations generated by using an Excel plot makes the results from Table 18 clearer when evaluating value-added by teacher as shown in the ranking from Diagram 4 below.
Diagram 4: Teacher Ranking based on PreScore Model only for Continuous Scores
Comparability of Results

After comparing teachers’ ranking under by using both methods as shown in Table 19 below, under both scoring systems it appears that the ranking of the confidence intervals and of the teachers’ averages from method 1 for teachers whose students entered the year in the 3-score category model are very similar to the ranking based on their error variance of the intercept in the level 2 equation (see Diagrams 3 and 4). Teachers 7 and 20 are outperforming teachers 11, 13 and 20 under both methods not only when comparing averages but also when comparing them by using the non-overlapping confidence intervals against their counterparts with lower averages for the 3-score method. When using 84% confidence intervals which are smaller in range than 95% confidence intervals the ranking among teachers is even more pronounced as fewer 84% confidence intervals overlap in length as compared to the standard 95% confidence intervals.

As shown in Table 18, teachers’ averages produced by using method 1 for 2-score and 3-score scenarios are highly correlated and statistically significant at the 0.01 level with teachers’ scores obtained through using the 2nd method or the multilevel model.

Table 20: Teachers’ 2- and 3- score Average Correlations

<table>
<thead>
<tr>
<th>Correlations</th>
<th>3 scores averages with state weights</th>
<th>2 score averages with State Weights</th>
<th>EB Intercept1</th>
</tr>
</thead>
<tbody>
<tr>
<td>3 scores averages with state weights</td>
<td>Pearson Correlation</td>
<td>1</td>
<td>.960**</td>
</tr>
<tr>
<td></td>
<td>Sig. (2-tailed)</td>
<td>12</td>
<td>.000</td>
</tr>
<tr>
<td></td>
<td>N</td>
<td>12</td>
<td>12</td>
</tr>
<tr>
<td>2 score averages with State Weights</td>
<td>Pearson Correlation</td>
<td>.960**</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>Sig. (2-tailed)</td>
<td>.000</td>
<td>20</td>
</tr>
<tr>
<td></td>
<td>N</td>
<td>12</td>
<td>20</td>
</tr>
<tr>
<td>EB Intercept1</td>
<td>Pearson Correlation</td>
<td>.823**</td>
<td>.646**</td>
</tr>
<tr>
<td></td>
<td>Sig. (2-tailed)</td>
<td>.001</td>
<td>.002</td>
</tr>
<tr>
<td></td>
<td>N</td>
<td>12</td>
<td>20</td>
</tr>
</tbody>
</table>

**. Correlation is significant at the 0.01 level (2-tailed).
Table 21: Side-by-side Mean Comparisons for Teachers with Students in 2- and 3-score categories

<table>
<thead>
<tr>
<th>Teacher Recoded</th>
<th>Teacher</th>
<th>Number of Students</th>
<th>EB Intercept</th>
<th>Lower Limit of CI for EB Intercept</th>
<th>Upper Limit of CI for EB Intercept</th>
<th>3 scores averages with state weights</th>
<th>95% CI for 3-scores averages</th>
<th>94% Confidence Interval for Means with State Weights for 3 score categories</th>
<th>2 score averages with State Weights</th>
<th>95% Confidence Interval for Means with State Weights for 2 score categories</th>
<th>94% Confidence Interval for Means with State Weights for 2 score categories</th>
</tr>
</thead>
<tbody>
<tr>
<td>6</td>
<td>07_D2</td>
<td>22</td>
<td>-7.557</td>
<td>-12.8361</td>
<td>-2.27787</td>
<td>3.07</td>
<td>(2.84,3.22)</td>
<td>(2.89,3.15)</td>
<td>3.23</td>
<td>(2.93,3.50)</td>
<td>(3.03,3.44)</td>
</tr>
<tr>
<td>11</td>
<td>07_H1</td>
<td>21</td>
<td>-6.286</td>
<td>-12.35</td>
<td>-0.22202</td>
<td>3</td>
<td>(2.89,3.21)</td>
<td>(2.92,3.18)</td>
<td>3.26</td>
<td>(2.93,3.55)</td>
<td>(3.06,3.45)</td>
</tr>
<tr>
<td>12</td>
<td>07_H2</td>
<td>22</td>
<td>-5.6</td>
<td>-10.9153</td>
<td>-0.28466</td>
<td>2.92</td>
<td>(2.57,3.1)</td>
<td>(2.67,3.06)</td>
<td>3.24</td>
<td>(2.90,3.58)</td>
<td>(3.03,3.48)</td>
</tr>
<tr>
<td>13</td>
<td>07_I1</td>
<td>16</td>
<td>-5.38</td>
<td>-12.3985</td>
<td>1.638487</td>
<td>2.32</td>
<td>(2.57,3.1)</td>
<td>(2.67,3.06)</td>
<td>3.21</td>
<td>(2.90,3.50)</td>
<td>(3.04,3.42)</td>
</tr>
<tr>
<td>1</td>
<td>07_A2</td>
<td>17</td>
<td>-5.304</td>
<td>-11.166</td>
<td>0.508553</td>
<td>0.03</td>
<td>(2.82,3.14)</td>
<td>(2.85,3.11)</td>
<td>3.62</td>
<td>(3.19,3.81)</td>
<td>(3.38,3.74)</td>
</tr>
<tr>
<td>10</td>
<td>07_G2</td>
<td>22</td>
<td>-5.291</td>
<td>-10.5649</td>
<td>-0.01715</td>
<td>3</td>
<td>(2.82,3.14)</td>
<td>(2.85,3.11)</td>
<td>3.24</td>
<td>(2.90,3.58)</td>
<td>(3.03,3.48)</td>
</tr>
<tr>
<td>5</td>
<td>07_C2</td>
<td>24</td>
<td>-2.378</td>
<td>-7.6009</td>
<td>2.844903</td>
<td>3.03</td>
<td>(2.82,3.14)</td>
<td>(2.85,3.11)</td>
<td>3.21</td>
<td>(2.90,3.50)</td>
<td>(3.04,3.42)</td>
</tr>
<tr>
<td>4</td>
<td>07_C1</td>
<td>19</td>
<td>-1.562</td>
<td>-7.98021</td>
<td>4.856211</td>
<td>3.28</td>
<td>(2.99,3.41)</td>
<td>(3.05,3.37)</td>
<td>3.33</td>
<td>(2.98,3.59)</td>
<td>(3.09,3.54)</td>
</tr>
<tr>
<td>15</td>
<td>07_J2</td>
<td>25</td>
<td>-1.55</td>
<td>-6.84699</td>
<td>3.746985</td>
<td>3.42</td>
<td>(3.15,3.52)</td>
<td>(3.22,3.47)</td>
<td>3.66</td>
<td>(3.28,3.84)</td>
<td>(3.39,3.78)</td>
</tr>
<tr>
<td>8</td>
<td>07_F2</td>
<td>22</td>
<td>-1.457</td>
<td>-7.17333</td>
<td>4.259329</td>
<td>3.24</td>
<td>(3.34)</td>
<td>(3.04,3.30)</td>
<td>3.23</td>
<td>(2.86,3.53)</td>
<td>(3.04,3.47)</td>
</tr>
<tr>
<td>18</td>
<td>07_L2</td>
<td>16</td>
<td>-0.821</td>
<td>-6.91099</td>
<td>5.268992</td>
<td>3.42</td>
<td>(3.07,3.68)</td>
<td>(3.19,3.60)</td>
<td>3.47</td>
<td>(3.12,3.70)</td>
<td>(3.24,3.68)</td>
</tr>
<tr>
<td>3</td>
<td>07_B2</td>
<td>20</td>
<td>-0.196</td>
<td>-5.87412</td>
<td>5.482117</td>
<td>3.47</td>
<td>(3.20,3.54)</td>
<td>(3.23,3.48)</td>
<td>3.54</td>
<td>(3.23,3.74)</td>
<td>(3.31,3.68)</td>
</tr>
<tr>
<td>19</td>
<td>07_M2</td>
<td>18</td>
<td>0.687</td>
<td>-5.41669</td>
<td>6.790688</td>
<td>3.47</td>
<td>(3.20,3.54)</td>
<td>(3.23,3.48)</td>
<td>3.54</td>
<td>(3.23,3.74)</td>
<td>(3.31,3.68)</td>
</tr>
<tr>
<td>2</td>
<td>07_B1</td>
<td>17</td>
<td>0.932</td>
<td>-6.13827</td>
<td>8.002275</td>
<td>3.43</td>
<td>(3.19,3.45)</td>
<td>(3.21,3.41)</td>
<td>3.44</td>
<td>(3.09,3.63)</td>
<td>(3.23,3.59)</td>
</tr>
<tr>
<td>17</td>
<td>07_K2</td>
<td>18</td>
<td>2.63</td>
<td>-3.1436</td>
<td>8.403601</td>
<td>3.37</td>
<td>(3.19,3.45)</td>
<td>(3.21,3.41)</td>
<td>3.44</td>
<td>(3.09,3.63)</td>
<td>(3.23,3.59)</td>
</tr>
<tr>
<td>16</td>
<td>07_K1</td>
<td>19</td>
<td>3.345</td>
<td>-3.05367</td>
<td>9.743666</td>
<td>3.44</td>
<td>(3.21,3.56)</td>
<td>(3.28,3.53)</td>
<td>3.65</td>
<td>(3.27,3.79)</td>
<td>(3.40,3.75)</td>
</tr>
<tr>
<td>9</td>
<td>07_G1</td>
<td>21</td>
<td>4.639</td>
<td>-1.52546</td>
<td>10.80346</td>
<td>3.48</td>
<td>(3.21,3.56)</td>
<td>(3.28,3.53)</td>
<td>3.65</td>
<td>(3.27,3.79)</td>
<td>(3.40,3.75)</td>
</tr>
<tr>
<td>14</td>
<td>07_J1</td>
<td>23</td>
<td>6.532</td>
<td>0.330083</td>
<td>12.73392</td>
<td>3.34</td>
<td>(3.04,3.50)</td>
<td>(3.10,3.43)</td>
<td>3.48</td>
<td>(3.11,3.74)</td>
<td>(3.23,3.66)</td>
</tr>
<tr>
<td>20</td>
<td>08_A1</td>
<td>16</td>
<td>11.994</td>
<td>4.837087</td>
<td>19.15091</td>
<td>3.57</td>
<td>(3.12,3.74)</td>
<td>(3.21,3.72)</td>
<td>3.7</td>
<td>(3.16,4.0)</td>
<td>(3.33,3.86)</td>
</tr>
</tbody>
</table>
Table 22: Side-by-side Mean Comparisons for teachers with students in 2- and 3-score categories with unadjusted end of year average included

<table>
<thead>
<tr>
<th>Teacher Recoded</th>
<th>Teacher</th>
<th>Number of Students</th>
<th>EB Intercept1</th>
<th>Lower Limit of CI for EB Intercept1</th>
<th>Upper Limit of CI for EB Intercept1</th>
<th>3 scores averages with state weights</th>
<th>95% CI for 3-scores averages</th>
<th>84% Confidence Interval for Means with State Weights for 3 score categories</th>
<th>2 score averages with State Weights</th>
<th>95% Confidence Interval for Means with State Weights for 2 score categories</th>
<th>84% Confidence Interval for Means with State Weights for 2 score categories</th>
<th>Unadjusted average at the end of year 1</th>
</tr>
</thead>
<tbody>
<tr>
<td>6 07_D2</td>
<td>22</td>
<td>-7.557</td>
<td>-12.8361</td>
<td>-2.27787</td>
<td>3.07</td>
<td>(2.84,3.22)</td>
<td>(2.89,3.15)</td>
<td>3.26</td>
<td>(2.93,3.55)</td>
<td>(3.06,3.45)</td>
<td>3.14</td>
<td>3.41</td>
</tr>
<tr>
<td>11 07_H1</td>
<td>21</td>
<td>-6.286</td>
<td>-12.35</td>
<td>-0.22202</td>
<td>3.07</td>
<td>(2.84,3.22)</td>
<td>(2.89,3.15)</td>
<td>3.26</td>
<td>(2.93,3.55)</td>
<td>(3.06,3.45)</td>
<td>3.14</td>
<td>3.41</td>
</tr>
<tr>
<td>12 07_H2</td>
<td>22</td>
<td>-5.6</td>
<td>-10.9153</td>
<td>-0.28466</td>
<td>2.92</td>
<td>(2.57,3.1)</td>
<td>(2.67,3.06)</td>
<td>3.24</td>
<td>(2.90,3.58)</td>
<td>(3.03,3.48)</td>
<td>3.41</td>
<td>3.14</td>
</tr>
<tr>
<td>13 07_I1</td>
<td>16</td>
<td>-5.38</td>
<td>-12.3985</td>
<td>1.638487</td>
<td>2.92</td>
<td>(2.57,3.1)</td>
<td>(2.67,3.06)</td>
<td>3.24</td>
<td>(2.90,3.58)</td>
<td>(3.03,3.48)</td>
<td>3.41</td>
<td>3.14</td>
</tr>
<tr>
<td>1 07_A2</td>
<td>17</td>
<td>-5.304</td>
<td>-11.1166</td>
<td>0.508553</td>
<td>3.03</td>
<td>(2.82,3.14)</td>
<td>(2.85,3.11)</td>
<td>3.21</td>
<td>(2.90,3.50)</td>
<td>(3.04,3.42)</td>
<td>3.27</td>
<td>3.41</td>
</tr>
<tr>
<td>10 07_G2</td>
<td>22</td>
<td>-5.291</td>
<td>-10.5649</td>
<td>-0.01715</td>
<td>3.03</td>
<td>(2.82,3.14)</td>
<td>(2.85,3.11)</td>
<td>3.21</td>
<td>(2.90,3.50)</td>
<td>(3.04,3.42)</td>
<td>3.27</td>
<td>3.41</td>
</tr>
<tr>
<td>5 07_C2</td>
<td>24</td>
<td>-2.378</td>
<td>-7.6009</td>
<td>2.844903</td>
<td>3.03</td>
<td>(2.82,3.14)</td>
<td>(2.85,3.11)</td>
<td>3.21</td>
<td>(2.90,3.50)</td>
<td>(3.04,3.42)</td>
<td>3.27</td>
<td>3.41</td>
</tr>
<tr>
<td>4 07_C1</td>
<td>19</td>
<td>-1.562</td>
<td>-7.98021</td>
<td>4.856211</td>
<td>3.03</td>
<td>(2.82,3.14)</td>
<td>(2.85,3.11)</td>
<td>3.21</td>
<td>(2.90,3.50)</td>
<td>(3.04,3.42)</td>
<td>3.27</td>
<td>3.41</td>
</tr>
<tr>
<td>8 07_F2</td>
<td>22</td>
<td>-1.457</td>
<td>-7.17333</td>
<td>4.259329</td>
<td>3.24</td>
<td>(3.34,3.34)</td>
<td>(3.04,3.30)</td>
<td>3.23</td>
<td>(2.86,3.53)</td>
<td>(3.04,3.47)</td>
<td>3.32</td>
<td>3.41</td>
</tr>
<tr>
<td>18 07_L2</td>
<td>16</td>
<td>-0.821</td>
<td>-6.91099</td>
<td>5.268992</td>
<td>3.42</td>
<td>(3.07,3.68)</td>
<td>(3.19,3.60)</td>
<td>3.44</td>
<td>(3.24,3.68)</td>
<td>(3.24,3.68)</td>
<td>3.5</td>
<td>3.41</td>
</tr>
<tr>
<td>3 07_B2</td>
<td>20</td>
<td>-0.196</td>
<td>-5.87412</td>
<td>5.482117</td>
<td>3.47</td>
<td>(3.12,3.70)</td>
<td>(3.24,3.68)</td>
<td>3.5</td>
<td>(3.24,3.68)</td>
<td>(3.24,3.68)</td>
<td>3.5</td>
<td>3.41</td>
</tr>
<tr>
<td>19 07_M2</td>
<td>18</td>
<td>0.687</td>
<td>-5.41669</td>
<td>6.790688</td>
<td>3.61</td>
<td>(3.25,3.77)</td>
<td>(3.37,3.74)</td>
<td>3.61</td>
<td>(3.37,3.74)</td>
<td>(3.37,3.74)</td>
<td>3.61</td>
<td>3.41</td>
</tr>
<tr>
<td>2 07_B1</td>
<td>17</td>
<td>0.932</td>
<td>-6.13827</td>
<td>8.002275</td>
<td>3.43</td>
<td>(3.20,3.54)</td>
<td>(3.23,3.48)</td>
<td>3.54</td>
<td>(3.2,3.74)</td>
<td>(3.31,3.68)</td>
<td>3.24</td>
<td>3.41</td>
</tr>
<tr>
<td>17 07_K2</td>
<td>18</td>
<td>2.63</td>
<td>-3.1436</td>
<td>8.403601</td>
<td>3.43</td>
<td>(3.20,3.54)</td>
<td>(3.23,3.48)</td>
<td>3.54</td>
<td>(3.2,3.74)</td>
<td>(3.31,3.68)</td>
<td>3.24</td>
<td>3.41</td>
</tr>
<tr>
<td>16 07_K1</td>
<td>19</td>
<td>3.345</td>
<td>-3.05367</td>
<td>9.743666</td>
<td>3.37</td>
<td>(3.19,3.45)</td>
<td>(3.21,3.41)</td>
<td>3.44</td>
<td>(3.09,3.63)</td>
<td>(3.23,3.59)</td>
<td>3.32</td>
<td>3.41</td>
</tr>
<tr>
<td>9 07_G1</td>
<td>21</td>
<td>4.639</td>
<td>-1.52546</td>
<td>10.80346</td>
<td>3.48</td>
<td>(3.21,3.56)</td>
<td>(3.28,3.53)</td>
<td>3.65</td>
<td>(3.27,3.79)</td>
<td>(3.40,3.75)</td>
<td>3.48</td>
<td>3.41</td>
</tr>
<tr>
<td>14 07_J1</td>
<td>23</td>
<td>6.532</td>
<td>0.330883</td>
<td>12.73392</td>
<td>3.34</td>
<td>(3.04,3.50)</td>
<td>(3.10,3.43)</td>
<td>3.48</td>
<td>(3.11,3.74)</td>
<td>(3.23,3.66)</td>
<td>3.35</td>
<td>3.41</td>
</tr>
<tr>
<td>20 08_A1</td>
<td>16</td>
<td>11.994</td>
<td>4.837087</td>
<td>19.15091</td>
<td>3.57</td>
<td>(3.12,3.74)</td>
<td>(3.21,3.72)</td>
<td>3.7</td>
<td>(3.16,4)</td>
<td>(3.3,3.86)</td>
<td>3.38</td>
<td>3.41</td>
</tr>
</tbody>
</table>
In order to determine how teachers’ averages at the end of year when we do not include any covariates or unadjusted averages compare to their adjusted averages at the end of year 1 which include covariates, we added a column showing the straight unadjusted average for each teacher at the end of year 1 and further correlated that with the same teachers adjusted averages. As shown in Diagram 5 there does not appear to be any correlation between the adjusted and the unadjusted teachers’ mean at the end of year 1. This is further proof that a VA approach with pre-score adjustment is a better way of ranking teachers than simpler straight averages.

Diagram 5:

The unadjusted averages for each of the twenty teachers do not appear to follow any pattern being sometimes below and sometimes above the adjusted state averages for each teacher (shown in Diagram 6). Therefore the ranking we determined by using the value-added methods discussed is quite different as compared to the one resulting from teachers unadjusted straight averages.
When looking at teachers adjusted averages at the end of year 1, we noticed that teachers raking at the lowest (teachers 6, 11 and 12) and also at the highest (teachers 14, 7, and 20) scores range based on the ranking produced by the models for value-added appear to score consistently under both methods proposed by our study while the rest of the teachers perform in the middle category.

The correlations among the methods results are decent but theoretically we expect them to improve with larger samples. In fact, if one would use two to three years data for the same teacher in order to enlarge the sample scores at the teacher level we would expect to see higher correlations.
Diagram 7: Raw Means Diagram for Teacher Rankings with Scores in Two Categories

Diagram 8: Bootstrapped Means Diagram for Teacher Rankings with Scores in Two Categories
X. Findings and Limitations

When comparing teachers’ value-added scores using both proposed methods, the teachers’ ranking based on their averages and confidence intervals appear to be consistent across methods unlike when using teachers’ unadjusted averages at the end of year 1 for ranking.

Therefore we recommend the use of the first method over the second one, since the first one is more accessible and simpler to program using basic statistical software packages such as Excel. Yet, the first method requires the use of R code to calculate the errors used to determine the confidence intervals for each teacher. To make it easier for interested practitioners we make this R code available in the appendix.

From the results of this study we can identify three groups of teachers, those clustered at the bottom, in the middle and at the top of the overall score ranges. Thus, we could identify these ranking teacher groups as being effective, developing and ineffective. If more score categories are desired a larger sample of schools will be required for statistical analyses. Smaller samples such as the one used in this research make more challenging the finding of clear cut-off scores especially when trying to classify teachers in three or four performance categories. For this instance and for instances related to statistical power we recognize that our small sample of 20 teachers and 400 students is a limitation of our study.

For larger samples one could identify three boundaries or cut-off points in order to rank teachers in four different categories as required by current NYS education policy related to APPR plans, highly effective, effective, developing and ineffective (NYSED, 2012).
Another limitation of the study is conducting statistical analyses on the scoring scale produced by NYSED without conducting transformations to eliminate ceiling and floor effects especially for students scoring in the 1 and 4 categories for method 1. Practitioners or other researchers who wish to replicate this study may want to consider transforming the data especially for students who score at the upper end of the scale, in the 4’s in order to manage the ceiling effects more appropriately and to identify whether or not they could be significant enough to address.

One other limitation of our study consists of using only math pre-test score as a covariate in determining teachers’ ranking based on their post-test math scores. We would encourage practitioners and other researchers to possibly include students’ NYS pre-test scores in ELA (English Language Arts) to the extent available when determining teachers’ value-added.
XI. The practical significance of this study and further research

In order to complement the information provided by state assessments, school districts and other testing centers developed proprietary assessments attempting to calculate value-added over time. Such systems may measure students’ performance several times a year, in addition to the state assessments to provide access to more data and to produce more accurate estimations of students’ growth. Yet, such systems continue to remain fairly expensive and certain schools will not be able to afford them.

While the value-added systems are superior to mean-achievement approaches it is at times more difficult for practitioners to interpret their results in context or to establish the effect of value-added as causal given the complexity of the statistical analyses, the measurement scales used or their alignment to content and to subject measurement across years or, in some cases for absence of expertise in the organization. Researchers who have used school data sets for growth analyses express their findings in technical terms, which practitioners may not understand or may not find easy to interpret.

Currently there is very little research about how NYS assessment data for grades 3-8 in how math is used by the local administrations of schools and of school districts as well as by how subject-area teachers link it to curriculum and/ or to their instruction. NYS and local BOCES institutions compile the testing data in these subjects for grades 3-8 and provide school districts with descriptive statistics. In the absence of other tools, these statistical summaries are what districts may choose to use to inform curriculum and instruction or other staff development efforts. In most cases, data are collected, distributed to parents, or teachers, schools are judged
based on performance or students’ growth, average results by subject are published in the local media and then, all may go on a shelf.

To our knowledge, the new implementation of APPR regulations in New York State scheduled for September 2011, which allow for implementation of local growth or value-added models find many practitioners lacking expertise in assessing student and educators’ growth. More recently, NYS Board of Regents voted to adopt Harold Doran’s growth model to evaluate teacher and principal yearly performance based on their students’ growth. (Lissitz & Doran, 2009)

Some are looking for simpler resources necessary to tackle complex mathematical structures while others find themselves overwhelmed or ready to compromise for any available model. For-profit companies offer their services to school districts, and in exchange for a fee they conduct data analyses by using growth models tailored to available data in a specific state. However, recent reduction in budgets place significant restrictions on money available for outside consulting.

We thus conclude that there is an abundant need for developing local expertise in analyzing and interpreting teacher value-added and in communicating such results to various stakeholders. The first step in developing local growth expertise may be through understanding simpler statistical models which could yield results close to those of more sophisticated multilevel models as in method 2 of our study.

From the data that school districts have as a result of New York State assessments, this research aims to provide accessible ways for school districts or for local practitioners to measure growth or value-added through one simpler and one more complex approach, and to develop or to use existing code necessary to run the hierarchical models. We further showed detailed
explanations, examples and the necessary code for each method in order to offer the practitioner the necessary tools to adapt these methods to their local needs when calculating teachers’ value added scores as part of a composite yearly score.

To account for false positives and for false negatives we evaluated and compared the results of both methods and determined that teachers’ ranking remain consistent under each of the two methods using adjusted teacher averages. Further we discussed their advantages and limitations in light of current policy context.

More research should be conducted on larger data sets with teachers’ data from multiple years and possibly from multiple cohorts in order to identify more categories of teachers’ performance and to test if results/rankings from smaller samples hold true when samples become larger.

Additionally, interested researchers may want to replicate our model by including students’ math and English Language Arts NYS assessments scores to see how they both contribute to determining teachers’ value-added.

We should also note that while ordinal measures are somewhat coarse in nature making it difficult to monitor smaller changes in one’s performance, with larger sample sizes one may want to further research how inclusion of half-level scores may impact the calculations of teacher and school value-added models.

Researchers as well as educators who wish to begin using a growth model in analyzing their own students’ data in a smaller setting should find these models accessible enough to begin implementing or replicating this methodology without great difficulty. This document will hopefully constitute a guide for local practitioners who wish to develop local expertise in understanding and modeling with greater accuracy their students’ growth from local or from
state assessment data to produce additional feedback to curriculum and instruction. This study may also constitute a good resource for districts who wish to further develop their own data teams.
Appendix 1 – SPSS Sorting Feature and Format Change

SPSS long format into wide format

Restructure the data set from the long format into the wide format:

SORT CASES BY Student_ID Year.

CASESTOVARS
/ID=Student_ID
/INDEX=Year
/GROUPBY=INDEX.

Sort database by Teacher ID:

DATASET ACTIVATE DataSet1.

SORT CASES BY rT_ID.

SPLIT FILE SEPARATE BY rT_ID.

SAVE OUTFILE='D:\Dissertation\wideDissert.data.12122011.sav'
/COMPRESSED.

DATASET ACTIVATE DataSet1.

SORT CASES BY rT_ID.

SPLIT FILE SEPARATE BY rT_ID.

Compare teachers’ scores between years 1 and 0 by creating cross tabs:

CROSSTABS

/TABLES=r_St_Level.0 BY r_St_Level.1 BY rT_ID.1

/FORMAT=AVALUE TABLES

/CELLS=COUNT

/COUNT ROUND CELL.
Appendix 2 – Excel Procedure to Generate Counts for Cross Tables

In Excel there are at least two options useful to generate counts for cross tabs also known as pivotal tables. The first option is to select the Pivotal Table and highlight the necessary range of data for which to generate these counts. The second option, simpler in a sense uses bin arrays, in our case represented by the ordinal scores, 1, 2, 3, and 4. Anywhere on the spreadsheet, we can display these value scores, and then highlight the cells next to the bin arrays. In the command line, use code “=FREQUENCY(DATA ARRAY,BINS ARRAY)”, and then select the section of the table showing the students’ scores for the teacher we want the counts for. This could be used to find the total counts by teacher at the beginning and also at the end of the year.

SPSS or SAS also offer ways of creating cross tabs from the menu embedded in each software or by using code. Yet, since they are not free to download we will not display that code here.

Appendix 3 – NYS Math Weights used in Method 1 Procedure

Below are the state weights used in the Excel program for calculating teachers’ weighted scores.

<table>
<thead>
<tr>
<th>Grade</th>
<th>N-Count</th>
<th>Level 1</th>
<th>Level II</th>
<th>Level III</th>
<th>Level IV</th>
</tr>
</thead>
<tbody>
<tr>
<td>3</td>
<td>200071</td>
<td>4.09</td>
<td>10.61</td>
<td>55.97</td>
<td>29.33</td>
</tr>
<tr>
<td>4</td>
<td>1999181</td>
<td>6.02</td>
<td>13.97</td>
<td>52.52</td>
<td>27.49</td>
</tr>
<tr>
<td>5</td>
<td>203670</td>
<td>5.78</td>
<td>18.01</td>
<td>54.10</td>
<td>22.11</td>
</tr>
<tr>
<td>6</td>
<td>205976</td>
<td>8.71</td>
<td>19.94</td>
<td>51.33</td>
<td>20.02</td>
</tr>
<tr>
<td>7</td>
<td>213165</td>
<td>7.46</td>
<td>26.06</td>
<td>48.13</td>
<td>18.35</td>
</tr>
<tr>
<td>8</td>
<td>215108</td>
<td>12.21</td>
<td>28.90</td>
<td>46.97</td>
<td>11.92</td>
</tr>
</tbody>
</table>

Sample:

<table>
<thead>
<tr>
<th>Last year</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>Sum</th>
</tr>
</thead>
<tbody>
<tr>
<td>4</td>
<td>0</td>
<td>1</td>
<td>2</td>
<td>7</td>
<td>10</td>
</tr>
<tr>
<td>3</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>0</td>
<td>6</td>
</tr>
<tr>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>0</td>
<td>6</td>
</tr>
<tr>
<td>1</td>
<td>2</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>3</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Conditional row proportions</th>
<th>Equal Wts</th>
<th>State Wts</th>
</tr>
</thead>
<tbody>
<tr>
<td>4 0 0.1 0.2 0.7 3.6</td>
<td>0.25 0.27</td>
<td></td>
</tr>
<tr>
<td>3 0.166667 0.333333 0.5 0</td>
<td>2.333333</td>
<td>0.25 0.53</td>
</tr>
<tr>
<td>2 0.333333 0.333333 0.333333 0</td>
<td>2</td>
<td>0.25 0.14</td>
</tr>
<tr>
<td>1 0.666667 0.333333 0 0</td>
<td>1.333333</td>
<td>0.25 0.06</td>
</tr>
</tbody>
</table>

Equal Weights Average: 2.316667
State weights average: 2.568667
Appendix 4 – Bootstrap Procedure for 2- and 3- score estimates

# Calculator for 2 score categories to find precision of estimate by using bootstrap with equal weights
# USED

a. <- c(0,0,1,7)  # Previous year's score = 4
b. <- c(0,0,6,4)  # 3
c. <- c(0,.25,0,0) # 2
d. <- c(0,0,.25,0) # 1

n.a <- sum(a.)   # Total number getting 4 in prior year
n.b <- sum(b.)
n.c <- sum(c.)
n.d <- sum(d.)

p.a <- a./n.a     # Proportion getting 1, 2, 3,4 this year,
p.b <- b./n.b     # given prior year's score = a, b, c, d
p.c <- c/(n.c +.025)     # (i.e., a = 4 in prior year, etc.)
p.d <- d./(n.d+.025) # I inserted .025 here to eliminate 0/0;

prob <- cbind(p.a,p.b)  # Put probs into matrix
ave.sco <- prob %*% c(.25,.25)
#prob <- cbind(p.a,p.b,p.c,p.d)  # Put probs into matrix
# ave.sco <- prob %*% c(.25,.25,.25,.25)
t(ave.sco) %*% c(1,2,3,4)   # level of this year's score, [1]
(t(ave.sco) %*% c(1,2,3,4))/.50       # adjusted level of this year's score, [1]

# find bootstrap distribution
boot <- NULL
for(i in 1:1000)
{
aa.2 <- rmultinom(1,n.a,a.)  # generate data for a, b, etc.
bb.2 <- rmultinom(1,n.b,b.)
c.2 <- rmultinom(1,n.c+1,c.+ .025)
d.2 <- rmultinom(1, n.d+1, d. + .025)

prob.boot <- matrix(c(na.2/sum(na.2),nb.2/sum(nb.2),
                        nc.2/sum(nc.2),nd.2/sum(nd.2)),c(4,4),byrow=F)

# boot <- c(boot,t(ave.sco.boot) %*% c(1,2,3,4))
ave.sco.boot <- prob.boot %*% c(.25,.25,.25,.25)
boot <- c(boot,t(ave.sco.boot) %*% c(1,2,3,4))
}
hist(boot)
mean(boot)
quantile(boot,c(.025, .05, .10, .25, .50, .75, .90, .95, .975))
# Calculator for finding mean for 3 score categories with state weights. Version 1 (used)

a. <- c(0,0,8,1)  # Previous year's score = 4
b. <- c(0,1,8,1)  #   "   "   "  3
c. <- c(0,1,1,0)  #      2
d. <- c(0,.25,0,0)  #      1

n.a <- sum(a.)  # Total number getting 4 in prior year
n.b <- sum(b.)
n.c <- sum(c.)
#n.d <- sum(d.)

p.a <- a./n.a  # Proportion getting 1, 2, 3,4 this year,
p.b <- b./n.b  # given prior year's score = a, b, c, d
p.c <- c./n.c  # (i.e., a = 4 in prior year, etc.)
p.d <- d./(n.d+.025)  # I inserted .025 here to eliminate 0/0;

#prob <- cbind(p.a,p.b,p.c)  # Put probs into matrix
#ave.sco <- prob %*% c(.27,.53,.14,.06)
#(t(ave.sco) %*% c(1,2,3,4))/.94  # level of this year's score

prob <- cbind(p.a.p.b.p.c)  # Put probs into matrix
ave.sco <- prob %*% c(.27,.53,.14)
(t(ave.sco) %*% c(1,2,3,4))/.94  # level of this year's score adjusted [1]
# find bootstrap distribution using all 4 category scores
boot <- NULL
for(i in 1:1000)
{
  na.2 <- rmultinom(1,n.a,a.)  # generate data for a, b, etc.
  nb.2 <- rmultinom(1,n.b,b.)
  nc.2 <- rmultinom(1,n.c,c.)
  nd.2 <- rmultinom(1, n.d+1, d. +.025)

  prob.boot <- matrix(c(na.2/sum(na.2),nb.2/sum(nb.2),
                       nc.2/sum(nc.2),nd.2/sum(nd.2)),c(4,4),byrow=F)

  ave.sco.boot <- prob.boot %*% c(.27,.53,.14,.06)
  boot <- c(boot,t(ave.sco.boot) %*% c(1,2,3,4))
}
hist(boot)
mean(boot)
quantile(boot,c(.025, .05, .10, .25, .50, .75, .90, .95, .975))
# Calculator for finding mean when 3 categories with equal weights.

a. <- c(0,0,8,1)  # Previous year's score = 4
b. <- c(0,1,8,1)  # " " " 3
c. <- c(0,1,1,0)  #
#d. <- c(0,.25,0,0)  #

n.a <- sum(a.)  # Total number getting 4 in prior year
n.b <- sum(b.)
n.c <- sum(c.)
#n.d <- sum(d.)

p.a <- a./n.a  # Proportion getting 1, 2, 3,4 this year,
p.b <- b./n.b  # given prior year's score = a, b, c, d
p.c <- c./n.c  # (i.e., a = 4 in prior year, etc.)
# p.d <- d./(n.d+.025)  # I inserted .025 here to eliminate 0/0;

#prob <- cbind(p.a,p.b,p.c,p.d)  # Put probs into matrix
ave.sco <- prob %*% c(.25,.25,.25,.25)
prob <- cbind(p.a,p.b,p.c)  # Put probs into matrix
ave.sco <- prob %*% c(.25,.25,.25)

t(ave.sco) %*% c(1,2,3,4)  # level of this year's score, [1]

# find bootstrap distribution
boot <- NULL
for(i in 1:1000)
{
  na.2 <- rmultinom(1,n.a,a.)  # generate data for a, b, etc.
  nb.2 <- rmultinom(1,n.b,b.)
  nc.2 <- rmultinom(1,n.c,c.)
  nd.2 <- rmultinom(1,n.d+1, d. + .025)

  prob.boot <- matrix(c(na.2/sum(na.2),nb.2/sum(nb.2),
                       nc.2/sum(nc.2),nd.2/sum(nd.2)),c(4,4),byrow=F)

  ave.sco.boot <- prob.boot %*% c(.25,.25,.25,.25)
  boot <- c(boot,t(ave.sco.boot) %*% c(1,2,3,4))
}
hist(boot)
mean(boot)
quantile(boot,c(.025, .05, .10, .25, .50, .75, .90, .95, .975))

# Calculator for finding mean for 3 score categories with state weights for 84% and for 95% CI. Version 1 (used)

a. <- c(0,0,1,2)  # Previous year's score = 4
b. <- c(0,1,0,6)  # " " " 3
c. <- c(0,1,5,0)  #

d. <- c(0,.25,0,0)  #

n.a <- sum(a.)  # Total number getting 4 in prior year
n.b <- sum(b.)
n.c <- sum(c.)
n.d <- sum(d.)

p.a <- a./n.a  # Proportion getting 1, 2, 3, 4 this year,
p.b <- b./n.b  # given prior year's score = a, b, c, d
p.c <- c./n.c  # (i.e., a = 4 in prior year, etc.)
p.d <- d./(n.d+.025)  # I inserted .025 here to eliminate 0/0;

#prob <- cbind(p.a,p.b,p.c,p.d)  # Put probs into matrix
#ave.sco <- prob %*% c(.27,.53,.14,.06)
#t(ave.sco) %*% c(1,2,3,4)  # level of this year's score

prob <- cbind(p.a,p.b,p.c)  # Put probs into matrix
ave.sco <- prob %*% c(.27,.53,.14)

ave.sco.boot <- prob.boot %*% c(.27,.53,.14,.06)

boot <- c(boot,t(ave.sco.boot) %*% c(1,2,3,4))

hist(boot)
mean(boot)
quantile(boot,c(.025, .08, .10, .25, .50, .75, .90, .92, .975))

# Calculator for 2 score categories to find precision of estimate by using bootstrap with state weights USED
a. <- c(0,0,1,2)  # Previous year's score = 4
b. <- c(0,1,0,6)  # " " " 3
c. <- c(0,0,0.25,0.25)  #
d. <- c(0,0,.25,0.25)  #

n.a <- sum(a.)  # Total number getting 4 in prior year
n.b <- sum(b.)
n.c <- sum(c.)
n.d <- sum(d.)

p.a <- a./n.a  # Proportion getting 1, 2, 3,4 this year,
p.b <- b./n.b  # given prior year's score = a, b, c, d
p.c <- c./(n.c +.025)  # (i.e., a = 4 in prior year, etc.)
p.d <- d./(n.d+.025)  # I inserted .025 here to eliminate 0/0;

prob <- cbind(p.a,p.b)  # Put probs into matrix
ave.sco <- prob %*% c(.27,.53)

prob <- cbind(p.a,p.b,p.c,p.d)  # Put probs into matrix
ave.sco <- prob %*% c(.27,.53,.14,.06)
t(ave.sco) %*% c(1,2,3,4)  # level of this year's score, [1]

(t(ave.sco) %*% c(1,2,3,4))/.80  # adjusted level of this year's score, [1]

# find bootstrap distribution
boot <- NULL
for(i in 1:1000) {
  na.2 <- rmultinom(1,n.a,a.)  # generate data for a, b, etc.
  nb.2 <- rmultinom(1,n.b,b.)
  nc.2 <- rmultinom(1,n.c+1,c.+ .025)
  nd.2 <- rmultinom(1, n.d+1, d. + .025)

  prob.boot <- matrix(c(na.2/sum(na.2),nb.2/sum(nb.2),
                       nc.2/sum(nc.2),nd.2/sum(nd.2)),c(4,4),byrow=F)

  boot <- c(boot,t(ave.sco.boot) %*% c(1,2,3,4))
}

ave.sco.boot <- prob %*% c(.27,.53,.14,.06)

boot <- c(boot,t(ave.sco.boot) %*% c(1,2,3,4))
}
hist(boot)
mean(boot)
quantile(boot,c(.025,.08,.10,.25,.50,.75,.90,.92,.975))
References


25(3), 287-298. doi:10.3102/01623737025003287

doi:10.1016/S0272-7757(00)00039-X

century* (No. CSE Technical Report 549). Los Angeles: Center for the Study of
Evaluation (CSE), National Center for Research on Evaluation, Standards, and Student
Testing (CRESST).

6(1), 83-102.

systems* (No. CSE Technical Report 539). Los Angeles: Center for the Study of
Evaluation (CSE), National Center for Research on Evaluation, Standards, and Student
Testing (CRESST).

NC: SAS Institute Inc.

introduction for Wisconsin educators* (Wisconsin Department of Public Instruction).


Lord, F. M. (1980). Applications of Item Response Theory to Practical Testing Problems. LEA,
NJ.


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


Sanders, W. L., & Horn S. P., (1998). Research findings from the Tennessee value-added assessment system (TVAAS) database: Implications for educational evaluation and

http://beteronderwijsnederland.net/files/cumulative%20and%20residual%20effects%20of  
%20teachers.pdf


Wright, S. P., White, J. T., Sanders, W. L., & Rivers, J. C.


(Ed.), *Longitudinal and value added models of student performance*. Maple Grove, MN: JAM Press.