The Evaluation of Teacher and School Effectiveness Using Growth Models and Value Added Modeling:

Hope Versus Reality

Abstract

Robert W. Lissitz

University of Maryland

Presentation to Division H, AERA, 2012

Abstract

This paper begins with a short overview of the literature regarding Student Growth Modeling (SGM) and Value Added Modeling (VAM). Included are comments regarding the motivation and goals for this interest, from NCLB and RTTT to a more general interest from those concerned with measuring the effectiveness of teaching on student performance. It progresses to a review of some of the literature that addresses the assessment of teachers and schools using more formal statistical methods. The paper continues to address issues of reliability and validity as applied to evaluating teachers and evaluating schools. From there, it moves to a data-based comparison of SGM that permits aggregating across teachers or schools to provide teacher and school performance information (VAM). The focus of this analysis is on relatively simple models. Finally, the paper closes with some conjectures about the future and the likelihood of success with this endeavor. The overall tone of the paper is one of skepticism, pessimism and advising caution.

The Evaluation of Teacher and School Effectiveness Using Growth Models and Value Added Modeling:

Hope Versus Reality

Robert W. Lissitz

University of Maryland

Opening Remarks

First, I want to thank the creators of this symposium for the opportunity to address this audience on the subject of Student Growth Models (SGM) and Value Added Modeling (VAM), and particularly Burcu Kaniskan, with whom I communicated and who just received his Ph.D. a few months ago. I also want to thank a number of people who are associated with my center, particularly the State of Maryland that provides the funding for The Maryland Assessment Research Center for Education Success (MARCES). The members of the MARCES Center are Laura Reiner, Yuan Zhang, Xiaoshu Zhu and Dr. Bill Schafer. These folks helped in numerous ways with this presentation. Drs. Xiaodong Hou and Ying Li were instrumental in previous work related to this project. They received their Ph.Ds and moved on. Additional student input includes comments from Yong Luo, Matt Griffin, Tiago Calico, and Christy Lewis. Dr. Bill Schafer has offered excellent advice throughout this project.

In this paper, I would like to quickly summarize a little history, then talk about some of the literature that attempts to study reliability and validity. Then I want to examine the application of some very simple value added models (VAM). Finally, I want to briefly mention the direction in which I think this field is going, both from an applied viewpoint and from a psychometric viewpoint.

Introduction and A Little History

Before I start on a very brief history, I want to note that once again the federal government, this time through Race to the Top, and previously with No Child Left Behind or what should have been termed Race to the Middle, have come to psychometricians to help the educators in the belief that we will provide the basis for tough decisions about teachers. In my experience, most school personnel have a pretty good idea which teachers are good and which ones are really at the bottom. What is the problem that is being solved by evaluating teachers using very complex psychometric models? Mainly, I think, the government wants us to try to develop a system that will pressure the educational administration to do the right thing for students and combat the teacher's union that is perceived as preventing us from doing that. Unfortunately, I think we are asked to do something that we are not actually able to do yet, but that will not likely stop some from using value-added modeling in high-stakes decision making.

The federal government, being supremely well intentioned, imposed data requirements on the schools with the NCLB Act. Sadly, the school administrators were seen as needing an attorney general to order them to collect data about the performance of the students in their charge and the associated success of their schools. In many cases that perception

was based on some reality. Many school systems had been too slow to adopt formal approaches to evaluating the success of their enterprise. So NCLB required data collection that would measure a school's status (where students are when they finish the year, regardless of where they started). Now, we are asked to implement growth modeling within a computer-based assessment framework. No small task, although quite an interesting endeavor.

Since the federal government voted to repeal the law of individual differences and required everyone to be proficient by 2014, some seemed to rest confident that the teachers and school administrators would respond to the pressure brought to bear and do the right thing for American children. The assumption seemed to be that the only reason the students have not been more successful is that the teachers have not wanted them to be. Of course that is all nonsense, but that does not seem to have much impact on the federal government and their insatiable appetite for centralized control of the educational establishment.

About 10 years ago a number of schools were approved to try some very simple growth modeling. Their models are included on the web site: http://www.ed.gov/admins/lead/account/growthmodel/index.html and researchers have been examining what was proposed. The current effort is more ambitious. Value Added Modeling (VAM) means that a formal system has been adopted that we hope will permit the determination of the effectiveness of some mechanism or entity (usually teachers or schools or some poorly defined education program) to improve performance. The most popular models include the application of simple regression, recording the transition from one performance level to the next in adjacent grades, and mixed effects (or multi-level) regression models with the teachers or school as a level 2 effect. The results are usually aggregated so that summaries for every student associated with each teacher are provided. In this way, evaluators hope to be able to show whether students exposed to a specific teacher are performing above or below the statistical expected performance or the performance levels of students associated with other teachers (perhaps an artificial "average" teacher). Most VAM models are inherently normative in nature.

So let's remind ourselves of some of the history here in America, upon which we owe so much. Back in the early 1980s there were researchers such as Garrett Mandeville (1987) who asked important questions about school effectiveness and the reliability of related indicators. He did find that some schools were better than others, at least temporarily so, but the differences were not very consistent across years or even within schools across grades or subject matter. In other words, a school might have a very good 5th grade English class and a not so very good 3rd grade math class. When one hopes to label a school as effective or ineffective, it is implied that there is some consistency across grades and subject matter. This lack of consistency does not support, as an example, a belief that the principal is the key ingredient in a "good" school. It became obvious that, to the extent that student performance is influenced by school-based factors, the causal model capturing that process must become a great deal more complex. At the same time, some of us were left wondering whether the problem was the statistical model itself or the phenomena being measured. Maybe there is no model (no matter how complex) that is going to be able to capture teacher or school effectiveness.

In the mid-1990s we have two very important developments. At around the same time that the influential "TVAAS" VAM model was being implemented in Tennessee, a different system for the same purpose was set up in Dallas, Texas. The original version of the Dallas system went into effect in 1994 and examined school effects (Webster & Mendro, 1994); this was expanded to include teacher effects in the 1995-1996 school year (Webster, Mendro, Bembry, & Orsak, 1995). The model was composed of two stages. The first stage used multiple regression to control the effects of "fairness variables," which were defined as student differences in gender, ethnicity, English proficiency, socioeconomic status and any other variable that was considered to be an unfortunate influence beyond the school's or teacher's ability to control. A multiple regression was used to remove these student variables by creating residual values, linearly independent of them. The second stage of the analysis used a hierarchical linear model (HLM) to control the effects of prior achievement, attendance, and school-level variables and to measure the conditional growth in student performance.

The staff in Dallas was interested in doing something high stakes with their statistical results. Bonuses were provided based on these models, for example. One additional purpose of Dallas's VAM was to help determine how often to observe individual teachers. They adopted a policy that those estimated to be most effective were to be observed every two to three years, while those least effective received extensive observation and interventions during the year. Teachers in the middle category of effectiveness would be observed annually (Webster, Mendro, Bembry, & Orsak, 1995). Dallas' system also set the stage for combining VAM with more qualitative systems of evaluation,that seem to be the same direction that the federal government is encouraging us to pursue now.

Meantime, Sanders and his team were working hard to characterize the quality of teachers in Tennessee. The Sanders' approach, as it is sometimes called, was a great deal more statistically sophisticated. It involved a "layered" multiple regression model called TVAAS, that looked for the effects of teachers (and past teachers, hence the term layered) that deflected the student gains from their expected performance levels so they were either above or below predicted performance. Many models have been used, but the one embraced by Tennessee is a mixed effects model using longitudinal performance measures. Multiple prior years' performance scores on several subject matter exams were used as a way of covarying out the effect on student growth of student characteristics that were undesirable. It was assumed that by doing this, the variables that made teaching some students more challenging than others were eliminated. However complex interactions could not be statistically removed and may account for some of the disappointing efforts to validate this assumption. In other words, it is possible that effects that influence bright students may not affect less capable students and eliminating the effect with simple statistical adjustments is probably not possible. Also, perhaps, some teachers are simply better with some groups of students than others. In addition to these interactions, it is likely that there are latent classes, or relatively homogenous groupings (subsets) of teachers and students. This observation gives rise to the current interest in mixture modeling and I predict this approach will become much more popular in the near future.

The differences in the difficulty of teaching students who differ across classrooms or systems, might be called the differential challenge problem. It is a little like measuring height with rulers having different-sized units and then comparing the results. Unfortunately, the use of past student performance does not seem to completely control for such factors. Rothstein (2008a, 2008b), in fact, has shown that the effect of non-random assignment of students to teachers is not controlled by the use of the prior performance level. There is, on the other hand, some evidence that using multiple prior measures, as Sanders does, reduces the bias that results in estimating teacher effectiveness, but I think this is not a settled problem. It certainly appears that there is an association between teacher effectiveness and student characteristics, even when we use prior performance as a covariate. There is a "dynamic" (to use Rothstein's term) interaction between student and teacher and controlling for and even estimating such an effect is very challenging. He also shows that it is an advantage to have a class of students who performed more poorly than expected in the last year. They are more likely to make gains in your class this year. Regression effects often account for observations such as these.

Many factors confound the teacher effects that do exist and the dynamic, interactive nature of the classroom and school system further complicate the assessment problem. Using the prior test performance to serve as a control for all sorts of other effects is discussed in the Newton, et al (2010) paper. Some analyses of their data show a relation between effectiveness and percent minority, even after controlling for prior performance, for example. The problem, at least in part, is that such factors are not just main effects easily controlled by recording performance levels at the beginning of the year. They interact with the teacher's ability to be effective all year long and they interact with other student factors, as well.

Kupermintz (2003) is but one more of the voices that raised issues with the accuracy of this work and associated claims. Another very critical paper is by Amrein-Beardsley (2008) and is followed by a rejoinder by Sanders and Wright (2008). Rothstein (2008), as mentioned above, is another good paper to read that calls into question VAM models, in general. McCaffrey, et al (2003) provide a balanced discussion of their model as well as others. Other states have adopted variations, such as the model used in Pennsylvania called PVAAS.

As I stated, the VAM approach to assessing teacher and school effectiveness is not undisputed. (e.g., Braun, 2005; Glass, 2004; Kupermintz (2003); McCaffrey, Koretz, Lockwood, & Hamilton, 2004). This is particularly important for us to hear in the case where the work becomes high-stakes for teachers or schools involving bonuses, public humiliation, or even the threat or reality of being fired. Many teachers are working in areas that do not involve standardized testing and therefore it is hard to calculate VAM based effectiveness indicators for them. Florida (Prince, et. al. 2009, page 5), for example, has calculated that 69 % of its teachers are teaching non-tested subjects and grades. In Memphis, Tennessee the current testing program does not apply to about 70 % of the teachers, according to Mathematica (Lipscomb, et. al. 2010). This is a problem that is quite common today and it is hard to know what to do about it, although it is not the only methodological problem. For example, most teachers do not actually work alone

with students. They have other teachers, other support personnel including the librarian and counselors, plus parent volunteers, plus co-teachers making the inference or attribution of effectiveness to the teacher more confounded and doubtful.

I think it is safe to say that many feel that VAM is not really ready for high-stakes decision making, although perhaps ready to be partnered with additional data gathering efforts to contribute to a multiple-measures view of teacher effectiveness. Overall, VAM is a very challenging endeavor and probably well worth pursuing, but so challenging as to make high-stakes applications a very high risk. I would like to focus now on reliability, as an issue, and then talk briefly about validity issues, before presenting some of our own data analyses.

Reliability

If one thinks of the reliability of VAM as a generalizability problem, you can ask about the stability or consistency of teacher or school effectiveness outcomes as a function of some intervening condition, such as time, test, course, grade in school, etc. In other words, we can ask if the effectiveness estimates are sensitive or stable in the face of changes in when the test is given, which test is analyzed, what course the teacher is responsible for, and what grade are the students enrolled in, to name just a few relevant facets. If we want to characterize a teacher as effective and another as ineffective, we need to be concerned with whether such a characterization is justified as a main effect, or whether teachers are actually effective in some circumstances and ineffective in others. In other words, if interactions exist, the problem for the principal changes from "who is ineffective?" to "are there conditions in which this teacher can be effective?" The following comments are a very brief summary of some of the results in the literature, at least in my opinion.

Stability over a one-year period: One of the first sets of papers that explored the issue of reliability (stability) was that by Garrett K. Mandeville (e.g., 1988), as mentioned previously. He explored the estimation of effectiveness as a school residual from the expectation of a regression model. He looked at the stability of these estimates across a one-year time period and found that schools were stable in the 0.34 to 0.66 range of correlations, although large differences across grade level and subject matter were observed. .

McCaffrey, et al (2009) have looked closely at the stability of the estimates of teacher effects and found them to be modest. They report 0.2 to 0.3 as the correlation between estimates one year apart. That is troublesome. McCaffrey et al (2009) did make a very useful distinction between the reliability of teacher characterizations across a year in time and the reliability of the measures themselves. Others who have looked at this form of the reliability question include Newton, et al (2010) and Corcoran (2010). We certainly know that there are many sources of unreliability that can negatively impact the stability of characterizations of individual teachers. Test reliability is just one source. It is not clear that teaching can be considered a stable phenomenon. That is, there is some evidence that teacher effects are at least partly a function of an interaction with the nature of the students and changes in the teachers themselves. If the instability is due to

sampling error or some statistical issues, at least it might be reduced by increasing sample size and averaging. If the variability is due to actual performance changes from year to year, then the problem may be intractable (McCaffrey, et al 2009).

Stability over a short period of time: Sass (2008) and Newton, et al (2010) found that estimates of teacher effectiveness defined from what amounts to test-retest assessments over a very short time period was reasonably high. Correlations in the range of 0.6, for example, have been reported in the literature. This shows that teacher effectiveness may be somewhat consistent if we look the second time shortly after our first view of the teacher**.** We usually demand greater than.8 reliability for high stakes testing, so these results should cause us some alarm, but they do seem to indicate that something real is occurring.

Stability across grade and subject: Mandeville and Anderson (1987) and others (e.g. Rockoff, 2004; Newton, et al, 2010) found that stability fluctuated across grade and subject matter. Some limited stability was more often found with mathematics courses and less often with reading courses. Again, this is all quite troublesome if one is interested in doing something with the teacher effect estimates. If your success depends upon what class you are assigned and not just on your ability, that raises serious issues of fairness and comparability. It also challenges our effort at intervention.

Stability at the school level: The perception that entire schools are either good or bad is a very popular belief that I once explored informally in two sets of conversations. When I conducted the evaluation of the Court Ordered Desegregation Plan in Saint Louis back in the early 1990s, I challenged the 30 or so school-based members on my advisory committee to find a top school that remained at the top three years in a row. No system out of the half dozen or so that reported back had even one such school. Then I called the body that governs Blue Ribbon schools in DC and asked the same question regarding stability. Amazing candor was demonstrated when the person I talked to said they had noticed that the winning school in one year was typically not at the top a year or two later. The bottom line is that rankings or groupings of schools (e.g., quintiles) are not that stable (Sass, 2008).

Stability across test forms: Sass (2008) provides a short summary of some of the literature on stability by citing McCaffrey, et al (2008), Koedel and Betts (2007), and others, if the reader is interested in a nice review. Sass compares quintiles of performance and found that the top 20% and the bottom 20% seemed to be the most stable based on both a low-stakes and a high-stakes exam. The correlation of teacher effectiveness for these data was 0.48 across comparable examinations. Note that this correlation was based on two different, but somewhat related exams over a short time period and limited to classification of teachers into five quality categories (quintiles). When the time period was extended to a year's duration between tests, the correlation of teacher effectiveness dropped to 0.27. Papay (2011) also looked at the issue of stability across test forms and explored VAM estimates using three different tests. Rank order correlations of teacher effectiveness across time ranged from 0.15 to 0.58 across the different tests. Test timing and measurement error are credited with causing some of the relatively low level of stability of the teacher effect sizes. Note that which assessment device you use can make a difference to the estimation of teacher effectiveness.

Stability across statistical models: Tekwe, et al (2004) did a very interesting study in which they compared four regression models. The main take away point, I believe, was that unless the models involve different variables, the results tend to be quite similar. Three of the models gave very similar results and one model that involved variables not included in the other models (poverty and minority status) resulted in somewhat different estimates of effectiveness. Linear composites seem to be pretty much the same regardless of how one gets the weights (Dawes, 1979). Hill, et al (2011) discuss this as a convergent validity problem.

Stability across Classrooms: Newton, Darling-Hammond, Haertel, and Thomas (2010) have looked at factors that affect teacher effectiveness and found that stability of teacher ratings can vary as a function of classes taught. They also found that teaching students who are less advantaged, ESL, in a lower track, and/or low income students can have a negative impact on teacher effectiveness estimates. They also looked at the stability of teacher effectiveness ratings across statistical models and across courses taught. In many cases they even found inverse relationships among courses taught by the same teacher, although these results were generally not significant. Their study also tried to match VAM scores with extensive information about teaching ability. Multiple VAM models were used, and the success of matching teacher characteristics to VAM outcomes was judged to be modest. It is tempting to consider the VAM score as a criterion to be used to judge other variables, but their questionable validity makes that a doubtful approach.

The effort to develop a fair and equal system for scoring two teachers who come to the table with the same teaching skill, despite teaching two different groups of students (perhaps one with language challenged and learning disabled students and the other not) is certainly a worthy goal. Will we find stability, or fairness to be present in such a system? In my opinion, we probably will not. We simply do not have models that are so accurate that they can ignore or compensate for the context of the instruction. I also doubt, as I indicated above, that we should assume that effective teaching is a simple constant (i.e. a main effect with no interactions – either you are a good teacher or you are not) no matter the characteristics of the students you are teaching or the context of the classroom. Does anyone really believe that all students are equally difficult to teach?

Sources of unreliability: Persistent effects, non-persistent effects, and non-persistence due to sampling error is a distinction that is discussed by Kane and Staiger (2002), and McCaffrey, Sass, and Lockwood (2009) as they try to understand this lack of consistency. Thirty to 60% of variation is due to sampling error associated with teachers (i.e., in part due to small numbers of students as the basis of their estimates). Some time-varying factors seem to be present, but it is not completely clear what they are. Atteberry's dissertation (2011) covers elements of this very interesting topic. The effect of regression to the mean can also complicate the interpretation of effectiveness. The sample size plays a role in this phenomenon and that may differentially affect teacher's indices since class sizes may vary within a school or a district (Kupermintz, 2001 and 2003, and Amrein-Beardsley, 2008). Bayes estimates used in multi-level modeling also introduce bias that is a function of sample size.

So, in summary, we seem to know that effectiveness is not very highly correlated with itself over a one-year period, across different tests, across different subject matter or

across different grades, but we might ask if we should have been surprised. Glazerman, et al (2010) briefly summarize the stability indices for a few other occupations and it turns out that the lack of consistency for teachers is typical of complex professions. When compared to baseball players, stock investors, and several other complex professions, we find the same resulting lack of consistency. Their argument is that while teacher effectiveness does not seem to correlate from year to year particularly well, teachers are no less reliable than other professionals working in complex industries. Perhaps we are just setting different expectations for ball players than we do for teachers.

Validity

Reliability is the easy thing to study. Validity is a much more complex concept and it is not altogether clear how we should verify the validity of the work on teacher or school effectiveness.

The work funded by the Gates foundation (Measures of Effective Teaching, Kane, et. al, 2009) and involving a number of partners moves us a bit forward in the process of understanding VAM, but we have a long way to go. In that work they used the VAM outcome as a dependent variable looking for the sort of teacher information that is supplied in teacher evaluations by students to identify what is correlated with high VAM scores and what is not. Their emphasis is upon what teachers can do to be effective. There is, unfortunately, a big jump to go from finding a small correlation to defining a causal model. This work is also not above criticism, as evidenced by Rothstein's comments (January, 2011). Their work is correlational for now, although they hope to move to random assignment of students, according to Rothstein.

Job Applications as measures predicting effectiveness: As a related issue, it would be nice if there were valid associations between teacher effects and the typical information associated with a teacher's application for employment. Unfortunately, while some evidence for the utility of such factors exists, they are, at best, weak indicators. As Sass (2008) has said, such variables as years of experience and advanced degrees have low relationships, if any, to teacher effectiveness. Sanders, Ashton, and Wright (2005) did find a weak relationship between effectiveness and possession of an advanced degree. But, this result was described in the Sanders and Wright (2008) paper as little better than a coin flip between teachers with National Board for Professional Teaching Standards certification and those without. Hill, Kapitula and Umland (2011) did find that for teaching mathematics, knowledge of mathematics was positively correlated with effectiveness. That is certainly comforting. They found that VAM scores correlate with math knowledge and the characteristics of the students they are teaching. Hanushek (1986) notes that teacher education and teacher experience effects are usually not significant. Goldhaber and Hanson (2010) found with North Carolina data that VAM estimates seem to provide better measures of teacher impact on student test scores than do measures obtained at the time a teacher applies for employment. They include such measures as degree, experience, possessing a master's degree, college selectivity, and licensure in addition to VAM estimated teacher effect. Frankly, none of these variables

have a particularly strong relationship to student achievement. Perhaps assessment centers or other such selection procedures might hold more promise.

Triangulation of multiple indicators: Goe et al (2008) provides another excellent survey and summary of the literature in the VAM area. They talk about other ways of evaluating teachers, specifically using some form of observation and identifying the factors that lead to effectiveness. They reference Danielson's (1996) Framework for Teaching as a common source for collecting relevant information about teachers. One implication, as Goe, Bell and Little (2008) say, is that teachers should be compared to other teachers who teach similar courses in the same grade in a similar context and assessed by the same or similar examination. That is certainly consistent with the literature on VAM stability, referenced above, and what is probably necessary to eventually establish validity. It also acknowledges the complex interactions that seem to exist.

Comparability: In the literature, it is often assumed that status is actually independent or at least uncorrelated with growth. As Kupermintz (2003) suggests, it seems obvious that ability is very likely to be correlated with growth and status. Does anyone really believe that dull students will learn at the same rate as gifted students? Smart students have a tremendous advantage over dull students, as do the teachers who have them in their class. It is likely, as Kupermintz indicates, that there is also an interaction between student ability and the ability of teachers to be effective. We do not like thinking about such realities and, as I said before, the federal government, through NCLB, repealed the law of individual differences in their requirement that every student will become proficient in just a few more years. Obvioiulsy, they are finding that changing human nature is more difficult than anticipated. Yumoto (2011) has just completed a dissertation where he examines modeling a class in which some students are affected by the teacher and some are completely unaffected. We are only beginning to develop such mixture models and they can only help us in our understanding of teacher effectiveness. The estimation of teacher effects seems to present us with a very complex interaction involving mixtures of students and teachers.

Causality, research design and theory: We are trying to infer causality and it is unlikely that we will succeed in doing that, using statistical adjustments. Whether we do or not can be considered a question of the validity of our effort. Rubin (2004) has approached the problem of research design in a very different way from Campbell and Stanley (1963), but regardless of how one looks at the problem of validity and the desire to infer causality, it seems very problematic. There are numerous reasons for that being the case. Rubin focuses on a few reasons in his paper in the special issue of the Journal of Educational and Behavioral Statistics regarding VAM. One reason is that missing data from our schools are not missing at random. In fact, as Rubin shows in his analysis of several VAM papers in that special issue, the evidence is overwhelming that missing data are missing in a way that confounds the results and complicates inferences. Even worse for the purposes of teasing out individual teacher effects, there is communication across classrooms and therefore the effects of one teacher are confounded with the effects of other teachers, and in some cases, communication even occurs across schools.

Rubin also suggests that we should have a clear idea of what our hypothesis is. We do not. In other words, we need to know exactly what parameters we are trying to estimate, as distinct from how we are trying to obtain estimation? To put it another way, we have no theory. We have multiple operational definitions of growth, but seem to have no developmental science for the phenomenon at which we are looking. Without a theory it is hard to determine just how we would validate teacher or school effectiveness and their associated causality, if in fact there is any.

In the absence of carefully controlled experiments, the conditions in a school are very complex and it is probably impossible to tease out the effects for which we are looking. The fact that most students, even in elementary school, have multiple teachers is just one example of this complexity. Others include the fact that where a student is today in school and what influences him or her is in part a function of where he or she was before, as Sanders has indicated. There are some papers by McCaffrey and others that attempt to assess the effects of the past upon the current learning environment but again that poses a high level of complexity and our models and general understanding of the process are not really adequate to the challenge yet, in my opinion.

We seem to have no concrete idea what we even mean by the notion that teachers or schools have a causal effect upon the students. How do they impart this affect and how is this effect internalized or received by the student, for example? I do not think we know the answers. Even if we eventually find some behaviors that seem to be correlated with student effects, understanding how the effect actually happens seems to be beyond our ability to study at this time, unless we started to do actual experimental studies and we seem to be very reluctant to do so.

Lord's paradox provides an example of the problem of making inferences from simple statistical models. Rubin also addresses this issue. Lord's paradox involved the use of ANCOVA, of which Sander's models, and others as well, are a special case. The general conclusion from Lord's paradox is that without a clear sense of the theory we are trying to validate, ANCOVA does not lead to unambiguous interpretations. Even propensity models are probably not adequate to this task. My guess is that only experimental efforts will satisfy our desires and provide adequate results.

One time I asked an eminent faculty member about her interest in teacher decision-making. She talked about observation data that indicated teachers made a variety of decisions regarding student instruction. I asked her what she knew about optimizing teacher's decisions to help a student learn as much as he or she can possibly learn? In other words, knowing that teachers are not very systematic is useful information, but being able to tell teachers how to optimize a student's performance is a much more important goal. We are perhaps on the verge of some sort of paradigm shift in education. I certainly hope so. The work on cognitive models and neuroscience may help us accomplish this shift. Understanding what we mean by saying someone learned something will be an important step forward for optimizing teaching. It is time to develop a learning science.

Why should we care: It is not unusual to find a statement that indicates that teachers are the most important factor to determining student achievement. That is

obviously wrong, though. What difference does it make if a teacher is effective? Nye, Konstantopoulos and Hedges (2004) found that only 11% of variation in student test score gains was explained by teacher effects. As Jonah Rockoff (2004) notes in his study "Across subject areas, the upper bound estimates range from 5.0-6.4% for teacher effects, 2.7-6.1% for school-year effects, and 59-68% for student fixed effects. (Page 20)." Examining teacher and school effects is trying to study something that is at best subtle and just one element of many that result in a student learning something. It is not even clear that it is the most important element, although that can be debated.

There is a body of research that suggests that the context of the classroom beyond the characteristics, demographics and homogeneity of the students in the classroom, may be more central to learning than currently thought and perhaps even rivals the effect of the teacher. For example, the recent paper by Kennedy (2010) suggests that situational factors may be parallel in importance to the teacher factors that influence learning. She mentions time, materials, and work assignments as three factors that influence a teacher's potential success with his or her students. More important, in my opinion, are factors such as controlling or eliminating behavioral issues in the classroom and mainstreaming only students who are capable or willing to enter into a contract for learning in a non-disruptive manner. The goal should be to maximize student learning, but I do not think our profession would even agree to that statement. Maximizing the context for learning may be the best way to achieve that. Of course, teachers are a part of the context, but so are a lot of other factors. Again, I am asking the question Rubin asked: What is the theory? What is the conceptual basis for believing in the existence of "teacher" effects? What exactly are the effects we think teachers are responsible for?

The greater importance of student ability compared to teacher effectiveness and the importance of context and the general education environment suggests a very different orientation toward the teaching and learning process. In this paradigm the teacher would be there to optimize the context of the classroom, including adding to motivation, preventing disruption and generally providing an opportunity for enhanced learning engagement to take place. That is quite different from thinking the teacher is there to teach. In other words, I am not sure that we are not conceptualizing the whole problem incorrectly. The development of a learning science may help us a great deal in conceptualizing what is really going on when someone seems to have learned something. The introduction of assistive teaching devices is also likely to play a very strong role in the future of education and the result will be a change in the teacher's role.

In other words, we should be creating a laboratory for education sciences to examine the effects of teachers in an environment in which random assignment and intervention with serious control is present. And, I do not mean creating a lab school. I mean actually doing highly controlled experiments. Such an approach does not seem to be a natural one for many in education. The emphasis on external validity and immediate generality (gratification) to the field seems to me to be the dominant paradigm, or at least its primary motivation. Unfortunately, I do not see another way to begin to develop a deep understanding of what influences (causes) a student to learn, unless we begin to conduct experiments or at least create highly controlled environments.

Finally, we need to be concerned, as Kupermintz (2003) indicates, with the issue of being fair if we want to use value added models in a high stakes testing context. There is little evidence that VAM is yet worthy of such a task. On the other hand, is using VAM less fair than the use of traditional personnel selection that focuses on getting an advanced degree or more credit hours, and working more years in the system, getting an advanced certificate, and having the principal come by and observe your classroom periodically? Perhaps the effort to create a systematic evaluation system will prove to be an improvement, but I would say that we are certainly not at a comfort level yet with our models.

Our Study Comparing Models Using Real Data

Now I would like to abruptly shift to some data analyses we have just completed at MARCES. The full report is available at: marces.org/completed.htm The MARCES Center embarked on a study of some of the simplest models that might be contenders for application in the field. This study involves four cohorts from one large and diverse county, in which each cohort has three years of data obtained from each student from 3rd to 8th grade. One cohort ends in 5th grade; a second ends in 6th grade, the third ends in 7th grade, and the fourth cohort ends in 8th grade. These math and reading data are from a statewide assessment system administered at the end of each spring from 3rd to 8th grade. The assessments do not have a vertical scale, although they are horizontally equated from year to year. The lack of such a scale prompts us to examine 11 simple models that do not require vertical linking from year to year. Nine models are used to characterize growth from the first year to the second and two additional models are used to model growth from the first and second year to the third. The study design is summarized in Figure 1.

------ Insert Table 1 about here

The following table summarizes our data and the models that were used.

------ Insert Table 2 about here

The following are brief descriptions of the models that we selected for our study.

Betebenner's model (QRG1, QRG2 and ConD) is quite popular, being used in Colorado, for example. It is a simple model that looks at the conditional percentile of each student's performance in the second year compared to other students who started at the same percentile in the initial year. So we look to see whether a teacher moved a student to a relatively high or low performance level compared to other students who started at the same initial level. Aggregating the conditional percentiles of students exposed to each teacher gives a value added measure to that teacher. The hope is that conditioning the percentiles will yield a fair comparison across teachers. We looked at three models. One used the prior year (QRG1) and the second used two prior years (QRG2) to condition the percentile in the third year and the third model is a simplification created by aggregating students into deciles (ConD) on year one and then

looking at deciles in the second year. These models are very easy to explain and easy to compute.

Thum's model (ConZ) involves a similar approach, but looking at an effect size rather than a percentile. It amounts to a z score that identifies a student's performance level compared to the average student in the first year. In the second year, we compare the z score of students in their relative position compared to students who started at the same z position in the prior year. So again, the effectiveness is measured as being able to move students relative to students who started out at conditionally the same position. We simplified Thum's model in several ways, and one simplification was introduced by categorizing student performance estimated by the z score conditional on their prior decile. The conditional z scores were aggregated for each teacher to provide a measure of the effectiveness for that teacher. The simplification proceeds as follows:

1. Rank order all students' year-one scale scores and divide them into 10 deciles (i.e., with 10% of all the cases in each decile). This step is to divide students into different groups with the assumption that students in the same decile have similar prior achievements.
2. Compute the mean of year-two scale scores for students within each decile.
3. Compute the deviation scores of year-two scale scores from the decile mean for students within each decile.
4. Compute the pooled within-decile standard deviation of year-two scale scores.
5. Compute the growth z score for each student by dividing the deviation scores obtained from step 3 by the pooled within-decile standard deviation obtained from step 4.

Ordinary least squares regression deviation models with one predictor and with two predictors (OLS1 and OLS2) were also used. We looked at the errors of prediction and aggregated these errors across teachers to see which teacher's students tended to perform above prediction which below prediction. In the first case (OLS1), the independent variable was the prior year scale score performance and the deviation from expected scale score performance on the next year of testing was the effectiveness measure. In the second case (OLS2), we predicted the third year's scale score performance from the first two years' scale score performance and looked at the deviations.

Regression using spline scores (OLSS and DIFS) (Schafer, et. al, 2009) was calculated with scores that had been transformed by a spline function that was created to give "moderated" or relational meaning to various points along the performance continuum across grades, as though the data were at least somewhat vertically scaled. The transformation was matched to the cut scores for the three proficiency levels (basic, proficient, and advanced). The spline function is essentially a piecewise curve fitting model that allows us to rescale the data. We are building a quasi-vertical scale, without using common items. The first model (OLSS) applied ordinary least squares to the spline scale scores and looked at deviations from predicted, and the second model (DIFS) used the spline function transformed scale scores and simply subtracted the transformed score at the first grade from the transformed score at the second grade, as though they were from a true vertical scale.

Transition models (TRSG, TRUG, and TRUD) were applied with one approximately borrowed from Delaware (TRUD) and a second that approximates one from Arkansas (TRUG) and the TRSG model that Bill Schafer designed. These models start with the classification of students into categories, based on basic, proficient and advanced in one grade. In this case we divided each of these three categories of performance into three subcategories. The students are followed into the next grade and we observe which category of performance they fall into on the next annual test, conditional on where they were in the prior year. There is also a matrix of values associated with each transition from the level of the initial grade (the rows) to the designated level of the column of the next grade (the columns). In practice, these values are the result of a complex judgment task involving educators making decisions about the relative importance of each transition. We experimented with three such models and aggregated the values for each teacher to see which teachers were most successful in transitioning their students to the highest category of performance given where they began.

------ Insert Tables 3, 4, and 5 about here

To focus on TRSG, note that it rewards students for making progress both within and across performance levels (i.e., basic, proficient, and advanced) and it also compensates them for their effort to maintain their previous performance status. Moreover, the reward increases with the status of one's performance levels. Successive cells along the main diagonal are awarded one more value point. For example, remaining in category Basic 1 earns a growth score of 9 whereas the value increases to 10 if one remains in category Basic 2 for two consecutive years. As can be seen from Table 3, improving from level Basic 3 to Proficient 1 gains a growth score of 13 whereas growing from Proficient 3 to Advanced 1 (also a jump of one performance level) wins a growth score of 16. A value of 12 was selected to represent achieving and maintaining the "proficiency" level (i.e., remaining in Level 4 for two consecutive years). Thus those schools (or other levels of organization) that have an average growth value of 12 may be considered as having reached the target growth level. The reader should notice that TRSG rewards both growth and status.

In this presentation, I want to take a closer look at the transition models and demonstrate that it does make a difference how you model the growth and then the value added. As was mentioned above (see e.g. Tekwe, et al, 2004), if the variables are not particularly different and you are using a model that is in effect a linear composite, such as a regression model, you would expect that the results will be similar. The selection of the three transition matrices was executed to illustrate that different results can be obtained if you create truly different models for growth.

To determine how different all these models are we performed a number of calculations, summarized as follows.

1. Inter-correlation of student growth scores and their Dimensionality

Each student had a growth calculation for year 1 to year 2 and from year 2 to year 3. Using a simple factor analysis of the student growth indices, using each growth model, we found that one dimension accounts for the largest percentage of variance, although there is clearly a great deal of noise in these results. The growth from year 1 to year 2 was intercorrelated and the same was done for year 2 to year 3. The resulting intercorrelation matrix was factored and we found that the approaches display marked unidimensionality for each pair of years for each content (reading and mathematics), although the variance accounted for by even the first few factors is not at all high. Over 80% of the variance remains undefined by the first dimension. Here is one example. The rest are equivalent.

------ Insert Figure 1 about here

The scree plot makes you think the models are essentially the same since they are unidimensional. You would be wrong, as we show below.

2. Relation to demographic variables and pre and posttest scores

We looked at many additional relationships that might represent something interesting in these data. We looked at the correlation between the growth measures and students' background variables such as gender, race, SES, accommodation, and English proficiency and found them to be low. However, students' growth in reading tends to be slightly more correlated with SES and race than those in math.

Overall, the correlations between TRSG and pre- and post-tests are strongest among all the models, with the correlation between TRSG and pretest oscillating around 0.5 and that between TRSG and posttest varying around 0.8. If we exclude TRSG, the correlations between pretest and growth measures are low for regression-based models and medium for transition models whereas the relationship actually reverses in terms of the correlations between posttests and regression- or transition-based growth measures.

3. The correlation between growth in math and growth in reading.

------ Insert Table 6 about here

Note that the correlation between the student's scale scores for reading and math is reasonably high, as expected (0.64 to 0.74). But the correlations between the growth scale scores for math and reading is much more modest, ranging around 0.2, except for one of the transition models, which is close to 0.4. The exception is the TRSG model that values a combination of growth and status.

4. The correlation between the two growth periods (year 1 to 2 and year 2 to 3).

Notice that growth from the first to the second year is negatively correlated with growth from the second to the third year, with one exception. This makes sense. Showing a lot of growth in the first to the second period means that your post-test scale score tends to

------ Insert Table 7 about here

be higher and therefore when that score becomes the pre-test scale score for the next growth period it is more likely to be followed by a lower post-test scale score. That relationship is not true for the TRSG model that shows a positive correlation for the two growth measures on mathematics and essentially a zero correlation for reading. Remember that TRSG values both growth and status.

5. Teacher effectiveness and teacher reliability

Each model provided a student growth index and that index can be aggregated for teachers to give a teacher effectiveness index. We can see from Table 8 that the Intra-Class Correlation (ICC, we use the square root so that it will be on the same scale as the correlations) indicates that teachers seem to have a moderate impact upon the growth of students. The effect is stronger in mathematics than reading and TRSG stands out as emphasizing that effect. This is consistent with the literature.

------ Insert Table 8 about here

Because we have three years of data we can look at growth for different students twice for the same teachers. In Table 9, we looked at the reliability of these teacher effectiveness indices and we found a moderate correlation between the two aggregated measures for the same teacher. Teachers who appear to be reasonably effective at producing growth for their students the first time it is measured, tend to be able to produce growth the second time. Notice that math provides higher reliability (more stable) effectiveness growth measures for teachers in 6^{th} grade. For reading the highest reliability is obtained for teachers of 7^{th} grade. Also, notice that TRSG provides a much stronger measure of stability than the other measures, with the exception of ConZ measure applied to 7^{th} grade reading.

------ Insert Table 9 about here

6. **School effectiveness and school reliability**

------- Insert Table 10 about here

The student growth can also be aggregated across schools and then analyzed to see how effective schools are. In this case, we again calculate an Intra-class correlation and take the square root of it to put it on the same scale as we did with our correlations. In Table 10 we see that schools seem to have a modest effect, particularly with respect to mathematics. Again, TRSG seems to stand out by displaying a higher level of effect, when organized by school.

Because we have three years of data, again we can look at the reliability (stability) of the impact, this time of schools. Notice (Table 11) that the stability for schools in 6^{th} grade is higher than found for 5^{th} or 7^{th} grades, with the exception of TRSG. For schools, math reflects greater reliability than reading in grades 5 and 7, again with the exception of

TRSG in math. Also, notice that reading has some negative correlations of the growth reliability indices for schools in 7th grade indicating a reversal of the level of effectiveness of schools if three particular indices are selected for examination of growth. Again, TRSG stand out as displaying a larger effect and a somewhat different pattern.

------- Insert Table 11 about here

7. Comparison between School and Teacher Effect

-------- Insert Figure 2 about here

As can be seen from the two diagrams of Figure 2, the conclusions about the school effectiveness and teacher effectiveness obtained from various models are very similar and the school effect is almost always smaller than the teacher effect. Besides that, it can also be noticed that the growth effectiveness measures obtained from TRSG are the largest among all the models, presumably due to the fact that this model rewards both students' growth and their status.

8. Methodological Issues

------ Insert Figure 3 about here

Many methodological issues abound in the work on VAM. One of them has to do with the variance of the incoming data. For example, correlations and ICCs characterizing effectiveness depend heavily on the amount of variance that resides among the teachers and in the schools. An example of this is contained in Figure 3, which systematically eliminated the teachers who were at the low end of the effectiveness scale on mathematics. As this figure shows, the index of effectiveness of teachers of mathematics decreases greatly as the variance of the teachers is reduced from where it is in our original data and where it can be if the lowest teachers were actually eliminated. The irony is that the effectiveness of teachers will become statistically quite low if the less successful teachers were eliminated. This is similar to the observation that if all schools receive roughly equivalent funding levels, then funding will no longer be an important source of variance in their success. Variance accounted for measures have the property of being largely a function of the diversity of the scores.. In other words, if we eliminate all the poor teachers, we will find that teachers account for very little variance in student growth or their status for that matter.

General Conclusions

* The model you use can make a difference. You can find things, not find things, find large things or small, find inverse relationships or direct relationships depending upon the model you use. Deciding how to balance status against growth is just one of the key decisions that drives the models we decide to adopt. There has been no standardization for the modeling of VAM. Knowing that we are having a hard time is, of course, always better than not knowing. The contradictions within these quantitative approaches are an

issue, although at least it is possible to know there are issues. The traditional qualitative approaches used by principals, are not likely to be an improvement on VAM. Using any of these approaches for high stakes testing and decision making seems premature at this time, though. Combining two procedures that are not likely to be valid by themselves is not necessarily going to result in a valid system when used jointly.

* More sophisticated growth models would be nice to be able to explore. For example, I wish we could apply a 4 level model with many vertically scaled time points from many subject matter assessments, nested within students nested within all their teachers as level 3, and nested within their school context as level 4. Of course, I would like to have all the relevant student, teacher and school characteristics to bring into the model, as well. I would also like to be able to examine mixture models that treat students and teachers as members of a few discrete groups that interact. Perhaps some day we will have such data to explore.

* Interactions should be modeled. Why should we insist on modeling teacher effects as though all students reacted the same way or even that all teachers are the same from day to day or over a year's time, independent of the school and the nature of students? I believe there will be more interest in such models in the future.

* An increase in the exploration of school context effects and classroom context effects should be on our agenda, as well. I believe this has great implications for how we eventually create a learning science. It is not at all clear that there is actually a significant phenomenon here to be studied if we narrowly focus on teacher effects. Our results are quite modest, but I must admit they are more impressive than I thought they would be. There does seem to be an effect worth studying, but right now, I do not think we can be confident that we know what that effect looks like. That will come from developing theory driven research.

* The change in instruction that involves a great deal of supportive technology is probably closer than we think. That transition is going to enable a more scientific approach to our profession. When I was young, I loved the story of John Henry and his contest with the steel-driving machine. Another one I really liked was Mike Mulligan and his steam shovel. The cognitive, computer, econometrician, engineering professional is moving into the study of education and our field will change accordingly. Can technology-based instruction be far behind? I think not and I think the nature of teachers and the nature of instructional decision-making is in for radical changes for the better in the not too distant future. Perhaps one of us will write a book called "The intuitive teacher versus the robot instructor."

* Right now, I do not encourage anyone to use VAM in a high stakes endeavor. If you have to because of federal or state policy, then my recommendation is to use a two-step process to initially use statistical models to identify low-performing teachers and then verify these results with additional data. Triangulating (or what might be called looking for reliability or even redundancy of information) is usually a good idea. Remember that it makes a difference what VAM model we implement. Different teachers may be identified and their effectiveness may be estimated at different levels. Of course, we can

use more than one model at a time. Also, we can and should choose our model based on policy decisions that capture the goals and intent of a school system.

* Beginning to relate VAM to what teachers are actually doing is an important direction in which this research should proceed. Creating causal models and exploring them with experiments is also a worthwhile activity, I believe.

* Anyone interested in implementing a VAM approach might want to read Finlay and Managi (2008) addressing some of the practical political issues related to using VAM in the schools. Dealing with unions, the federal government, and special education advocates all combine to make effective teaching (and its measurement) a great challenge. Humility is certainly warranted, and in this area, seems particularly well justified

References

Amrein-Beardsley, A. (2008). Methodological concerns about the Education Value-Added Assessment System. *Educational Researcher, 37*(2), 65-75.

Atteberry, A. (2011). *Understanding the Instability in Teachers' Value-Added Measures Over Time.* Preliminary draft. Stanford University School of Education. NOT for CIRCULATION.

Baker, E. L., Barton, P. E., Darling-Hammond, L., Haertel, E., Ladd, H. F., Linn, R. L., Shepard, L. A. (2010). *Problems with the use of student test scores to evaluate teachers.* Economic Policy Institute, Washington, DC.

Bill & Melinda Gates Foundation (2011). *Learning About Teaching: Initial Findings from the Measures of Effective Teaching Project.*

Braun, H. I. (2005). Value-added modeling: What does due diligence require? In R. W. Lissitz (Ed.), *Value added models in education: Theory and applications* (pp. 19–39). Maple Grove: JAM Press.

Briggs, D., & Domingue, B. (2011). *Due diligence and the evaluation of teachers: A review of the value-added analysis underlying the effectiveness rankings of Los Angeles Unified School District teachers by the Los Angeles Times.* National Education Policy Center, Boulder, CO.

Buddin, R. (2010). How effective are Los Angeles elementary teachers and schools? http://www.latimes.com/media/acrobat/2010-08/55538493.pdf

Buddin, R., McCaffrey, D. F., Kirby, S. N., & Xia, N. (2007). *Merit Pay for Florida Teachers: Design and Implementation Issues.* Prepared for the Florida Education Association. RAND Corporation.

Calvert County Public Schools (2009). Observation/Evaluation Procedures. http://www.calvertnet.k12.md.us/departments/hr/observations/documents/OE.pdf

Campbell, D. T., and Stanley, J.C. (1963) "Experimental and Quasi-Experimental Designs for Research on Teaching." In N. L. Gage (ed.), Handbook of Research on Teaching. Chicago: Rand McNally

Cavanagh, S., & Zehr, M. (2010). Future of D.C. Reforms Is Uncertain Following Rhee's Plan to Resign. *Education Week, 30*(8), 6.

Corcoran, S. P. (2010). Can teachers be evaluated by their students' test scores? Should they be? The use of value-added measures of teacher effectiveness in policy and practice. Annenberg Institute for School Reform, Providence, RI.

Danielson, C. (2010). Evaluations that help teachers learn. *Educational Leadership, 68*(4), 35-39.

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

De Vise, D. (2010) Montgomery County to weigh student performance as a third of teachers' reviews. *Washington Post.* http://www.washingtonpost.com/wp-dyn/content/article/2010/04/20/AR2010042005120.html

District of Columbia Public Schools (2011a). *The District of Columbia Public Schools guide to value-added*. Washington, D.C.

District of Columbia Public Schools (2011b). *IMPACT: The District of Columbia Public Schools Effectiveness Assessment System for School-Based Personnel 2010-2011.* Group 1: General education teachers with individual value-added student achievement data. Washington, D.C.

Dawes, R. (1979) The robust beauty of improper linear models in decision making. American Psychologist, 34, 7, 571-582

Finlay, C. & Manavi, S. (2008) Problems in Implementing Value-Added Models: An Analysis of Current K-12 Value-Added Systems in the U.S. U. of California, LA, School of Public Affairs, Dept. of Public Policy

Goe, L., Bell, C., & Little, O. (2008) Approaches to Evaluating Teacher Effectiveness: A Research Synthesis. Washington, DC, National Comprehensive Center for Teacher Quality.

Glass, G. V. (2004). *Teacher evaluation* (Policy Brief). Tempe: Arizona State University, Education Policy Studies Laboratory. Retrieved May 9, 2008, from http://epsl.asu.edu/epru/documents/EPSL-0401-112-EPRU.doc

Glazerman, S, Loeb, S, Goldhaber, D., Staiger, D., Raudenbush, S., & Whitehurst, G. (2010) Evaluating Teachers: The important role of value-added. Brown Center on Education Policy at Brookings November 17, 2010.

Goldhaber, D., & Hansen, M. (2010). *Assessing the Potential of Using Value-Added Estimates of Teacher Performance for Making Tenure Decisions.* Working paper 31. Washington, D.C.: CALDER.

Hanushek, Eric A., "The Economics of Schooling: Production and Efficiency in Public Schools," Journal of Economic Literature, September, 1986.

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

Hinchey, P. H. (2010). *Getting teacher assessment right: What policymakers can learn from research.* National Education Policy Center, Boulder, CO.

Jaffe, I. (2010). *'L.A. Times' Database Angers Teachers, Union.* http://www.npr.org/templates/story/story.php?storyId=129537137

Kane, T. J. (2009) Learning about teaching: Initial findings from the measures of effective teaching project. A project funded by the Bill and Melinda Gates Foundation.

Kennedy, M. M. (2010). Attribution error and the quest for teacher quality. *Educational Researcher, 39*(8), 591-598.

Koedel, Cory and Julian R. Betts. 2007. "Re- Examining the Role of Teacher Quality in the Educational Production Function." Working Paper #2007-03. Nashville, TN: National Center on Performance Initiatives.

Kupermintz, H. (2003). Teacher effects and teacher effectiveness: A validity investigation of the Tennessee Value Added Assessment System. *Educational Evaluation and Policy Analysis, 25*(3), 287–298.

Lipscomb, S, Teh, B., Gill, B., Chiang, H., & Owens, A. (2010) Teacher and principal value-added research findings and implementation practices. Mathematica Policy Research, Inc.

Lissitz, B., & Doran, H. (2009). *Modeling growth for accountability and program evaluation: An introduction for Wisconsin educators* (White paper). Wisconsin Department of Public Instruction.

Los Angeles Times (2010). *Los Angeles Teacher Ratings.* http://projects.latimes.com/value-added/

Mandeville, G. K. (1988). School effectiveness indices revisited: Cross-year stability. *Journal of Educational Measurement, 25*(4), 349-356.

Mandeville, G. K., & Anderson, L. W. (1987). The stability of school effectiveness indices across grade levels and subject areas. *Journal of Educational Measurement, 24*(3), 203-216.

Maryland State Department of Education (2009). *Maryland's longitudinal data system: Continuing the vision.*

McCaffrey, D. F., Lockwood, J. R., Koretz, D. M., & Hamilton, L. S. (2003). *Evaluating Value-Added Models for Teacher Accountability.* Santa Monica, CA: RAND Corporation.

McCaffrey, D. F., Koretz, D., Lockwood, J. R., & Hamilton, L. S. (2004). *The promise and peril of using value-added modeling to measure teacher effectiveness* (Research Brief No. RB-9050-EDU). Santa Monica, CA: RAND Corporation. Retrieved May 9, 2008, from http://www.rand.org/pubs/research_briefs/2005/RAND_RB9050.pdf

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

National Board Resource Center at Stanford University. (2010). A Quality Teacher in Every Classroom: Creating a Teacher Evaluation System that Works for California. Perspectives from Accomplished California Teachers. National Board Resource Center at Stanford University.

National Council on Teacher Quality (2009). *State Teacher Policy Yearbook.* National Summary.

National Council on Teacher Quality (2010). *State Teacher Evaluation Policies.* www.stateinnovation.org/Events/Event-Listing/PDAM-09/.../Jacobs.aspx

Newton, X. A., Darling-Hammond, L., Haertel, E., & Thomas, E. (2010). Value-added modeling of teacher effectiveness: An exploration of stability across models and contexts. *Education Policy Analysis Archives, 18*(23), 1-22.

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

Papay, J.P. (2011) Different tests, different answers: The stability of teacher value added estimates across outcome measures. *American Educational Research journal*, 48,163-193.

Prince, C. D. Schuermann, P. J., Guthrie, J. W., Witham, P. J., Milanowski, A. T., & Thorn, C. A. (2009) The other 69 percent: Fairly rewarding the performance of teachers of non-tested subjects and grades. Center for Educator Compensation Reform http://www.cecr.ed.gov/guides/other69Percent.pdf (page 5).

Ravitch, D. (2010). The problems with value-added assessment. *Education Week: Bridging Differences.* http://blogs.edweek.org/edweek/Bridging-Differences/2010/10/dear_deborah_you_asked_what.html

Rockoff, J E. (2003) The impact of individual teachers on student achievement: Evidence from Panel Data. A project from a Harvard Seminar.

Rothstein, J. (2008a) Student sorting and bias in value added estimation: Selection on observables and unobservables.

Rothstein, J. (2008b) Teacher Quality in Educational Production: Tracking, Decay, and Stu- dent Achievement. Unpublished manuscript, Princeton University.

Rothstein, J. (January 2011) Review of Learning about Teaching. National Education Policy Center, U. of Colardo.

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

Russel, T., Putka, D., & Waters, S. (2007) Review of the Psychometric Quality of the National Board Assessments. http://www7.nationalacademies.org/bota/NBPTS-Russell_Putka_Waters_Paper.pdf

Sander, W. L., and Wright, S. P. (2008) A Response to Amrein-Beardsley (2008) "Methodological Concerns About the Education Value-Added Assessment System. A SAS White paper.

Sanders, W. L., & Horn, S. P. (1994). The Tennessee Value-Added Assessment System (TVAAS): Mixed-model methodology in educational assessment. *Journal of Personnel Evaluation in Education, 8*, 299-311.

Sanders, W. L., & Horn, S. P. (1998). Research findings from the Tennessee Value-Added Assessment System (TVAAS) database: Implications for educational evaluation and research. Journal of Personnel Evaluation in Education, 12(3), 247-256.

Sass, T. R. (2008) The stability of value-added measures of teacher quality and implications for teacher compensation policy. Calder, Brief 4, November.

Schafer, W. D., Hou, X., & Lissitz, R. W. (2009). *Consideration of test score reporting based on cut scores* (Technical report). College Park, MD: MARCES.

Strauss, V. (2010). What Rhee's successor should do first. The Answer Sheet. http://voices.washingtonpost.com/answer-sheet/dc-schools/a-job-for-the-new-dc-schools-c.html

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

United States Department of Education (2009). *Race to the Top Program Executive Summary.* Washington, DC.

Webster, W. J., & Mendro, R. L. (1994). *Identifying and rewarding effective schools: The Dallas School Acountability Program.* Paper presented at the Center for Research on Educational Accountability and Teacher Evaluation (CREATE), National Evaluation Institute, Gatlinburg, TN.

Webster, W. J., & Mendro, R. L., Bembry, K. L., & Orsak, T. H. (1995). *Alternative methodologies for identifying effective schools.* Paper presented at the Annual Meeting of the American Educational Research Association, San Francisco, CA.

Wright, S. P., White, J. T., Sanders, W. L., Rivers, J. C. (March 25, 2010) SAS EVAAS Statistical Models. SAS Institute Inc. World Headquarters.

Yeatman, J. (2010). St. Mary's, Charles teacher evaluations to be tied to student scores. *Washington Post.* http://www.washingtonpost.com/wp-dyn/content/article/2010/08/10/AR2010081005877.html

Yumoto, F. (2011) Effects of unmodeled latent classes on multilevel growth mixture estimation. Dissertation.

Table 1

Study Design

We were given 3 years of data on the same students, linked to their teachers.

We were given data from 3, 4, 5, 6, 7, and 8th grade. We divided the students into four cohorts.

Cohort 1: 3, 4, and 5th grade on the same students

Cohort 2: 4, 5, and 6th grade on the same students

Cohort 3: 5, 6, and 7th grade on the same students

Cohort 4: 6, 7, and 8th grade on the same students.

Sample size is around 5000 per cohort.

Mathematics and Reading Performance scale scores were obtained.

The data were from 2008, 2009, and 2010. That gave us two growth calculations

We studied 9 models for growth in the first two years and 11 models defining growth for the second two years.

None of the data were vertically scaled.

Table 2

Data used in our study

Variable	Label
QRG1	Quantile regression with one predictor
QRG2	Quantile regression with two predictor
ConD	Deciles conditional on deciles
ConZ	Z scores conditional on deciles
OLS1	Ordinary least squares with one predictor
OLS2	Ordinary least squares with two predictors
OLSS	Ordinary least squares using spline scores
DIFS	Difference between spline scores
TRSG	Transition model with values reflecting both status and growth
TRUG	Transition model reflecting upward growth only
TRUD	Transition model reflecting upward and downward change
Pre	Pretest
Post	Posttest
Sub1	Math subscore 1 (Algebra) or Reading subsocre1 (General Reading)
Sub2	Math subscore 2 (Geometry and Measurement) or Reading subscore 2 (Literatry Reading)
Sub3	Math subsocre 3 (Statistics and Probability) or Reading subscore 3 (Informational Reading)
Sub4	Math subscore 4 (Numbers and Computations)
Sub5	Math subscore 5 (Processes)
Gender	Male – 0 ; female – 1
SPED	Special Ed : no – 0 ; yes – 1
LEP	ELL code : no – 0 ; yes – 1
FARMs	Free and reduced meals : no – 0 ; yes – 1
ACC	Accommodated : no – 0 ; yes – 1
Indian	No – 0 ; yes – 1
Asian	No – 0 ; yes – 1
African	No – 0 ; yes – 1
White	No – 0 ; yes – 1
Hispanic	No – 0 ; yes – 1

Table 3
TRSG (rewards students for their status and their growth)

Table 3
Value Table for TRSG

	B1	B2	B3	P1	P2	P3	A1	A2	A3
B1	9	11	13	15	17	19	21	23	25
B2	8	10	12	14	16	18	20	22	24
B3	7	9	11	13	15	17	19	21	23
P1	6	8	10	12	14	16	18	20	22
P2	5	7	9	11	13	15	17	19	21
P3	4	6	8	10	12	14	16	18	20
A1	3	5	7	9	11	13	15	17	19
A2	2	4	6	8	10	12	14	16	18
A3	1	3	5	7	9	11	13	15	17

Table 4
TRUD: (rewards students for their academic progress and penalizes them for their achievement regress but does not reward for status)

Table 4
Value Table for TRUD

	B1	B2	B3	P1	P2	P3	A1	A2	A3
B1	0	0.5	1	1.5	2	2.5	3	3.5	4
B2	-1	0	0.5	1	1.5	2	2.5	3	4
B3	-1	-1	0	0.5	1	1.5	2	2.5	3
P1	-2	-1	-1	0	0.5	1	1.5	2	3
P2	-2	-2	-1	-1	0	0.5	1	1.5	2
P3	-3	-2	-2	-1	-1	0	0.5	1	2
A1	-3	-3	-2	-2	-1	-1	0	0.5	1
A2	-4	-3	-3	-2	-2	-1	-1	0	1
A3	-4	-4	-3	-3	-2	-2	-1	-1	0

Table 5
TRUG (rewards growth, but does not distinguish amounts of growth above the next several performance level and does not punish for regressing)

Table 5
Value Table for TRUG

	B1	B2	B3	P1	P2	P3	A1	A2	A3
B1	1	3	4	4	4	4	4	4	4
B2	0	1	3	4	4	4	4	4	4
B3	0	0	1	3	3	4	4	4	4
P1	0	0	0	1	2	3	4	4	4
P2	0	0	0	0	1	2	4	4	4
P3	0	0	0	0	0	1	3	3	4
A1	0	0	0	0	0	0	1	2	3
A2	0	0	0	0	0	0	0	1	2
A3	0	0	0	0	0	0	0	0	1

Figure 1

A representative Scree Plot

Scree Plot for Math 2008-2009

Cohort 1

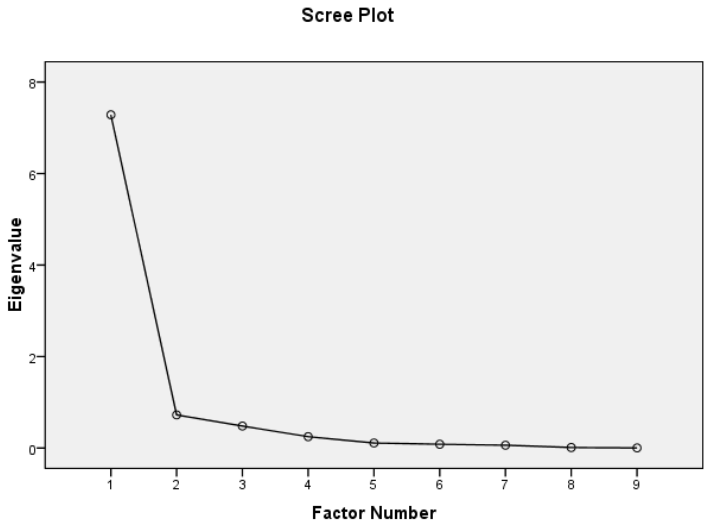


Table 6

Correlation between Math and Reading Student Growth Scores

Year 2008-2009

	Cohort 1	Cohort 2	Cohort 3	Cohort 4
Scale score[1]	0.74	0.66	0.64	0.65
QRG1	0.25	0.17	0.19	0.15
ConD	0.25	0.18	0.19	0.16
ConZ	0.27	0.19	0.23	0.17
OLS1	0.26	0.18	0.23	0.16
OLSS	0.25	0.19	0.20	0.14
DIFS	0.20	0.14	0.15	0.08
TRSG	0.42	0.30	0.36	0.32
TRUG	0.13	0.03	0.11	0.07
TRUD	0.16	0.07	0.14	0.08

Year 2009-2010

	Cohort 1	Cohort 2	Cohort 3	Cohort 4
Scale score[1]	0.69	0.67	0.68	0.65
QRG1	0.19	0.22	0.18	0.16
QRG2	0.18	0.21	0.18	0.14
ConD	0.19	0.22	0.18	0.16
ConZ	0.20	0.24	0.20	0.18
OLS1	0.20	0.24	0.18	0.17
OLS2	0.19	0.23	0.18	0.16
OLSS	0.21	0.25	0.16	0.15
DIFS	0.17	0.16	0.09	0.09
TRSG	0.30	0.40	0.35	0.32
TRUG	0.05	0.12	0.08	0.08
TRUD	0.07	0.16	0.09	0.08

[1] These correlations are between Math and Reading scale scores of each cohort in 2009 and 2010, respectively. They are not growth scores.

Note the Results for TRSG

Table 7

Correlation between 2008-2009 and 2009-2010 Student Growth Scores

Stability of Student Growth measures

Math

	Cohort 1	Cohort 2	Cohort 3	Cohort 4
Scale Score[2]	0.84	0.83	0.86	0.87
QRG1	-0.30	-0.25	-0.29	-0.25
ConD	-0.24	-0.20	-0.24	-0.19
ConZ	-0.23	-0.20	-0.23	-0.18
OLS1	-0.31	-0.26	-0.32	-0.28
OLSS	-0.31	-0.24	-0.32	-0.28
DIFS	-0.45	-0.36	-0.44	-0.36
TRSG	0.18	0.32	0.33	0.37
TRUG	-0.39	-0.28	-0.37	-0.29
TRUD	-0.48	-0.33	-0.46	-0.33

Reading

	Cohort 1	Cohort 2	Cohort 3	Cohort 4
Scale score[2]	0.77	0.68	0.69	0.71
QRG1	-0.25	-0.26	-0.26	-0.27
ConD	-0.19	-0.22	-0.22	-0.21
ConZ	-0.20	-0.19	-0.18	-0.19
OLS1	-0.26	-0.26	-0.26	-0.26
OLSS	-0.27	-0.22	-0.25	-0.27
DIFS	-0.42	-0.47	-0.44	-0.50
TRSG	0.12	0.02	0.03	0.03
TRUG	-0.36	-0.37	-0.38	-0.38
TRUD	-0.38	-0.45	-0.46	-0.46

[2] These correlations are between 2009 and 2010 post-test scale scores of each cohort on Math and Reading, respectively. They are not growth scores.

Note the results for TRSG

Table 8 Teacher Intraclass Correlations[3] for Year 2008-2009

	Math				Reading			
	Cohort 1	**Cohort 2**	**Cohort 3**	**Cohort 4**	**Cohort 1**	**Cohort 2**	**Cohort 3**	**Cohort 4**
QRG1	0.48	0.44	0.39	0.44	0.42	0.37	0.28	0.26
ConD	0.47	0.44	0.38	0.43	0.43	0.37	0.28	0.28
ConZ	0.47	0.45	0.38	0.44	0.43	0.38	0.31	0.28
OLS1	0.48	0.45	0.39	0.45	0.42	0.37	0.31	0.27
OLSS	0.47	0.45	0.41	0.43	0.42	0.39	0.32	0.26
DIFS	0.44	0.41	0.38	0.40	0.37	0.34	0.29	0.23
TRSG	0.55	0.60	0.51	0.57	0.53	0.42	0.37	0.35
TRUG	0.41	0.36	0.33	0.35	0.35	0.34	0.31	0.23
TRUD	0.42	0.40	0.35	0.37	0.36	0.33	0.28	0.24
# of Teacher[4]	292	262	96	120	268	107	122	122
# Mean	19.48	21.13	57.99	48.26	21.29	51.79	45.39	47.18
# SD	15.66	14.00	82.32	89.14	14.43	31.78	25.62	26.90

For Year 2009-2010

	Math				Reading			
	Cohort 1	**Cohort 2**	**Cohort 3**	**Cohort 4**	**Cohort 1**	**Cohort 2**	**Cohort 3**	**Cohort 4**
QRG1	0.44	0.35	0.38	0.36	0.33	0.29	0.30	0.24
QRG2	0.42	0.35	0.38	0.38	0.32	0.27	0.25	0.22
ConD	0.43	0.36	0.37	0.36	0.34	0.29	0.31	0.26
ConZ	0.43	0.36	0.38	0.37	0.34	0.31	0.31	0.26
OLS1	0.44	0.38	0.38	0.37	0.34	0.30	0.31	0.25
OLS2	0.42	0.38	0.39	0.39	0.33	0.29	0.26	0.21
OLSS	0.44	0.40	0.38	0.36	0.34	0.28	0.28	0.25
DIFS	0.40	0.40	0.35	0.32	0.30	0.23	0.23	0.20
TRSG	0.59	0.50	0.59	0.57	0.42	0.34	0.34	0.34
TRUG	0.36	0.32	0.31	0.31	0.31	0.20	0.22	0.20
TRUD	0.40	0.36	0.32	0.31	0.32	0.22	0.23	0.19
# of Teacher[4]	306	283	94	103	291	91	97	95
# Mean	18.33	16.97	50.61	49.27	19.33	53.81	48.84	53.61
# SD	8.99	9.24	34.50	38.35	12.11	33.19	30.60	34.49

[3] This ICC is also the square root of the traditional ICC.

[4] The last three rows are descriptive statistics of sample size of teachers. The number of teacher is the number of teachers for each cohort. The number mean is the average number of students taught by each teacher and number SD is the SD of the number of students attached to each teacher.

Table 9

Year to Year Reliability of Teacher Effectiveness[5]

Between 2008-2009 and 2009-2010

	Math			Reading		
	Grade 5	Grade 6	Grade 7	Grade 5	Grade 6	Grade 7
QRG1	0.42	0.73	0.50	0.28	0.51	0.61
ConD	0.44	0.73	0.52	0.31	0.53	0.63
ConZ	0.46	0.74	0.56	0.36	0.53	0.68
OLS1	0.47	0.75	0.55	0.34	0.49	0.67
OLSS	0.43	0.72	0.52	0.32	0.58	0.58
DIFS	0.42	0.65	0.50	0.13	0.08	0.30
TRSG	0.61	0.82	0.73	0.42	0.71	0.68
TRUG	0.36	0.58	0.53	0.22	0.01	0.34
TRUD	0.40	0.62	0.50	0.20	0.10	0.29
# of Teacher	177	69	82	185	57	55

Note the results for TRSG

5 The teacher effectiveness measures based on the growth scores of students in the same grade in consecutive years are correlated and presented here.

Table 10 School Intraclass Correlation[6] for Year 2008-2009

	Math				Reading			
	Cohort 1	Cohort 2	Cohort 3	Cohort 4	Cohort 1	Cohort 2	Cohort 3	Cohort 4
QRG1	0.37	0.36	0.28	0.34	0.31	0.27	0.21	0.18
ConD	0.37	0.36	0.27	0.34	0.32	0.27	0.20	0.19
ConZ	0.37	0.36	0.26	0.34	0.32	0.27	0.22	0.18
OLS1	0.38	0.35	0.26	0.34	0.31	0.27	0.23	0.18
OLSS	0.36	0.36	0.27	0.35	0.30	0.28	0.24	0.18
DIFS	0.36	0.34	0.26	0.24	0.26	0.25	0.27	0.16
TRSG	0.37	0.41	0.31	0.43	0.38	0.27	0.25	0.22
TRUG	0.31	0.29	0.24	0.19	0.25	0.22	0.28	0.16
TRUD	0.32	0.31	0.24	0.22	0.26	0.23	0.24	0.17
# of School[7]	103	102	27	28	103	100	27	27
# Mean	55.23	54.27	206.19	206.82	54.47	48.03	176.19	187.96
# SD	15.61	13.91	76.78	91.74	10.38	12.53	30.27	28.97

Year 2009-2010

	Math				Reading			
	Cohort 1	Cohort 2	Cohort 3	Cohort 4	Cohort 1	Cohort 2	Cohort 3	Cohort 4
QRG1	0.35	0.26	0.28	0.25	0.24	0.19	0.19	0.16
QRG2	0.34	0.26	0.27	0.28	0.24	0.19	0.17	0.15
ConD	0.34	0.25	0.28	0.25	0.24	0.19	0.20	0.17
ConZ	0.35	0.26	0.28	0.26	0.24	0.20	0.20	0.17
OLS1	0.34	0.27	0.27	0.25	0.24	0.19	0.20	0.16
OLS2	0.34	0.28	0.27	0.28	0.25	0.20	0.18	0.14
OLSS	0.35	0.28	0.27	0.24	0.23	0.20	0.18	0.16
DIFS	0.33	0.28	0.22	0.20	0.21	0.19	0.14	0.14
TRSG	0.40	0.31	0.37	0.40	0.26	0.23	0.21	0.20
TRUG	0.26	0.24	0.19	0.21	0.21	0.17	0.13	0.14
TRUD	0.30	0.26	0.20	0.21	0.23	0.17	0.14	0.13
# of School[7]	103	27	27	28	103	27	27	27
# Mean	55.40	205.22	205.07	205.57	54.61	181.37	175.44	188.63
# SD	15.17	73.19	76.85	93.65	15.48	73.50	82.69	90.40

[6]The definition of the ICC value is presented on pp.47 of the full report. It is the square root of the traditional ICC.
[7] The last three rows are descriptive statistics of sample size of schools. The number of school is the number of schools for each cohort. The number mean is the average number of students within each school and number SD is the SD of the number of students within each school.

Table 11

Year to Year Reliability of School Effectiveness[8]

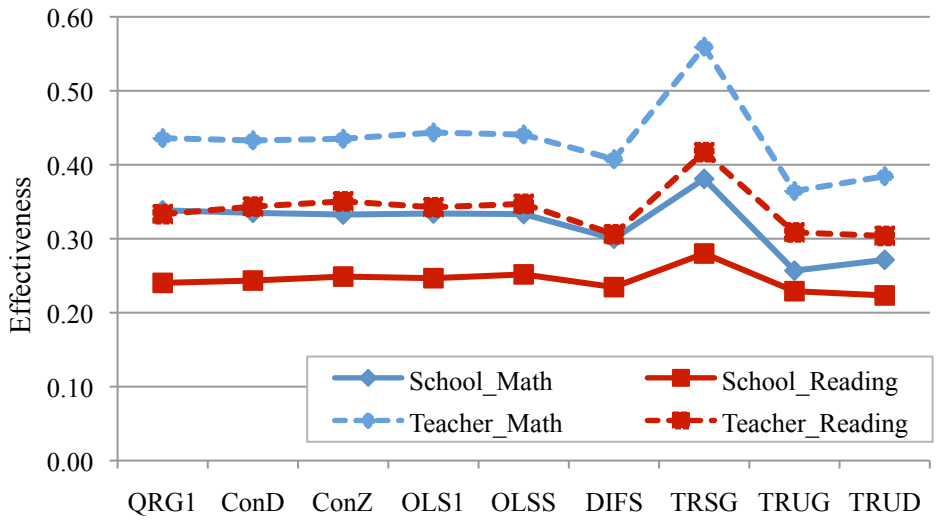
Between 2008-2009 and 2009-2010

	Math			Reading		
	Grade 5	Grade 6	Grade 7	Grade 5	Grade 6	Grade 7
QRG1	0.53	0.77	0.60	0.33	0.74	0.37
ConD	0.53	0.77	0.61	0.39	0.72	0.43
ConZ	0.55	0.76	0.61	0.41	0.72	0.43
OLS1	0.58	0.76	0.63	0.37	0.76	0.44
OLSS	0.56	0.78	0.62	0.46	0.79	0.28
DIFS	0.48	0.77	0.30	0.25	0.86	-0.15
TRSG	0.79	0.86	0.90	0.61	0.81	0.53
TRUG	0.52	0.75	0.23	0.31	0.88	-0.20
TRUD	0.53	0.73	0.31	0.30	0.89	-0.21
# of School	101	27	27	99	27	27

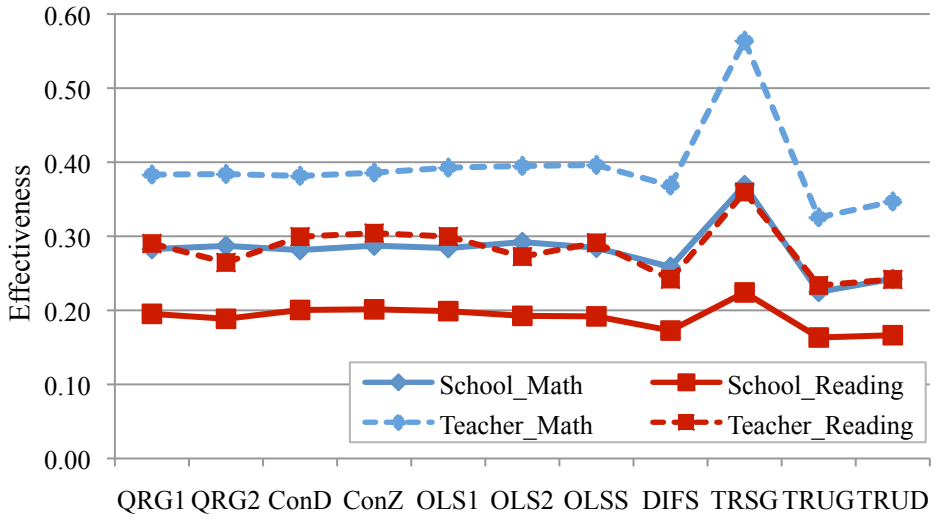
[8] The school effectiveness measures based on the growth scores of students in the same grade in consecutive years are correlated here

Figure 2 Levels of Effectiveness

Year 2008-2009



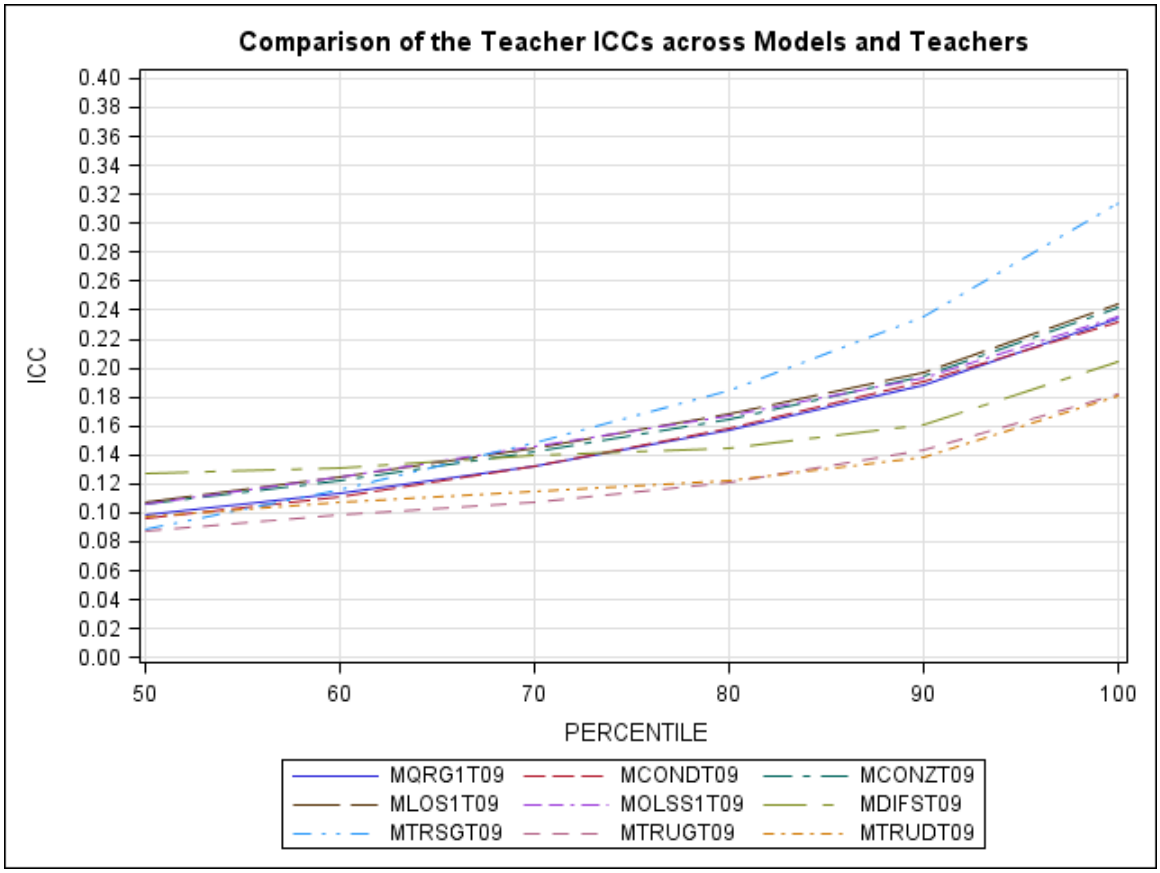
Year 2009-2010



Note the results for TRSG

Figure 3 The Effect of reducing the number of lower performing teachers

Math cohort 1 in Year 2008-2009[9]



[9] These are the original ICC values and are not the square root used in many of the tables above.