

Value-Added Modeling of Teacher Effectiveness:

An Exploration of Stability across Models and Contexts

Xiaoxia Newton, Linda Darling-Hammond

Edward Haertel, and Ewart Thomas

Abstract: Recent policy interest in tying student learning to teacher evaluation has led to growing use of value-added methods for assessing student learning gains linked to individual teachers. VAM analyses rely on complex assumptions about the roles of schools, multiple teachers, student aptitudes and efforts, homes and families in producing measured student learning gains. This article reports on analyses that examine the stability of high school teacher effectiveness rankings across differing conditions. We find that judgments of teacher effectiveness for a given teacher can vary substantially across statistical models, classes taught, and years. Furthermore, student characteristics can impact teacher rankings, sometimes dramatically, even when such characteristics have been previously controlled statistically in the value-added model. A teacher who teaches less advantaged students in a given course or year typically receives lower effectiveness ratings than the same teacher teaching more advantaged students in a different course or year. Models that fail to take student demographics into account further disadvantage teachers serving large numbers of low-income, limited English proficient, or lower-tracked students. We examine a number of potential reasons for these findings, and we conclude that caution should be exercised in using student achievement gains and value-added methods to assess teachers' effectiveness, especially when the stakes are high.

Value-Added Modeling of Teacher Effectiveness:

An Exploration of Stability across Models and Contexts

Growing interest in tying student learning to educational accountability has stimulated unprecedented efforts to use high-stakes tests in the evaluation of individual teachers and schools. In the current policy climate, pupil learning is increasingly conceptualized as standardized test score gains, and methods to assess teacher effectiveness are increasingly grounded in what is broadly called value-added analysis. The inferences about individual teacher effects many policymakers would like to draw from such value-added analyses rely on very strong and often untestable statistical assumptions about the roles of schools, multiple teachers, student aptitudes and efforts, homes and families in producing measured student learning gains. These inferences also depend on sometimes problematic conceptualizations of learning embodied in assessments used to evaluate gains. Despite the statistical and measurement challenges, value-added models for estimating teacher effects have gained increasing attention among policy makers due to their conceptual and methodological appeal.

While prior research provides some evidence concerning the accuracy and the stability of estimated teacher "value-added" effects, few studies have looked comprehensively and systematically at the variability of teacher effect estimates obtained using alternative models or using data from the same teachers over time or across different course offerings. In addition, many of the knotty issues associated with claims and measures of teacher effectiveness have as yet received little systematic treatment in the literature.

As both researchers and policy makers increasingly seek to use different sorts of teacher effectiveness measures, it is important to investigate the “hidden” judgments and dilemmas embedded in various approaches, including often tacit conceptualizations of teaching and assumptions about the sources of variation and influence affecting teaching practice. The most important of these assumptions may be the very existence of large, stable "teacher effects" independent of statistical modeling details or teaching contexts.

The present paper reports an empirical investigation of the stability of teacher effectiveness ratings based on Valued-Added Modeling (VAM), and of factors that might affect such stability. The analysis is intended to test the kinds of VAM that are likely to be used in the near-term by states or schools given the requirements of Race to the Top and kinds of data currently available in most locales: From a measurement perspective, these data are generally based on tests of specific grade-level standards, rather than vertically-scaled assessments designed to measure student growth reliably across consecutive grades (Anderman et al., 2010; Good et al., 2010) – an approach that has been reinforced by federal requirements under No Child Left Behind. From a data perspective, few states can tie teacher identities to large samples of students for whom they can track multiple years of data. From a methodological perspective, few states or districts have developed sophisticated models that use a full range of statistical controls or hierarchical methods.

While this study does examine a range of models, including some that are more sophisticated than those used by states and districts, we work with data sets that closely resemble those in most states as we examine the influences of different models, teaching years, and teaching contexts on teachers’ effectiveness ratings.

Review of Literature

Questions related to teacher effectiveness have a long intellectual history within the broader field of research on teaching and teacher education, as well as research on school effectiveness (Doyle, 1977; Raudenbush and Willms, 1995). Throughout its history, however, research on teacher effectiveness has found few consistent relationships between teacher variables and effectiveness measures, typically operationalized as student test scores (e.g., Barr, 1961; Morsh and Wilder, 1954; Rosenshine, 1970). Consequently, earlier scholars (e.g., Brophy, 1973) argued against “the indiscriminate use of student gain or general achievement tests for assessing teacher accountability” and stressed that “[u]ntil the sources of instability become known and are eliminated or controlled, this procedure is inappropriate, and certainly unfair to many teachers” (p. 251).

Despite such cautions, there has been a resurgence of interest in defining and measuring "teacher effectiveness." And, while the use of simple gain scores has long been questioned (e.g., Cronbach & Furby, 1970; Linn & Slinde, 1977), more sophisticated forms of gain score analysis addressing some of the earlier critiques have emerged in the context of what is broadly called value-added modeling (VAM).

Conceptually, VAM's promise of quantifying the “added value” that teachers and/or schools produce in terms of student learning offers intuitive appeal. Methodologically, VAM is attractive because it appears to offer a way to disentangle the effects of teachers and/or schools from those of other uncontrolled factors such as students’ demographic and socioeconomic characteristics, family education and language background, and neighborhood environment (income, employment, and the like).

However, isolating the teacher and school effects can be difficult because these other factors include omitted variables and variables that are imperfectly measured.

Furthermore, the unobservable mechanisms by which teachers are assigned to schools, pupils are assigned to schools, and pupils within schools are assigned to classrooms are nonrandom. Consequently, omitted and imperfectly measured factors vary systematically across teachers and schools, creating extreme methodological challenges to making definitive causal inferences about teachers' effectiveness based on their student scores (Kupermintz, 2003; Linn, 2008; Raudenbush and Willms, 1995; Rubin et al., 2004).

Despite its conceptual and methodological appeal, the use of VAM to estimate teacher effectiveness or to rank teachers for high stakes purposes poses daunting challenges stemming from many factors: the non-random assignment of students to teachers and schools, the limitations of particular tests both for measuring the full range of desired knowledge and skills and for measuring learning gains, and the difficulties of disentangling the contributions of many influences on learning – multiple teachers, parents, tutors, specific curricula, and the availability of useful learning materials, as well as other school resources like time, class sizes, and the like (For reviews, see Baker et al., 2010; Braun, 2005, and McCaffrey et al., 2005.) Furthermore, analysts have found that teachers' measured effectiveness can be quite different when results are measured on different tests (Lockwood et al., 2007) or when different statistical methods are used (Rothstein, 2007).

The analyst employing VAM models faces complex decisions involving trade-offs between competing values. One such decision that has been a point of debate in the VAM literature (McCaffrey et al., 2005) is whether student background variables should

be included as covariates in the models in addition to prior-year test scores. For example, the most prominent example of the VAM application, the Tennessee Valued Added Assessment System (TVAAS) (Sanders and Horn, 1998), does not control for student characteristics. The reason for this, developers of TVAAS have argued, is that TVAAS uses student gains to measure teacher effectiveness, which implicitly controls for socioeconomic status and other background variables that are related to initial levels of achievement (Ballou, Sanders, and Wright, 2004). This argument has raised concerns among educational scholars (Amrein-Beardsley, 2008; Kupermintz, 2003), who question why the effects of important student characteristics variables should have vanished in the TVAAS model. McCaffrey and his colleagues (2005) note that, “the importance of modeling student background characteristics when using VAM to estimate teacher effects remains an empirical question that must be addressed...” (p. 70).

Another important issue is related to disentangling school effects from effects of teachers. Some VAM applications (e.g., the TVAAS layered model) omit school effects from the models. One likely consequence is biased estimation of teacher effects. Researchers such as McCaffrey and colleagues (2005) argued that this bias might be reduced by including separate predictor variables for each school in the value-added regression models, i.e., by using the so-called school fixed effects model. However, these authors caution that additional empirical investigation is needed to determine the extent to which including or excluding school fixed effects changes inferences about teacher effectiveness. While school fixed effects may control for unmeasured aspects of the environment that differ across schools (for example, differences in class sizes, curriculum materials, availability of instructional supports, or the competence of

principals and peers), they also adjust away any school-level differences in average teacher quality. School fixed effects essentially compare teachers within schools only. Thus, with school fixed effects in the model, comparisons among teachers in different schools rest on the assumption that on average, all schools routinely hire equally capable teachers, an assumption that is unsupported in many cases.

A third fundamental issue VAM faces concerns the stability of teacher effects across time, an issue that has long been of interest to various researchers (e.g., Brophy, 1973; Doyle, 1977; and Rosenshine, 1970). These researchers' reviews of studies on teacher effectiveness were primarily concerned with the question of "whether a teacher who is effective or ineffective once is equally effective or ineffective a second time" (Rosenhine, 1970, p. 647). McCaffrey and his colleagues (2005) found a moderate correlation between value-added teacher rankings in different years for elementary and middle school mathematics teachers in several Florida school districts. However, Sass (2008) found considerable instability in VAM rankings of teachers in an analysis of five urban districts across the country. For example, among those ranked in the lowest quintile of effectiveness in one year, only 25% to 35% were similarly ranked a year later, depending on the district, while a comparable proportion had moved up to the top two quintiles. Among those initially ranked in the top quintile of effectiveness, only 20% to 30% were similarly ranked a year later, while a comparable proportion had dropped to the bottom two quintiles. In sum, empirical work addressing the consistency of teacher effectiveness rankings over time is inconclusive, and this important issue bears further empirical investigation (Campbell et al., 2004; McCaffrey et al., 2005).

A fourth issue is related to how the curriculum is structured, particularly at the secondary level. U.S. high school teachers typically offer different courses to different groups of students during any given year, and course assignment policies may dictate that a teacher typically teaches high-level courses to high-achieving students or less challenging courses to low-achieving students. Naturally, an important question to consider when thinking about teacher effectiveness is whether a given teacher is equally effective across different types of classes. Empirical investigation of this issue is relatively thin, both because the dominant underlying assumption is that a teacher's effectiveness is constant regardless of the content and/or classes he or she is teaching (Campbell et al., 2004), and because it is unusual to have data available that permit this kind of comparison.

In this paper, we address these important issues, examining the extent to which “teacher effectiveness” ratings are stable across different statistical models, across classes or courses that teachers teach, and across years.

Method

Study Context and Sample

This investigation is part of a longitudinal study that examines the relationships among teaching variables and pupil learning as part of the Teachers for a New Era (TNE) research initiative at Stanford University. In this broader study, we examined a sample of approximately 250 secondary teachers and roughly 3500 students taught by these teachers. All were from a set of six high schools in the San Francisco Bay Area. Because California did not at this time have a state longitudinal data system, student and teacher data had to be secured from individual schools and districts' electronic data files.

The present study focused only on the mathematics and English language arts teachers, because the curricular course sequence for science and social studies, evaluated through end-of-course examinations that measure only the standards for those domains, does not allow for a systematic study of pupil learning gains over the years. It is difficult to construct a series of value-added gains in chemistry when chemistry concepts do not appear on the other science tests and when students' chemistry courses may occur before or following integrated science, physics or biology courses, which are taken in no standard sequence across schools. On the other hand, English language arts and mathematics courses are generally taken in a reasonably common sequence in most California high schools and have some overlapping constructs and skills from year to year, if not perfect alignment.

Tables 1 and 2 describe the sample and the types of data that formed the basis of the value added analysis for the current study.

[Tables 1 and 2 Go About Here]

Conceptualization of Teacher Effectiveness

We base our measurement of “value added” on the variation in pupils' test scores on the California Standards Tests (CSTs), controlling for prior-year scores, rather than on variation in year-to-year test score gains, because the CSTs are not vertically scaled and, therefore, do not yield interpretable gain scores. Although the CSTs do use Item Response Theory to create scale scores, these scores are not vertically equated in California.

We recognize the problems with the use of such non-vertically equated tests for VAM purposes. This lack of vertical equating is also true in most other States, as only a

minority currently have tests that are vertically scaled across grade levels. While the use of such non-vertically scaled tests is a drawback for research studies of value-added methods, we believe analyses with extant data of this sort reflect contemporary realities in the field, as all 50 states are required under Race to the Top rules to use their current testing data for the evaluation of teachers. We discuss this issue more fully later.

We use ordinary least square (OLS) regression analyses to predict pupils' CSTs after taking into consideration prior year's achievement (CST scores in the same subject area). Some of our models also control for key demographic background variables (i.e., race/ethnicity, gender, free/reduced lunch status, English language learner status, and parent education), and some include school fixed effects. Additionally, we test a multi-level mixed-effects model to take into account the ways in which students are nested within classrooms and teachers are nested within schools.

With these different statistical controls, a teacher's effectiveness is then measured by the average difference between actual scores and predicted scores for all students assigned to that teacher (i.e., the average of the residual scores). This measure of teacher effectiveness has the advantage of transparency and is conceptually similar to estimates of the teacher fixed effects in more sophisticated VAM regression models. We recognize that stronger statistical controls would be possible using two or more prior years of student data for each teacher. However, the practical limitations of many district and state data systems, the high levels of student mobility in many districts, and the policy requirement that teachers from as many tested grade levels as possible be included in VAM analyses suggest that VAM implementations for some years to come will be limited to a single prior year of data.

Linking Students to Teachers

For each of the schools in our study sample, we obtained student course enrollment files. Based on these course files, we linked individual pupils with teachers from whom they took the English language arts or mathematics courses for both fall and spring semesters (i.e., during the entire academic year). Additionally, when a teacher was teaching different courses to different groups of students (e.g., algebra 1 and geometry; or regular English and honors English), we generated separate value-added estimates for the teacher (one for each course).

Because California students take different high school courses, each with its own end-of-course examination (e.g., algebra 1, geometry, algebra 2, etc.), and CST scale scores are not directly comparable across different course-specific tests, scale scores from each CST were converted to z scores prior to performing these regressions. We transformed raw scale scores into z-scores based on the sample mean and standard deviation of a particular grade (for English language arts, where students take grade-level tests each year) or of a particular subject test (for math, where students take subject-specific tests). In addition to enabling the pooling of prior-year scores across different CSTs, this linear transformation of raw scale scores also facilitated the presentation of study outcomes in a standardized metric.

Data Analysis

The data analyses were designed to investigate whether teacher rankings were consistent across different models, across different courses/classes for teachers who taught multiple types of ELA or math courses, and across two years for teachers for

whom we had three waves of pupil test scores. The analysis consisted of three major stages.

Estimation of Value-Added Models. First, we conducted a series of parallel ordinary least square (OLS) linear regressions with and without student controls and school fixed effects, separately for math and for ELA, and for years 2006 and 2007, respectively. These OLS analyses generated four residual (observed minus predicted) scores for each student. These residual scores for each student were aggregated to the teacher (or course within teacher) level. Based on the aggregated residual scores, teachers were assigned “effectiveness” rankings for each of the OLS models. Each teacher was assigned four rankings in each year for which the teacher had data, one for each of the OLS models that produced the aggregated residual gain scores. For each of the OLS models, separate rankings were assigned for ELA versus math teachers, and for 2006 versus 2007 student outcomes. (See table 3.)

In addition to these four models (i.e., OLS regression with or without student characteristics and with or without school fixed effect), we conducted a multilevel mixed-effect model with teacher as a random effects factor and school as a fixed effects factor (i.e., Model 5 in Table 3). Estimates of teacher effects under this model are Empirical Bayes estimators which take into account how much information is available from each cluster unit. In other words, the estimate for a teacher with a smaller number of students would be "shrunk" toward the overall mean to a greater extent than for a teacher with a greater number of students.

[Table 3 Goes About Here]

Models 1 and 2 – both of which control for prior achievement, and the second of which adds controls for student characteristics -- are most similar to those used thus far by states and districts involved in value-added modeling. Models 3 and 4 add school fixed effects, offering an approach that can be carried out with relative ease and is common in the research literature, but one that has rarely been used on-the-ground in the field for teacher evaluation purposes. As we describe later, this approach has the benefit of controlling for unobserved differences across schools but the pitfall of comparing teachers only within schools and making the unrealistic assumption that teacher quality is randomly distributed across schools, which can create a countervailing bias. Thus, its use would be debatable if a district or state were trying to compare teachers across a jurisdiction. Fewer analysts and virtually no states or districts currently use multilevel models like that employed in our Model 5. Such approaches require both more sophisticated statistical methods and tools and a potentially greater numbers of teachers and schools in the analysis than might be available in many small districts.

Some scholars have argued that value-added estimates can be improved by incorporating more than one prior test scores in the regression model (e.g., TVAAS). Having more information on a student's prior achievement certainly is an advantage; however, putting more than one year's prior test scores in the model has two disadvantages. First, the use of two prior years of data further limits the already small subset of teachers for whom value-added estimates can be obtained, because only teachers of students tested at each of two earlier grade levels can be included. (Note that use of two prior years of data is also highly unrealistic at the high school level, where annual testing is the exception not the rule, and where most courses are not part of three-

year sequences.) Second, insisting on putting more than one year's prior test scores in the model could lead to more missing data than using just one year's prior test score (i.e., the immediate adjacent year's pretest scores), because year-to-year test score linkage is imperfect, and because it is not uncommon for students to enter and exit the school system in a given year, especially in urban school settings. Given these trade-offs, we believe that our choice of using the immediately adjacent year's pretest scores is the most consistent with the models most likely to be seriously entertained in most states. In addition, this choice also helps to avoid the problem of discarding more cases than necessary from the analysis due to missing data problems (i.e., students with only one prior year's pretest scores would be dropped out from the analysis).

Analyses of Relationships. During the second stage, several types of descriptive and correlational analyses were conducted, using teacher ranks. These analyses included: (1) Spearman rank correlations (Spearman's rho) among teacher ranks using different models, (2) Pearson correlations between different teacher ranks and the student compositions of their classes (e.g. proportion of English language learners, free / reduced meal program participants, mean parental educational level, etc.), and (3) variation in teacher decile ranks across years and across courses. We use decile ranks to examine the extent of stability in teacher rankings, because deciles offer a familiar reporting metric, with units large enough to represent meaningful performance differences but fine-grained enough to avoid obscuring important information.

Examination of Within-Teacher Variance. Finally, for math and ELA teachers who taught the same sets of courses within the same school, a series of analyses of

variance (ANOVAs) was conducted to quantify the relative sizes of effects of teacher, course, and the interaction between teacher and course.

Findings

How Models Matter

To investigate variation in the estimates of teacher effectiveness using the five different models and what factors might be related to the variation, we examined: (1) Spearman rank correlations among teacher ranks using different models; and (2) the correlation between teacher ranks generated under the five models and key student demographic characteristics to evaluate whether some statistical models might be more sensitive to the composition of the student body than others.

Not surprisingly, teacher ratings from the four models were highly correlated with one another in both mathematics and English language (See table 4 showing data from 2007. Correlations were similar for 2006.) There were somewhat larger differences in rankings between random effects and fixed effects models than there were between models with and without controls for student demographics, within these categories.

[Table 4 Goes About Here]

Examinations of the patterns of correlation coefficients for math 2006, math 2007, ELA 2006 and ELA 2007 suggested that teacher ranks generated by the five models were significantly related to: (1) student racial / ethnic background, (2) student socioeconomic status proxies, including meal program participation and parents' educational level, and (3) proxy measures of student mathematics ability (on track or on the fast track) for mathematics and of student English language status for ELA.

Even though three of the five models controlled for student demographics as well as students' prior test scores, teachers rankings were nonetheless significantly and negatively correlated with the proportions of students they had who were English learners, free lunch recipients, or Hispanic, and were positively correlated with the proportions of students they had who were Asian or whose parents were more highly educated. (See Table 5.) In addition, English teachers were more highly ranked if they had a greater proportion of girls in their classes, and math teachers were more highly ranked if they had more students in their classes who were on a "fast track" in mathematics (that is, taking a given course at an earlier grade than it is generally offered in the school's usual math sequence). While the correlations with student demographics were generally slightly lower for the models that had controlled for student demographics in producing the rankings (M2, M4, and M5), they were still statistically significant. This suggests either that teachers who were teaching greater proportions of more advantaged students may have been advantaged in their effectiveness rankings, or that more effective teachers were generally teaching more advantaged students.

[Table 5 Goes About Here]

Furthermore, the size of the differences between teachers' rankings across models was also significantly related to student demographics, especially in English language arts, indicating that teachers with more African American, Hispanic, low-income, or limited English proficient students and teachers whose students' parents had lower levels of education were more likely to be ranked significantly lower when student demographics were not taken into account in the VAM model.

To evaluate this, we computed the differences in rankings generated under model 1 and under each of the other models (i.e., M2 to M5) for each teacher, and then we correlated these difference estimates with student demographic variables. In ELA, for example, the differences in teacher rankings between Model 1 and those of Models 3 through 5 were significantly correlated with the percentage of African American students taught. (Pearson r ranged from $-.38$ to $-.63$, $p < .01$, for 2006 and from $-.39$ to $-.58$, $p < .01$, for 2007). The patterns were similar for other demographic variables (e.g., Hispanic, Asian, ELL, meals, and parent education), although not all correlations were statistically significant.

The presence of significant correlations between teacher effectiveness rankings and the demographic composition of classrooms may signal the compositional or contextual effects that Bryk & Raudenbush (1992, 2002), among others, have described. Their research finds that individual students' achievement is affected not only by their individual background characteristics, but also by the characteristics of other students in their class. This has implications for estimating teacher value-added effects on student learning. For instance, a teacher teaching a class in which most students come from highly educated families might have higher value-added scores because each individual student's learning is boosted by presence of other well-supported and highly motivated students. Our analysis of relationships between classroom demographic characteristics supports the hypothesis of compositional or contextual effects on value-added achievement gains, as does our analysis of teachers teaching similar pairs of classes within the same school, which will be discussed in the next section.

How Classes Matter

An implicit assumption in the value-added literature is that measured teacher effects are stable across courses and time. Previous studies have found that this assumption is not generally met for estimates across different years. There has been less attention to the question of teacher effects across courses. One might expect that teacher effects could vary across courses for any number of reasons. For instance, a mathematics teacher might be better at teaching algebra than geometry, or an English teacher might be better at teaching literature than composition. Teachers may also be differentially adept at teaching new English learners, for example, or 2nd graders rather than 5th graders. It is also possible that, since tracking practices are common, especially at the secondary level, different classes might imply different student compositions, which can impact a teacher's value-added rankings, as we saw in the previous section.

To examine teachers' effectiveness rankings across classes, we examined the correlation between teacher ranks across courses for teachers who taught more than one type of course. Table 6 summarizes the intra-class correlation coefficients representing the extent to which a teacher's ranking for one course was related to his or her ranking for a different course. None of the correlations for the OLS models (models 1 to 4) was statistically significant, and most of the correlations were negative. The preponderance of negative correlations in our data is likely due to fact that, in some of the high schools we studied, teachers who teach two different math courses or two different English courses are assigned to one upper-track course and one lower-track course. Because in this sample, score gains tend to be larger in upper track courses, even after controlling for prior achievement, a teacher who gets a high (low) ranking for one course is likely to get a low (high) ranking for the other course.

The multi-level model generated one significant positive correlation for mathematics in 2007, suggesting that math teachers were similarly ranked in both the classes they taught that year. However, the model generated negative correlations for mathematics in 2006 and for ELA in both years, similar to the results from the other models.

[Table 6 Goes About Here]

We also ran a total of 16 ANOVAs using the student residualized achievement scores for teachers who taught the same set of math or ELA courses within the same school in each of two large comprehensive high schools. Table 7 displays the ANOVA table from one of the 16 analyses. This Table shows results for three mathematics teachers who were from one of the large comprehensive high schools and who taught the same pair of math courses, Geometry and Algebra 1, during the 2006-07 year. As shown in Table 7, the two statistically significant predictors of students' current mathematics achievement scores were: (1) students' prior mathematics achievement scores, and (2) the math course. In contrast, neither teacher nor the teacher-by-course interaction effect was related to students' residualized mathematics achievement scores.

This was true in most of the 16 ANOVA analyses: Students' prior achievement was a significant predictor in all cases, and the specific class was a significant predictor most of the time (in 11 of the 16 ANOVAs). However, the identity of the teacher was predictive in only three of the sixteen analyses, as was the teacher – course interaction. (See Table 8.) In other words, the “teacher effect” was generally less strong as a predictor of student achievement gains than the “course effect”: Student achievement

varied significantly across courses taught by the same teachers more often than it varied across teachers who taught the same courses within the same school.

[Tables 7 and 8 Go About Here]

Furthermore, the analyses suggested that teachers' rankings were higher for courses with "high-track" students than for untracked classes. Figure 1 shows rankings for three English teachers from the same large comprehensive high school when they taught the same pairs of courses, one designed for higher-track students and the other for lower-achieving students. Each teacher appeared to be significantly more effective when teaching upper-track courses (with rankings falling in the 7th to the 9th deciles) than the same teacher appeared when teaching low-track courses (with rankings falling in the 1st to the 3rd deciles).

[Figure 1 Goes About Here]

How Teacher Rankings Vary Across Years

To address the question of how teacher rankings vary across years, we correlated teacher rankings based on the five models for 2005-06 and 2006-07 with CST scores during these two years. The inter-year correlations are shown in Table 9. There was a modest correlation of about 0.4 for ELA teachers, regardless of the model used to derive the teacher rankings. For math teachers, the correlations are lower when Models 1 and 2 (without school fixed effects) are used, and moderate (i.e., about 0.6) when Models 3, 4, and 5 (school fixed effects models and the multilevel mixed-effect model) are used.

If the 2006 and 2007 rankings are conceived of as two replications of a teacher effectiveness measurement procedure, then these correlations may be interpreted as estimates of reliability coefficients. As can be seen, none of these correlations

approaches the level of reliability customarily demanded as a basis for consequential decisions affecting individuals.

[Table 9 Goes About Here]

Stability of Teacher Rankings

Together, our results indicate that teacher rankings vary somewhat across models, and vary substantially across courses and years. To further examine the extent to which teacher rankings are stable, we calculated the percentage of teachers who had stable rankings, whose rankings changed by one or more deciles in either direction, by two or more deciles, or by three or more deciles across the four models, across courses, or across years. Table 10 summarizes the results.

[Table 10 Goes About Here]

As evident in Table 10, teacher rankings fluctuated across models, courses, and years, and the fluctuation was greatest across courses and years. In some cases, there were teachers whose rankings changed as much as 8 deciles depending on the course they taught. The same patterns were observed for both math and ELA teachers. Although our results are based on small numbers of teachers, we have no reason to think the results would be markedly different with larger groups.

As noted earlier, teacher ranks generated by the five models were significantly related to key student characteristics, including racial / ethnic background, socioeconomic status, parents' educational level, math ability as indicated by fast-track status for mathematics courses, and English language status for ELA courses.

To illustrate just how dramatically student characteristics can impact teacher rankings, Figure 2 displays the student characteristics associated with a mid-career

English language arts teacher in one of the sampled large comprehensive high schools, whose ranking changed across two years from the bottom decile (1) to the top decile (10). In the first year, this teacher was teaching a course in which 58% of students were English language learners, 75% were Latino, and 42% were eligible for free or reduced price lunch. In the second year, by contrast, only 4% of the students were classified as English language learners and the proportions who were Latino and low-income were about half as much as the year before. The parent education level was also significantly higher in year two, with the average parent having at least some college as compared to the average parent education of year one students having less than a high school education. In this instance, the instability of the teacher effectiveness rankings appeared to be at least in part associated with student race/ethnicity, English language learners status, poverty, and parents' educational level.

[Figure 2 Goes About Here]

These examples and our general findings highlight the challenge inherent in developing a value-added model that adequately captures teacher effectiveness, when teacher effectiveness itself is a variable with high levels of instability across contexts (i.e., types of courses, types of students, and year) as well as statistical models that make different assumptions about what exogenous influences should be controlled. Further, the contexts associated with instability are themselves highly relevant to the notion of teacher effectiveness.

Discussion

Value added models (VAM) have attracted increasing attention in the research and policy community as a hoped-for means to isolate the effect of teachers upon student

learning, disentangling the effect of teachers' efforts from other powerful factors. Many hope that if VAM can isolate teacher effects on student learning, then various personnel decisions can be based on the teacher effect estimates. This study provided an opportunity to explore the possibilities and limitations of value-added methods for evaluating teacher effectiveness. We focused on testing the kinds of VAM that are likely to be used in the near-term by states and districts given the requirements of Race to the Top and data currently available to states or schools.

Our study confronted many of the limitations of current efforts to use VAM in many states: the lack of a statewide data system designed to support such studies, and the consequent need to assemble data sets from individual schools and districts; missing data created by high student and teacher mobility, especially in low-income communities; and the need to develop means for estimating pre- and post-measures of pupil learning in a context where states and districts lack fall-to-spring measures that are vertically scaled and reflect the full range of learning goals.

The measurement issues we confronted include the challenges of using end-of-course examinations for courses that are not linked by a clear learning progression. While we were able to make measurement adjustments to estimate learning gains in mathematics and English language arts, this was not possible to do with measurement integrity in fields like science, where students are measured at the end of the year on content in different courses (biology, chemistry, physics, integrated science) that are not taken in a standard order and are not constructed to represent a progression of learning.

The strengths of our study included: (1) matching students with teachers at the course level, which allowed us to investigate the stability of teacher ranking across

different courses; (2) matching students with teachers from whom they studied for the entire academic year, which avoided the potential problem of attributing a student's learning to a teacher when the student had not been under the care of the teacher over the entire academic year; and (3) studying teachers at the high school level, which illuminated several practical and conceptual issues that have important policy implications for using value added estimates to hold schools or teachers accountable for student performance.

This exploratory research does not offer definitive answers to questions concerning sources of instability in estimates of effectiveness. However, it offers some indications about potential sources of influence on measured student achievement gains that may have implications for the conduct of value-added analysis and the assessment of teacher effectiveness.

Influences of Student Characteristics on Value-Added Estimates

Through our analysis, we observed that judgments of teacher effectiveness depend on the statistical model used, on which specific class or course was examined, and on which year's data are used. The most important differences among statistical models were whether they controlled for student demographics and school fixed effects. The use of a multi-level model did not generally change outcomes noticeably. Teacher effectiveness ratings fluctuated most extensively by course and year. As shown by the examples in figures 1 and 2, student characteristics were found to impact teacher rankings, sometimes dramatically.

Even in models that controlled for student demographics as well as students' prior test scores, teachers rankings were nonetheless significantly and negatively correlated

with the proportions of students they had who were English learners, free lunch recipients, or Hispanic, and were positively correlated with the proportions of students they had who were Asian or whose parents were more highly educated. In addition, English teachers were more highly ranked when they had a greater proportion of girls in their classes, and math teachers were more highly ranked when they had more students in their classes who were on a “fast track” in mathematics (that is, taking a given course at an earlier grade than it is generally offered in the school’s usual math sequence).

While the correlations with student demographics were generally slightly lower for the models that had controlled for student demographics in producing the rankings, they were still statistically significant. Furthermore, the magnitude of the differences among teachers’ rankings across models was also significantly related to student demographics, especially in English language arts: Teachers with more students who were African American, Hispanic, low-income, limited English proficient students, and whose parents had lower levels of education were more likely to be ranked significantly lower when student demographics were omitted from the VAM model.

These findings suggest that teachers who were teaching greater proportions of more advantaged students may have been advantaged in their effectiveness rankings, or that more effective teachers were generally teaching more advantaged students. Our data set allowed us to test this possibility with a set of teachers who taught both upper-track and lower-track courses in the same year. In those analyses we found that students’ residualized achievement scores were, in most analyses, more strongly predicted by the students’ prior achievement and the course they were in than by the teacher him or

herself. Each teacher appeared to be significantly more effective when teaching upper-track courses than the same teacher appeared when teaching lower-track courses.

The default assumption in the value-added literature is that teacher effects are a fixed construct that is independent of the context of teaching (e.g., types of courses, student demographic compositions in a class, and so on) and stable across time. Our empirical exploration of teacher effectiveness rankings across different courses and years suggested that this assumption is not consistent with reality. In particular, the fact that an individual student's learning gain is heavily dependent upon who else is in his or her class, apart from the teacher, raises questions about our ability to isolate a teacher's effect on an individual student's learning, no matter how sophisticated the statistical model might be.

Our the correlations indicate that even in the most complex models, a substantial portion of the variation in teacher rankings is attributable to selected student characteristics, which is troubling given the momentum gathering around VAM as a policy proposal. Even more troubling is the possibility that policies that rely primarily on student test score gains to evaluate teachers – especially when student characteristics are not taken into account at all (as in some widely used models) -- could create disincentives for teachers to want to work with those students with the greatest needs.

The widespread acceptance of the default assumption that teacher effectiveness is a fixed construct, largely unaffected by other variables, is demonstrated by the recent action of the Los Angeles *Times*, which published a data base of 6000 teachers' value-added rankings. This action, while deplored by many researchers, was applauded by many policy advocates in commentary on the *Times* stories. (See, for example, opinions

posted at <http://www.nytimes.com/roomfordebate/2010/09/06/assessing-a-teachers-value>). In the LA *Times* articles on VAM, a Question and Answer section regarding the published analysis states unequivocally that teachers' scores will not be affected by "low-performing students, English-language learners or other students with challenges." (<http://www.latimes.com/news/local/la-me-qanda-20100816,0,4120439.story>).

This statement is not supported by our analysis, nor is there any indication in the report of the value-added analysis conducted for the *Times* that additional steps, beyond the kind we took, were taken to control for student characteristics. Buddin's model controlled for "gender, race, parent's education, special attitudes [sic] and needs" (Buddin, 2010, p. 4), and did not include controls for free or reduced lunch status and English learner status, as our study did. Neither does Buddin's technical report describe any use of student fixed effects (which could provide even stronger controls on student characteristics by essentially comparing students' value-added gains against their own learning trajectories in other years) or any examination of compositional or contextual effects (Buddin, 2010). Thus, there is no reason to believe the influences of student characteristics that we found would not be as large or larger in the LA *Times* study.

Buddin's model, like ours, adjusts students' test scores using just one prior year of data. Our analyses base each estimate of an effect for a teacher on data from a single year, whereas Buddin averages data over all available years (up to 7) for a given teacher. (Note that Buddin's use of multiple years of data is not the same as using more than one prior year's test score for an individual student. Only one prior score is used for each student in calculating an estimate for a given teacher. Whereas Buddin's pooling of

multiple years of data may reduce or mask the instability of teacher rankings, it cannot reduce any systematic bias due to omitted variables or model misspecification.)

Implications of the Findings

There are several ways to think about the implications of these empirical findings. Conceptually and theoretically, we might need to broaden our definition of teacher effectiveness from a generic perspective to a differentiated perspective, acknowledging that teacher effectiveness is context specific rather than context free. Several researchers in the United Kingdom (e.g., Campbell, Kyriakides, Muijs, and Robinson, 2004) have argued for developing a differentiated model for assessing teacher effectiveness which considers that teachers might be more effective teaching some students than others. The results of our empirical investigations are consistent with this line of thinking on teacher effectiveness.

However, it is also possible that a substantial share of what some would call a “teacher effect” actually measures other factors that are correlated with student characteristics. These might include aspects of students’ learning contexts that influence the rate of learning, which might include the influences of prior knowledge on ability to profit from specific grade level instruction, engagement factors like attendance and time for homework that may influence both prior and current achievement, the availability of parent help and/or tutoring, or even the curriculum and class sizes offered to students of different measured ability levels.

It is also possible that students’ gains and teachers’ effects are less well-measured by existing standardized tests for some student populations than others. For example, in California, where we conducted these studies, new immigrant students who have less

than a year of English language learning opportunity must take the same tests in English as other students, with fewer language modifications and accommodations than are permitted in other states.

In addition, tests geared strictly to grade-level standards may not measure the gains that students exhibit who begin the year far below (or for that matter, far above) grade level, as the areas and extent of their actual learning growth may not be measured on the test. This might be particularly true for students with exceptional needs, as well as new immigrants entering with little formal education, or others who have fallen behind academically. Thus, teachers who have large populations of such students in their classroom may appear less effective than they in fact are. These teachers' effectiveness might be better reflected in vertically-scaled tests that measure a much more extended continuum of learning, including the kind of "out-of-grade" testing currently prohibited by federal rules under No Child Left Behind.

Practically, the notion that the contexts of teaching (which are contributing factors to the instability of teacher effectiveness measures) are integral to the conception of teacher effectiveness has important implications for policy and practice. For instance, policies advocating the use of test scores to hold teachers accountable for student learning may need to take the contexts of teaching and the characteristics of students into consideration. They may also need to consider the development and use of adaptive student tests that measure a broader range of learning gains, and that do so validly for special populations of students.

The use of better-designed tests and more thoughtful statistical methods (including those that control for student characteristics, school effects, and the nested

nature of teaching and learning within schools and classrooms) could help to address some, but not all, of these concerns. Meanwhile, however, the policies that are embedded in Race to the Top guidance and a number of other federal and state laws expect - and in many cases require - that student gain scores be used to evaluate teachers without vertically scaled tests (currently in existence in a minority of states), implying the use of very simple gain measures that poorly control for student characteristics or school effects and in no case include hierarchical modeling. And these judgments about teachers must be made with small data sets in the many small school districts that predominate in most states. Thus, the issues that surfaced in our data from California - even as we have incorporated more sophisticated measures than most states plan to use - reflect the issues that will arise in the real world.

Furthermore, better tests, data systems, and statistical strategies alone will not solve the problems of measuring teacher effects on learning. We believe that the focus on improving data quality should include, as well, a commitment to increasing the synergies and alignment among curriculum, assessment, instruction, and accountability. Without a “growth-oriented” curriculum in which learning in later grades builds on that in earlier grades for each subject area, even vertically-scaled assessments are unable to measure real gains. Similarly, all available statistical methods are based on untestable and probably untenable assumptions about how the world works. Until our schooling system is set up in such a way as to allow for random assignment between students, teachers, and schools, statistical models that attempt to estimate “what would have happened for a student with teacher A in school 1 if he or she were with teacher B in

school 1 or teacher C in school 2” can do just so much, no matter how sophisticated they are. (For an excellent discussion, see Rubin et al., 2004).

By exploring these methodological challenges of measuring teacher effectiveness in terms of pupil achievement, we have tried to clarify some of the implications of various measures and approaches for research on teacher effectiveness and for policy on teacher accountability. Our conclusion is NOT that teachers do not matter. Rather, our findings suggest that we simply cannot measure precisely how much individual teachers contribute to student learning, given the other factors involved in the learning process, the current limitations of tests and methods, and the current state of our educational system. Other studies are needed to evaluate these issues further, and to develop strategies for taking into account the various factors that may influence student achievement gains, so that that effects of teachers on student learning can be properly understood.

Value-Added Modeling of Teacher Effectiveness

Table 1
Teacher and Student Samples for the VAM Analyses

Sample	2005-06	2006-07
Mathematics teachers	57	46
ELA teachers	51	63
Students		
Grade 9	646	881
Grade 10	714	693
Grade 11	511	789

Note. Some teachers taught multiple courses. There were 13 such math teachers for year 2005-06 and 10 for year 2006-07; and there were 16 such ELA teachers in 2005-06 and 15 in 2006-07.

Table 2
List of Variables

Variables	Scale
<i>Outcome measures:</i> CST math or ELA	CST scale scores were transformed to Z scores
<i>Student prior achievement:</i> CST math or ELA in previous year On track status (for math) Fast track status (for math)	CST scale scores were transformed to Z scores. Variable indicating that a student took a math course at the usual grade level it is offered in the school Variable indicating that a student took a math course at an earlier grade level than it is usually offered in the school
<i>Student demographic background:</i> Race or ethnicity Gender English language learners	Indicator variables for African American, Hispanic, Native American, Pacific Islander, or Asian Indicator variable for female Indicator variable for English language learner
<i>Student social economic status proxies:</i> Parent educational level Meal program	Indicator variable for high school or above Ordinal measure (0-4) from less than high school to education beyond college Indicator variable for free or reduced lunch meal participation
<i>School differences:</i>	Dummy indicator variable for each school

Table 3
 Predictors of CST Scores used in OLS Regression and Multilevel Mixed Effect Models

Models	Predictors
Model 1 (M1)	Prior achievement only
Model 2 (M2)	Prior achievement plus student characteristics
Model 3 (M3)	Model 1 plus school fixed-effect
Model 4 (M4)	Model 2 plus school fixed-effect
Model 5 (M5)	Three-level mixed-effect model with same predictors

Value-Added Modeling of Teacher Effectiveness

Table 4
Spearman's Rho Correlations of VAM Rankings across Models Math and ELA 2007

Models	M1		M2		M3		M4		M5	
	Math	ELA	Math	ELA	Math	ELA	Math	ELA	Math	ELA
M1	1.00	1.00	.93**	.89**	.91**	.90**	.82**	.83**	.76**	.79**
M2			1.00	1.00	.89**	.86**	.89**	.92**	.85**	.86**
M3					1.00	1.00	.92**	.92**	.85**	.92**
M4							1.00	1.00	.95**	.94**
M5									1.00	1.00

** $p < .01$.

Value-Added Modeling of Teacher Effectiveness

Table 5
Significant Correlations between Teachers' VAM Rankings and Their Students' Characteristics, 2007

	ELL		Meal		Asian		Hispanic		Parent Ed.	
	Math	ELA	Math	ELA	Math	ELA	Math	ELA	Math	ELA
M1	-.38 ^{***}	-.48 ^{***}	-.27 [*]	-.45 ^{***}	.27 [*]	.31 ^{***}	-.33 ^{**}	-.43 ^{***}	.34 ^{**}	.48 ^{***}
M2	-.37 ^{***}	-.31 ^{***}	-.25 [*]	-.20 [*]	.18	.24 [*]	-.27 [*]	-.26 ^{**}	.28 ^{**}	.32 ^{**}
M3	-.37 ^{***}	-.42 ^{***}	-.30 ^{**}	-.30 ^{**}	.31 ^{**}	.30 ^{**}	-.35 ^{**}	-.39 ^{***}	.35 ^{**}	.38 ^{***}
M4	-.31 ^{**}	-.31 ^{**}	-.31 ^{**}	-.18	.24 [*]	.31 ^{**}	-.32 ^{**}	-.30 ^{**}	.32 ^{**}	.31 ^{**}
M5	-.29 ^{**}	-.36 ^{***}	-.34 ^{**}	-.22 [*]	.29 ^{**}	.29 ^{**}	-.34 ^{**}	-.34 ^{***}	.34 ^{**}	.32 ^{***}

* $p < .10$. ** $p < .05$. *** $p < .01$.

Value-Added Modeling of Teacher Effectiveness

Table 6
Intra-class Correlations of Teacher Rankings across Courses

Models	Outcome	Math 06 (n=13)	Math 07 (n=10)	ELA 06 (n=16)	ELA07 (n=15)
M1		-.39	.05	-.54	-.29
M2		-.14	.29	-.34	.14
M3		-.11	.05	-.65	-.38
M4		-.14	.41	-.41	-.25
M5		-.16	.72*	-.52*	-.47

* $p < .05$.

Table 7

Analysis of Variance for Geometry and Algebra 1 Tests of Between-Subjects Effects—
Dependent Variable: CST math 06 Z scores

Source	Type III Sum of Squares	<i>df</i>	Mean Square	<i>F</i>	<i>p</i>
	96.37 ^a	7	13.77	27.69 ^{***}	.000
Prior year test score	65.31	1	65.31	131.34 ^{***}	.000
Teacher ID	2.09	2	1.04	2.10	.125
class	5.67	1	5.67	11.39 ^{***}	.001
Teacher ID * class	.18	2	.09	.18	.835
Error	94.98	191	.50		
Total	191.35	198			

Note. $R^2 = .50$ (Adjusted $R^2 = .49$).

^a The 7th *df* refers to the intercept for ANOVA.

*** $p < .01$.

Table 8
Significant Factors in ANOVA

Factors Predicting Student CST Scores	Number of Times Significant
Prior student achievement	16
Course	11
Teacher	3
Teacher by Course	3

Value-Added Modeling of Teacher Effectiveness

Table 9
Correlations of Teacher Rankings across Years

Models	Outcomes	Math (n=27)	ELA (n=31)
M1		.45*	.34
M2		.43*	.39*
M3		.63**	.39*
M4		.62**	.43*
M5		.59**	.48**

* $p < .05$. ** $p < .01$.

Value-Added Modeling of Teacher Effectiveness

Table 10
Percent of Teachers Whose Effectiveness Rankings Change

	By 1 or more Deciles	By 2 or more Deciles	By 3 or more Deciles
Across models ^a	56-80%	12-33%	0-14%
Across courses ^b	85-100%	54-92%	39-54%
Across years ^b	74-93%	45-63%	19-41%

^a Depending on pair of models compared.

^b Depending on the model used.

Figure 1

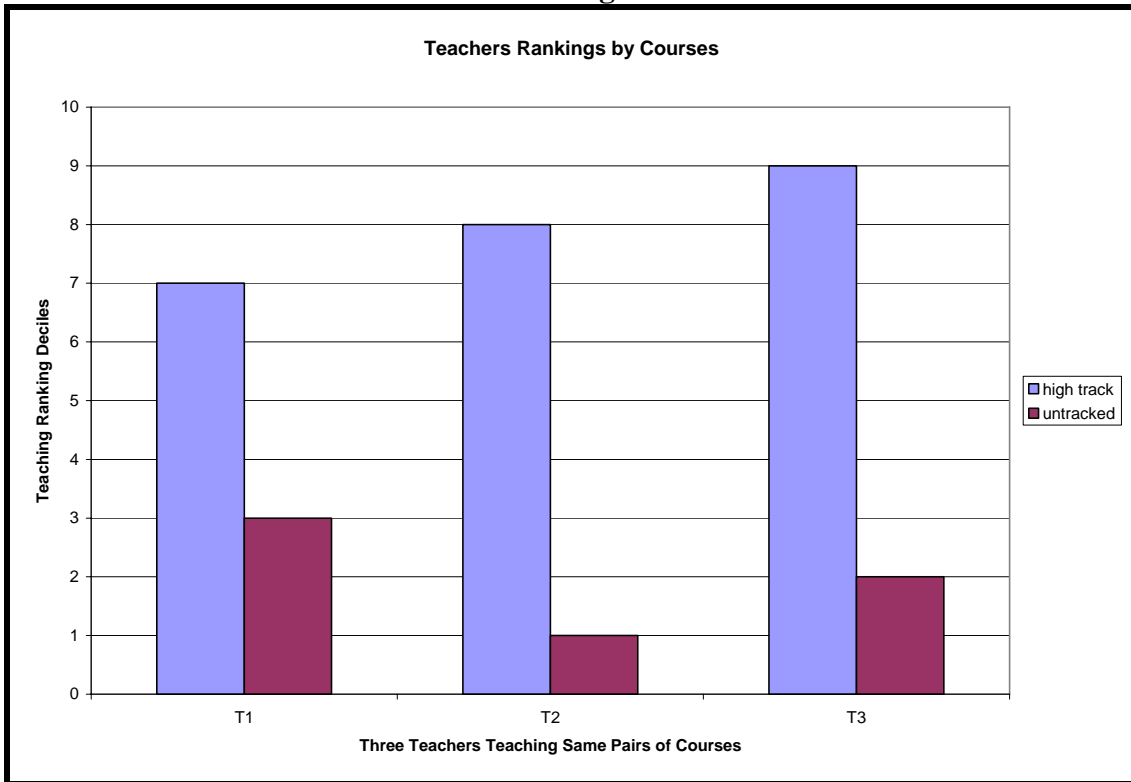
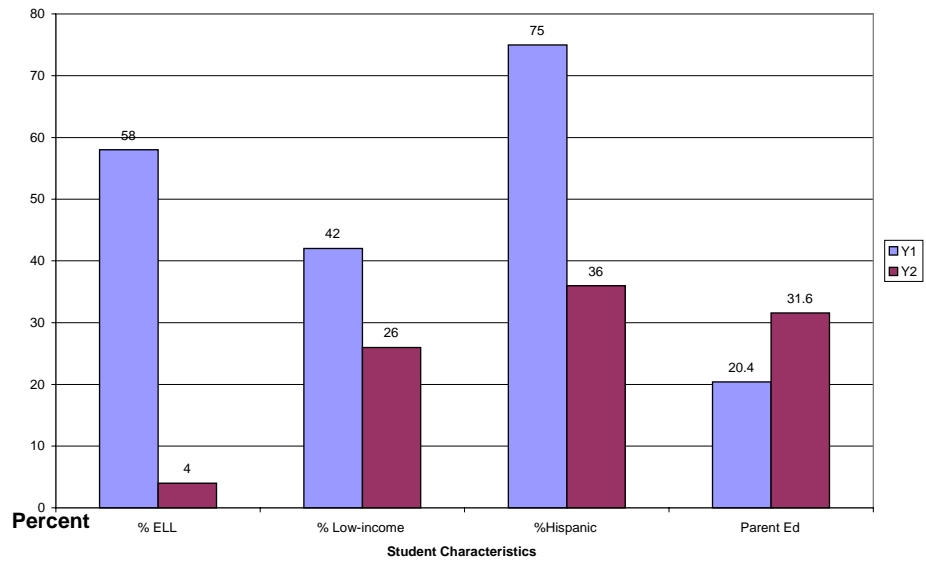


Figure 2 - Student Characteristics in Years 1 and 2 for a Teacher whose Ranking Changed from the 1st to the 10th Decile



References

- Amrein-Beardsley, A. (2008). Methodological concerns about the education value-added assessment system. *Educational Researcher*, 37(2), 65–75.
- Anderman, E. M., Anderman, L. H., Yough, M. S., & Gimbert, B. G. (2010). Value-added models of assessment: Implications for motivation and accountability. *Educational Psychologist*, 45(2), 123–137.
- Baker, E.L., Barton, P.E., Darling-Hammond, L., Haertel, E., Ladd, H.F., Linn, R.L., Ravitch, D., Rothstein, R., Shavelson, R.J., & Shepard L. (2010). *Problems with the use of student test scores to evaluate teachers*. Washington, DC: Economic Policy Institute.
- Ballou, D., Sanders, W., & Wright, P. (2004). Controlling for student background in value-added assessment of teachers. *Journal of Educational and Behavioral Statistics*, 29(1), 37–65.
- Barr, A. S. (1961). Teacher effectiveness and its correlates. In A. S. Barr, D. A. Worcester, A. Abel, C. Beecher, L. E. Jensen, A. L. Peronto, T. A. Ringness, & J. Schmid, Jr. (Eds.), *Wisconsin studies of the measurement and prediction of teacher effectiveness: A summary of investigation* (pp. 134-52). Madison, WI: Dembar Publications.
- Braun, H. I. (2005). *Using student progress to evaluate teachers: A primer on value-added models*. ETS Policy Information Center. Retrieved April 1, 2008, from http://www.ets.org/research/policy_research_reports/pic-vam.
- Brophy, J. E. (1973). Stability of teacher effectiveness. *American Educational Research Journal*, 10(3), 245–252.

- Bryk, A. S. & Raudenbush, S. W. (1992). *Hierarchical Linear Models: Applications and Data Analysis Methods (1st edition)*. Thousand Oaks, California: Sage Publications.
- Buddin, R. (2010). How Effective Are Los Angeles Elementary Teachers and Schools? Retrieved, September 10, 2010, from <http://www.latimes.com/media/acrobat/2010-08/55538493.pdf>.
- Campbell, J., Kyriakides, L., Muijs, D., & Robinson, W. (2004). *Assessing teacher effectiveness: Developing a differentiated model*. New York: RoutledgeFalmer.
- Cronbach, L. J., & Furby, L. (1970). How should we measure “change”—or should we? *Psychological Bulletin*, 74 (1), 68–80.
- Doyle, W. (1977). Paradigms for research on teacher effectiveness. *Review of Research in Education*, 5(1), 163–198.
- Good, T. L., Wiley, C. R. H., & Sabers, D. (2010). Accountability and educational reform: A critical analysis of four perspectives and considerations for enhancing reform efforts. *Educational Psychologist*, 45(2), 138–148.
- 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.
- Linn, R. L. (2008). Methodological issues in achieving school accountability. *Journal of Curriculum Studies*, 40(6), 699–711.
- Linn, R. L., & Slinde, J. A. (1977). The determination of the significance of change between pre- and post-testing. *Review of Educational Research*, 47(1): 121–150.

- Lockwood, J. R., McCaffrey, D. F., Hamilton, L. S., Stetcher, B., Le, V-N., & Martinez, J. F. (2007). The sensitivity of value-added teacher effect estimates to different mathematics achievement measures. *Journal of Educational Measurement, 44* (1), 47–67.
- McCaffrey, D. F., Lockwood, J. R., Koretz, D. M., Louis, T. A., & Hamilton, L. S. (2004). Models for value-added modeling of teacher effects. *Journal of Educational and Behavioral Statistics, 29*(1), 67–101.
- 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., Sass, T. R., Lockwood, J. R., & Mihaly, K. (2009). The intertemporal variability of teacher effect estimates. *Education Finance and Policy 4*(4), 572–606.
- Raudenbush, S. W. & Bryk, A. S. (2002). *Hierarchical Linear Models: Applications and Data Analysis Methods (2nd edition)*. Thousand Oaks, California: Sage Publications.
- Raudenbush, S. W. & Willms, J. D. (1995). The estimation of school effects. *Journal of Educational and Behavioral Statistics, 20*(4), 307–335.
- Rosenshine, B. (1970). The stability of teacher effects upon student achievement. *Review of Educational Research, 40*(5), 647–662.
- Rothstein, J. M. (2007). Do value-added models add value? Tracking, fixed effects, and causal inference (CEPS Working Paper No. 159). Princeton University and National Bureau for Economic Research (NBER).

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(1), 103–116.

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

About the Authors

Xiaoxia A. Newton

University of California, Berkeley

Xiaoxia A. Newton is an assistant professor in the division of Policy, Organization, Measurement, and Evaluation (POME), Graduate School of Education, University of California, Berkeley. Prior to Berkeley, she worked as a postdoctoral scholar for the Teachers for a New Era (TNE) research project at Stanford University. Her research focuses on using a variety of methodological tools to address educational and policy issues related to students' opportunities to learn mathematics, pipeline in Science, Technology, Engineering, and Mathematics (STEM) education, teacher learning and professional development. She can be reached at xnewton@berkeley.edu.

Linda Darling-Hammond

Stanford University

Linda Darling-Hammond is Charles E. Ducommun Professor of Education at Stanford University and faculty co-director of the Stanford Center for Opportunity Policy in Education (SCOPE). Her research, teaching, and policy work focus on issues of teaching quality, school reform, and educational equity. She has served as president of the American Educational Research Association and is a member of the National Academy of Education. Among her more than 300 publications, her most recent book, *The Flat World and Education*, outlines the policy approach needed for the United States to regain its position as a leading nation educationally. It was written while she served as director of President Obama's education policy transition team. She can be reached at ldh@stanford.edu.

Edward Haertel

Stanford University

Edward Haertel is Jacks Family Professor of Education and Associate Dean for Faculty Affairs at Stanford University, where his work focuses on educational testing and assessment. His research centers on policy uses of achievement test data; the measurement of school learning; statistical issues in testing and accountability systems; and the impact of testing on curriculum and instruction. Haertel has served as president of the National Council on Measurement in Education, a member of the National Assessment Governing Board, and as a member of the joint committee for 1999 edition of the Standards for Educational and Psychological Testing. He can be reached at haertel@stanford.edu.

Ewart Thomas

Stanford University

Ewart Thomas is a Professor of Psychology at Stanford University. He teaches undergraduate and graduate courses in Statistics and Research Methods. His research interests include the development and application of mathematical and statistical models in many areas, such as, signal detection, motivation, inter-rater reliability, parent-infant interaction, equity, and law as a social science. Thomas has served as Chair of the Psychology Department and as Dean of the University's School of Humanities and Sciences. He can be reached at ethomas@stanford.edu.