NATIONAL EDUCATION POLICY CENTER

AN ANALYSIS OF THE USE AND VALIDITY OF TEST-BASED TEACHER EVALUATIONS REPORTED BY THE *LOS ANGELES TIMES*: 2011

Catherine S. Durso

University of Denver

June 2012

National Education Policy Center

School of Education, University of Colorado Boulder
Boulder, CO 80309-0249
Telephone: (802) 383-0058

Email: NEPC@colorado.edu
http://nepc.colorado.edu

This is one of a series of briefs made possible in part by funding from
The Great Lakes Center for Education Research and Practice.

GREAT LAKES CENTER
FOR EDUCATION RESEARCH & PRACTICE
http://www.greatlakescenter.org
GreatLakesCenter@greatlakescenter.org

Kevin Welner
Project Director

William Mathis
Managing Director

Erik Gunn
Managing Editor

Briefs published by the National Education Policy Center (NEPC) are blind peer-reviewed by members of the Editorial Review Board. Visit http://nepc.colorado.edu to find all of these briefs. For information on the editorial board and its members, visit: http://nepc.colorado.edu/editorial-board.

Publishing Director: **Alex Molnar**

Suggested Citation:

Durso, C.S. (2012). *An Analysis of the Use and Validity of Test-Based Teacher Evaluations Reported by the Los Angeles Times: 2011*. Boulder, CO: National Education Policy Center. Retrieved [date] from http://nepc.colorado.edu/publication/analysis-la-times-2011.

The author would like to thank Howard Wainer for his support and many helpful discussions. Ben Domingue generously shared his insights into working with the LAUSD data. Gratitude also is extended to NEPC for the opportunity to work on this review, for their tireless pursuit of the current LAUSD data, and for their challenging and constructive editorial process.

An Analysis of the Use and Validity of Test-Based Teacher Evaluations Reported by the *Los Angeles Times*: 2011

Catherine S. Durso, University of Denver

Executive Summary

In May of 2011, the *Los Angeles Times* published, for the second time, results of statistical studies examining the variation in teacher and school performance in the Los Angeles Unified School District, based on the California Standards Tests for math and English Language Arts (ELA).[1] The studies use data from the seven academic years ending in 2009-2010. The *Times* published teachers' names along with their effect estimates. These estimates were then used to classify teachers into five categories: least effective; less effective; average; more effective; and most effective. The *Los Angeles Times* previously published the results of statistical analyses designed to address the same issues in August, 2010, using data from the period 2003-2009.

The earlier analyses were reviewed by Briggs and Domingue, who identified several serious concerns.[2] Yet the more recent analyses differ from the earlier ones in important ways and merit separate review. Accordingly, the focus in this review is on the properties of the teacher effects estimated in the more recent study.

Both collections of *LA Times* analyses were carried out by Dr. Richard Buddin. Descriptions of the work are published in the white papers "How Effective are Los Angeles Teachers and Schools?"[3] (for the August 2010 publication) and "Measuring Teacher and School Effectiveness in Improving Student Achievement in Los Angeles Elementary Schools"[4] (for the May 2011 publication). Both the 2010 and 2011 white papers address school effects and the relationship of teacher qualifications with estimated teacher effects.

The 2010 and the 2011 studies both use versions of what are known as value-added models (VAMs). Such models are gaining political favor, primarily as tools for teacher evaluation, but also for use in other personnel decisions, such as performance bonuses. Although the student-related variables differ between the 2010 publication and the 2011 publication, as do the statistical models and the criteria for including teachers in the study, there are consistent defining features of the VAM approach used in both analyses. Specifically, a statistical model is used to predict a student's current scores on a test on the basis of:

1. the student's test scores in prior years,
2. possibly additional information about the student, and
3. the identity of one of the student's teachers for the school year preceding the current test.

The value assigned to the teacher by the prediction method is often called the teacher effect, though generally the extent to which it is caused by the teacher, rather than factors out of the teacher's control, is difficult to determine. To acknowledge this difficulty explicitly, this review will use the term "teacher-linked effect." In the *Los Angeles Times* publications, the teacher-linked effect is taken at face value as a measure of the effectiveness of the teacher.

The publication of the teacher effects in Los Angeles, the recent release of teacher value-added effects for New York, and the growing movement toward use of VAMs in teacher evaluation increase the urgency of careful consideration of the interpretations these model results can sustain. For instance, one issue that must be addressed is the extent to which teachers with very different student populations are engaged in directly comparable

It is the single score of the teacher-linked effect, particularly the effectiveness classification categories published in the Teacher Ratings, that the lay reader will be drawn to, and these are unreliable.

activities.[5] Is it appropriate to compare, using a common metric, the effectiveness of teachers of struggling students to the effectiveness of teachers of high-achieving students? In a large school system such as the Los Angeles Unified School District, the differences among the groups of students in different teachers' classes can be large. For example, many of the teachers in the study work with classes with average incoming math scores rated "far below basic," and many other teachers work with classes with scores rated as "advanced." Though VAMs control for prior achievement, and may control for additional variables, teachers addressing very different needs are still compared directly without recognition of the underlying disparity in their jobs.

Another persistent concern about the estimates—and about high stakes standardized testing in general—is that it may reflect (and perhaps encourage) teaching to the test, rather than high-quality, comprehensive instruction.[6] Information accompanying the *Los Angeles Times* report points out that the effects are based on standardized tests which assess only a fraction of the content taught. If teachers can achieve better results by tailoring the form and content of their instruction to better test scores rather than more educated students, this may not only incentivize undesired behavior, but it might also limit the degree to which the estimates reflect "true teacher quality" rather than test preparation.

In addition to these issues, any specific implementation of VAMs concerning teacher performance faces the challenge of providing results that can, with reasonable precision and accuracy, distinguish between teachers and separate the effects on students' test scores attributable to the teacher, the school, the family, and other factors.

In the 2011 *Los Angeles Times* supporting documents (white papers), four different models are used to estimate teacher effects for math and English-Language Arts (ELA). Each model produces teacher-linked effects that differ significantly among teachers. There are also significant year-to-year fluctuations in the effects for a given teacher. The variability among students, after controlling for prior scores, demographic variables, and peer effects, is substantially larger than either the variability in teacher-linked effects or classroom effects. These features have been verified here.

Of the four models, it is Model 4 that is emphasized by the 2011 white paper and the *LA Times*' publication of teacher effects. This fourth model includes the most additional information about the student, the class, and the grade. It provides substantially more information than was used to generate the 2010 rankings published by the newspaper. Accordingly, this analysis focuses on Model 4.

As a result of a variety of factors, including year-to-year instability and the variability among students, Model 4 estimates teacher-linked effects with low precision. This means that the effect estimate for each teacher cannot be taken at face value. Rather, the estimate is a broad distribution of the likely effect of the teacher, represented by an "average" effect plus an "error band" which indicates the range of reasonably likely values for the effect. This is similar to the results of political polls—e.g., a person's approval rating, "plus or minus" a certain number of percentage points.

For example, Ms. Smith may be assigned an average teacher effect of 45, with an error band of +/-20, meaning that her score should be understood as somewhere in the range of 25 to 65, though more likely closer to 45 than to 25 or 65. The results of the models indicate that the error band (readers might think of this as the probable range) for many teachers is larger than the entire range of scores from the "less effective" to "more effective" designations provided by the *LA Times*.

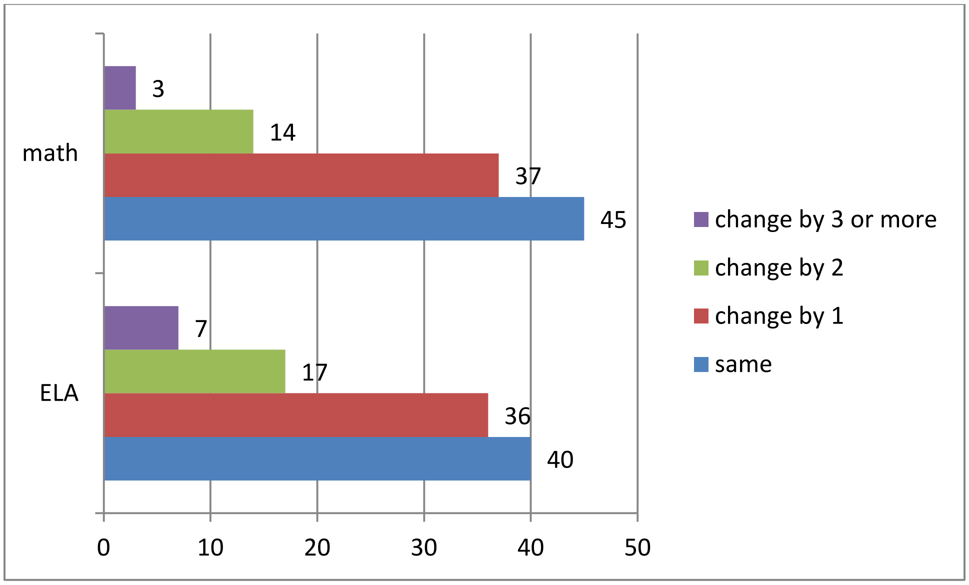

Figure 1. Percentages for rating changes of teacher-linked effects estimated for the same teacher in two separate three-year periods.

Because of this imprecision in any given year, the teacher-linked effect for individual teachers is also unstable over time. To illustrate this, the effects for teachers present in the data in all six years can be estimated using two separate three-year periods (i.e., treating teachers with six years of data as though they were two teachers with three years each). Using these three-year periods helps to judge the stability of effectiveness ratings. As shown in Figure 1, for English Language Arts only 40% of teachers fall into the same effectiveness category for the two periods, while 36% differ by one category, 17% differ by two, and 7% change by three or more categories. The stability for math is somewhat greater, with 45% staying in the same category. Note that a difference of three categories corresponds to a change from “least effective” to “more effective” or from “most effective” to “less effective.” These are not differences in degree; they are completely different conclusions.

The results are also unstable when predicting the effect in a single year from the effect averaged across multiple earlier years. Figure 2 shows the percentages of the single year effects that are in the category that would be predicted by the long-term estimate based on up to six years of data, and the percentages that differ by 1, 2, or 3 or more categories. This comparison shows how reliable the categories are for readers trying to predict an effect for a subsequent year on the basis of the published ratings. And remember that this instability is even greater when predicting “future” ratings using only one year of prior data, as is often the case, particularly among less experienced teachers.

When students are non-randomly assigned to schools and to teachers within schools, the researcher creating the statistical model attempts to include enough factors to capture all

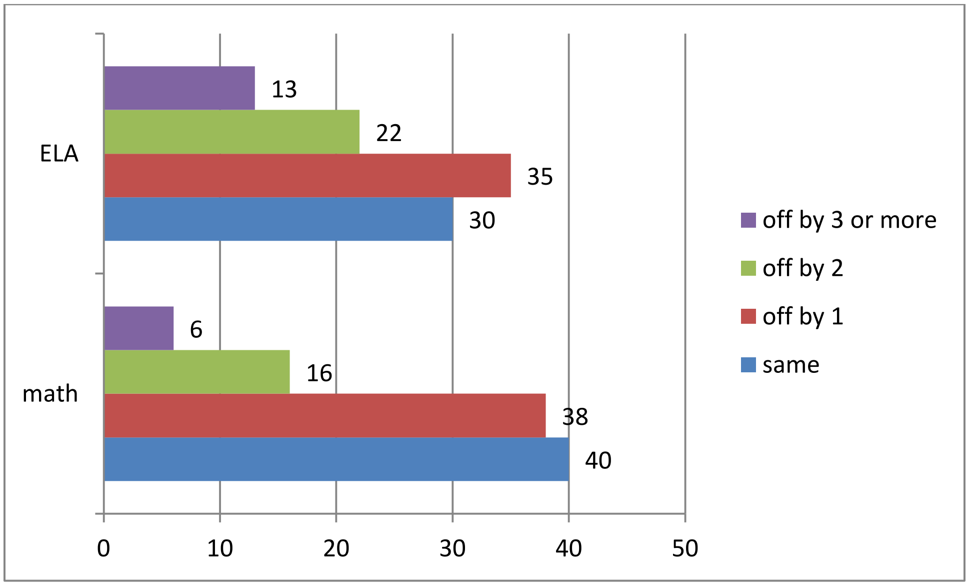

Figure 2. Percentages for rating changes of teacher-linked effects estimated for the base period and the following year.

the differences between the students in one teacher's class and the students in another teacher's class. Including the students' earlier test scores accounts for a great deal of those influences. But many important differences remain. Therefore, in addition to proper interpretation (e.g., looking at error bands), it's important to go back and verify the results by testing for hidden bias.

The extent to which Model 4 has separated effects on student learning due to the teacher from other effects is examined in several ways. Following Briggs and Domingue, the first approach is to test whether students are effectively assigned randomly to teachers with respect to variables that affect the previous year's test score but are not accounted for in the model for the previous year. There is strong evidence that this is not the case.

The second approach is based on data from teachers who changed schools. Comparison of teacher-linked effects estimated for the same teachers at two different schools show substantial changes in effect. For math, 34% of the paired effects are in the same category, while 29% differ by two or more categories. The corresponding results for ELA are that 30% are in the same category, while 33% of the pairs differ by two or more categories.

This change in estimated effectiveness among school switchers is associated with school characteristics. The difference between the average teacher-linked effect at the old school and the average at the new school is a statistically significant predictor of the difference between the teacher's effect at the old school and the teacher's effect at the new school. There is a tendency for the same teacher to have a higher teacher-linked effect at the school with the higher average effect.

To their credit, the 2011 white paper and the 2011 *LA Times'* Teacher Ratings do discuss the issues of variability and model sensitivity and provide readers with some information to enable them to take these issues into account when looking at individual teacher-linked effects. This is an improvement in the Teacher Ratings over the 2010 version.

However, it is the single score of the teacher-linked effect, particularly the effectiveness classification categories published in the Teacher Ratings, that the lay reader will be drawn to, and these are unreliable. The broad "error bands" and the cautions about model accuracy receive only peripheral attention. Thus, the way in which the *LA Times* presented the data may lead to its misuse, which can be harmful.

In summary, then:

- Teachers in the studied data have incoming classes with very different characteristics. A high teacher-linked effect must be understood as an estimate of test score improvement applying only to the range of students typically taught by that teacher.

- The model results indicate that teacher-linked effects estimates must take annual fluctuation beyond sampling error into account.

- The teacher-linked effect means and the effectiveness categories are not reliable for comparison or prediction. The large variability must be taken into account. This limits the detail available in comparisons among teacher-linked effects, particularly for effects estimated from three or fewer years of data.

- Large annual fluctuations make even effects calculated over longer time periods very approximate predictors of the effect size in the future. Parents should not rely on the published effect being reproduced in any given year.

- There is strong evidence that the teacher-linked effects include contributions to student learning not due to the teachers. These contributions are meaningful on the scale of the categories used in the *Los Angeles Times* effect report. Comparison of teacher-linked effects must be understood as comparing teachers and their work environments, not just teachers.

An Analysis of the Use and Validity of Test-Based Teacher Evaluations Reported by the *Los Angeles Times*: 2011

Introduction

In May of 2011, the *Los Angeles Times* published results of statistical models examining the variation in student performance among teachers in grades three through five and among elementary schools in the Los Angeles Unified School District based on the California Standards Tests (CST) for math and English Language Arts (ELA).[7] The studies include four models that use data from the academic years between 2004 and 2010 to predict each students' standardized test scores in a given year from the student's scores the previous year, a collection of auxiliary student-level, class-level, and grade-level data (the selection of which varies by model), and the identity of the teacher for the given year. The use of the previous year's scores makes these "value-added models" of teacher performance. The model outputs assign a distribution of values—a mean accompanied by an error margin[8]-to each teacher, which indicates the probable contribution of the individual teacher to the testing performance of his or her students. This range of values is technically termed the "effect" for that teacher, though it will reflect any factors associated with assignment to that teacher that are not accounted for by the previous year's test scores or the auxiliary data. Because the term "teacher effect" does not make explicit allowance for the possibility of other factors, the term used here will be "teacher-linked effect."

The *Los Angeles Times* publication of the study results included making 11,500 individual teacher-linked effect estimates available by teacher name on a searchable website. On the basis of the means of the teacher-linked effects, the website classifies the teachers into categories as least effective, less effective, average, more effective, and most effective, for both math and ELA.

This is not an isolated phenomenon. The *Los Angeles Times* previously, in August, 2010,[9] created a similar online database based on the results of statistical studies by the same analyst, using data from the Los Angeles Unified School District between 2003-2009. In February, 2012, the New York City Education Department released rankings of 18,000 teachers based on value-added analyses. Several New York news organizations, including *The New York Times*[10] and the *New York Post*[11] quickly made those ratings available in databases on their websites.

Reactions to the databases varied. Bill Gates, co-chair of the Bill and Melinda Gates Foundation, which supports the Measures of Effective Teaching Project, deplored this as "public shaming."[12] In contrast, a *New York Post* editorial heralded the "big victory" scored for "accountability and transparency."[13] These releases may have set a precedent that these data are public domain, suggesting that more such publications are possible. Both this possibility and the controversy surrounding the releases motivate this review of the 2011 *Los Angeles Times*' teacher ratings and researcher Richard Buddin's 2011 white paper "Measuring Teacher and School Effectiveness in Improving Student Achievement in Los Angeles Elementary Schools," which describes the study from which the teacher ratings are drawn.

Many researchers have raised concerns regarding the use of value-added models (VAM) for teacher evaluation. Briefly, VAM do not provide guidance for improvement[14], are comparative rather than absolute measures, assess a small part of teacher's responsibilities[15], force different kinds of teaching into one scale, do not produce consistent results for given teachers over time,[16] and may not identify effects actually caused by the teachers.[17]

The latter questions of variance and bias relate directly to the specific model chosen, and how its results are interpreted. Both the 2010 and the 2011 Los Angeles Teacher Ratings are based on studies carried out independently by Richard Buddin, a senior economist at the RAND Corporation. Some technical details of the work are published in attendant

> *The results suggest that teacher-linked effects do indeed vary by the school in which teachers work.*

white papers,[18] along with conclusions based on the model output. The 2010 analyses were reviewed by Briggs and Domingue.[19] That review demonstrated major sensitivity of the 2010 model to the inclusion of additional variables, a red flag for large-scale bias (i.e., unmeasured non-teacher factors influencing the results). Briggs and Domingue also demonstrated the lack of precision inherent in the *Times*' use of the estimates to assign teachers effectiveness categories.

The 2011 Buddin analyses differ from the earlier ones. This time, he presents results from four models, with increasing sets of variables (whereas the latter only estimated a single model). The additional models incorporate many of the variables used in the Briggs and Domingue sensitivity tests. Buddin's most extensive 2011 model shows little sensitivity to increasing the number of available variables. In addition, unlike the 2010 analysis, Buddin's 2011 models explicitly account for annual fluctuation in the teacher-linked effects (discussed in greater detail below).

Finally, the 2011 Los Angeles Teacher Ratings include nearly twice as many teachers as the 2010 ratings.[20] The primary reason for this is that the earlier round of estimates was only published in the database if teachers had taught at least 60 students. The new database includes all teachers who taught more than 10 students in at least one year. This leads to the inclusion of first-year teachers and others with very little data in the period of the study. (In recognition of the reduced precision of effects based on less data, the 2011 ratings information does include graphical indicators of the error margins of the teacher-linked effects, as well as the means.)

In short, then, the changes in the model and the reporting method between 2010 and 2011 mean that the analysis done by Briggs and Domingue does not apply directly to the 2011 ratings. This makes separate review of the new ratings worthwhile.

This review addresses and illustrates the variance of the new ratings—i.e., their imprecision and inconsistency—using several different approaches. First, it examines the imprecision of the teacher-linked effects in the context of the effectiveness categories used in the *LA Times*' teacher ratings, concluding that the implied accuracy of the effectiveness categories (e.g., very effective, etc.) is, in most cases, more than the model output can support.

The second approach to understanding the impact of the variance is to assess how the rankings change between years for the same teacher. The changes are substantial.

Finally, the consequences of the high variance and annual fluctuation for prediction are examined—for example, using the first six years of data to predict a teacher's rating in the seventh year. This comparison shows that the estimated persistent teacher-linked effects account for about 40% of the variation in the seventh year for math, and less than 20% for ELA. Put simply, even using multiple years of data, the estimates were only modestly useful for predicting a future year.

In addition to the imprecision and inconsistency of these estimates, there is also the question of whether they represent unbiased causal effects. Interpreting teacher-linked effects as teacher effectiveness ratings implies that the teacher-linked effects are, at least to a large extent, *caused* by the teachers. For this to be the case, the model must be largely successful in separating effects due to the teacher from all the other factors, inside and outside of schools, that influence testing outcomes.

Failure to do this means that a teacher's estimated effect might also reflect bias due to factors external to the teacher, and out of his or her control. In their review of Buddin's 2010 analysis, Briggs and Domingue applied a method described and used by Rothstein[21] to test for the presence of such confounding factors. This test is applied here, and the results indicate the potential for statistically significant external effects.

The question of the extent of this bias, however, remains. Briggs and Domingue approached this problem by examining the sensitivity of the teacher-linked effects to alternative models that include additional variables measuring influences external to the teacher.

Buddin's updated 2011 research paper, and the *LA Times*' Teacher Rankings, emphasize their full model (model 4), which includes the largest set of student, classroom, and grade variables. Since this model, unlike its predecessor in the 2010 analysis, includes a full set of variables, it's difficult to test to whether the results are sensitive to alternative specifications (since there are few variables left to add). Instead, this review tests the sensitivity of the results not to changes in the model, but rather to changes in the results when teachers switch schools. The results suggest that teacher-linked effects do indeed vary by the school in which teachers work.

Buddin's Estimation of the Teacher-linked Effects

The teacher-linked effects in the 2011 white paper and the *Los Angeles Times*' Teacher Rankings, which this review replicates in order to assess, are produced from a model that essentially focuses on predicting students' test scores. The prediction of student j's test score in grade k in year n is based on

1. the student's test scores the previous year (k-1),
2. additional information about the student,
3. a contribution from the student's teacher in grade j, say teacher m, and
4. a contribution for the student's teacher in grade j for that specific year.

These quantities comprise an equation:

$$\begin{aligned}\mathbf{TestScore(j,k,m,n) = a*MathScore(j,k\text{-}1,n\text{-}1)+ b* ELAScore(j,k\text{-}1,n\text{-}1)} \\ \mathbf{+c*ClassSize(k,m,n)+d*TitleI(j,n)+\ldots} \\ \mathbf{+TeacherLinkedContribution(j,m)} \\ \mathbf{+AnnualTeacherLinkedContribution(j,m,n)+Residual(j,k,m,n)}\end{aligned}$$

The annual teacher contribution is the same for every student of that teacher in that particular year. The residual term is just the amount needed to make the left hand side equal to the right hand side. The values a, b, c, d, and so on are the weights given to each of the factors describing the student, the class and the grade. The weights are the same for all students.

The tests are the California Standards Tests (CST). The smallest model, Model 1, uses just the variables shown above. Model 2 adds a factor for the student's English language proficiency, a factor for the parents' education level, a factor for whether the student attended kindergarten in the LA Unified School District, and a factor for whether the student is new to the school. In addition to all of those variables, Model 3 adds class-level variables: the proportion of the class that falls into each English language proficiency category, the proportion of the class with each parental education level, the proportion that attended kindergarten in the LA Unified School District, and the proportion that of the class that is new to the school. To address the effects associated with ethnicity, Model 3 also uses the grade share of Hispanic students, African-American students, and students that are Asian or Pacific Islanders. Model 4 uses all the variables in Model 3, together with the class mean for the prior year's math CST scores and the class mean for the prior year's ELA CST scores.

The data used for the 2011 report includes math and ELA scores for grades 2-5 for the academic years ending in 2004 through 2010. Thus the current year n in the formula can range from 2005 to 2010 and the prior year ranges from 2004 to 2009. The grades range from 3 through 5, because prior year scores are not available for grade 2.

Fitting the models to the data means determining the weights a, b, c, d, . . . , etc. in the formula above, the teacher contributions, the annual teacher contributions, and the residual terms that are "best" according to the criteria of the model. In 2011, Buddin uses a linear mixed effects model. The impact of the choice is that the persistent teacher-linked effect (that across multiple years) and the annual teacher-linked effect (that in any given year) are reported by the model as normal probability distributions specified by mean and variance, rather than as single numbers. Additional technical description is available in the technical appendix to this report.

Data

Most of the data used in the 2011 publications were released to the *Los Angeles Times* by the LAUSD. The grade level shares of racial and ethnic groups are available from LAUSD's public online database. This study restricted attention to schools having at least 100 students, classrooms having at least 10 students, and students having complete test information for the models. For the 2005-2010 period, in Buddin's analysis, these restrictions result in about 11,500 teachers being represented in the model. For the 2004-2009 period, application of these restrictions results in about 11,300 teachers in the model.

The test scores from the data are standardized before being used in the model to have a mean of 0 and a standard deviation of 1 by grade and year. Among other things, this permits the comparison of scores between grades.

There was one important additional issue with the data that had to be addressed. Discussions with LAUSD personnel revealed that the data released for the academic year 2009-2010 did not include the English Language Development (ELD) levels. Though these levels are used in Models 2, 3 and 4, the absence of these data need not prevent fitting the models if the missing data can be effectively estimated (imputed) from the available data. Tests performed for this review showed that the imputation can be done quite effectively,[22] indicating that results in the 2011 publications may not have been substantially affected by the absence of the ELD levels for 2010.

However, due to the unavailability of the technique used by Buddin to compensate for the missing data, most of the model analysis here uses a full data set for 2003-2009. The one-year shift in time period should not affect the model properties. By this method, the strengths and limitations of Model 4 can be investigated without the complication of considering the validity of the imputation method.

Results

Much of the interest in value-added models stems from the often meaningful size of the variation in teacher-linked effects. If the effects can be interpreted causally, large variation in effects corresponds to large variation in the effectiveness of teachers in preparing students for the tests.

The models in Buddin's 2011 white paper find statistically significant variation among the teacher-linked effects and statistically significant year-to-year fluctuation in those estimates. The standard deviation estimated for the teacher-linked effects in ELA is between 0.18 and 0.16 standardized points (SD units), with the size decreasing as the model includes more information (variables). In math, the standard deviation estimated by the models for the teacher-linked effects is around 0.25 SD units. The replication of the models for the period 2004-2009 carried out for this review reproduces these values.

To put them in context, 0.16 standardized points on the ELA test is about 9 points on the 150-600 original scale of the ELA test. In math, 0.25 SD units correspond to about 20 points on the 150-600 scale. For comparison, the range designated "proficient" on the original scale is 300-349. In math, about 68% of the teacher-linked effects lie in the range corresponding to ±20 points, half the range of "proficient." In ELA, about 68% of the teacher-linked effects lie in the range corresponding to ±9 points, less than a fifth of the range of "proficient."

The year-to-year fluctuations estimated by Buddin are meaningful relative to the scale of the teacher-linked effects. The standard deviations for the classroom effect, the annual fluctuation, are 0.14 SD units and 0.18 SD units for ELA and math in model 4 in the original analysis, 90% and 71%, respectively, of the size of the actual teacher-linked effect standard deviation. This estimation of the magnitude of these annual fluctuations is an important contribution of the 2011 white paper. Other researchers have also found annual fluctuation in teacher-linked effects in addition to the sampling variability in the students.[23] Value-added models that do not take this annual fluctuation into account risk producing overly divergent teacher-linked effects with prediction intervals that are too small.

Stability

The teacher-linked effects are represented by a mean and a variance, or, equivalently, a standard deviation. These standard deviations-which comprise the error margins[24]-are similar in size to the effects themselves. This creates a fundamental incompatibility or tension between the rating categories used by the *Los Angeles Times*' and the actual model results.

Remember that the margin of error (in standard deviations) for a given teacher's estimate tells you the range (confidence interval) into which the "true effect" probably falls. The mean is in the middle. The *Times*' teacher rating categories completely ignore the error margins and sort teachers into groups based exclusively on their means, with the categories separated by the 20th, 40th, 60th and 80th percentiles.

Yet these percentiles are separated by amounts comparable to typical standard deviations for the individual teacher-linked effects. In other words, the error is almost as large as the distance between categories.

As a result, the majority of the persistent teacher-linked effects cannot be assigned reliably to a single category. In fact, for ELA, 85% overlap 3 or more categories. For math, 75% overlap 3 or more categories.

Buddin notes the unreliability of the means as point estimates of the teacher-linked effects:

> . . . teacher effectiveness varies substantially from year to year and ... the standard error of the model residuals are large even in the most complete model. These factors translate into imprecise point estimates for individual teachers.[25]

Figure 3 shows the proportion of math persistent teacher-linked effects that have 90% intervals entirely below 0, including 0, and entirely above 0 SD units, zero being the value corresponding to an average teacher-linked effect. When the range is entirely above or below zero, this means we can have confidence that the teacher is "truly" above or below

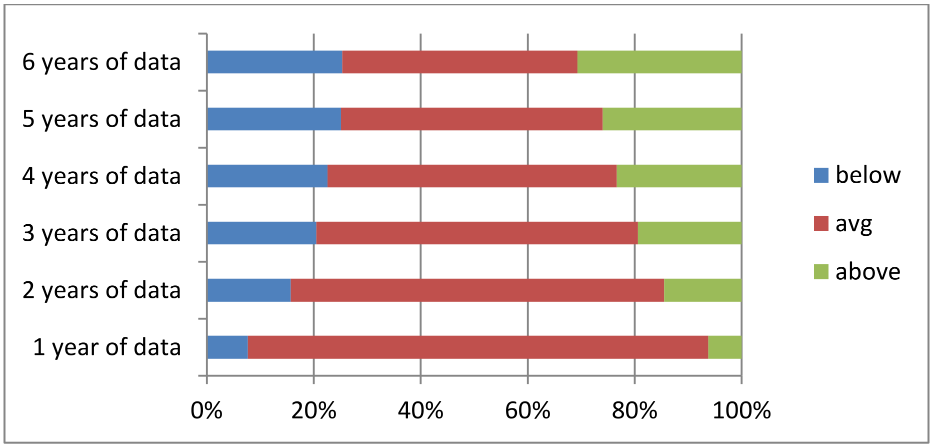

Figure 3. The proportion of teachers having 90% intervals lying below the average, including the average value, and lying entirely above the average value for math.

average, respectively. When the range includes zero, this indicates that the teacher is statistically indistinguishable from the average.

The proportions are calculated separately for effects based on 1, 2, 3, 4, 5, and 6 years of data. Most of the estimates based on fewer than 3-4 years of data are statistically no different from the average, to say nothing of whether we can reliably separate them into quintiles, as does the *Times*. The results for ELA are similar, but with a larger proportion of effects including 0 for all years, as seen in Figure 4. These results are qualitatively similar to those calculated by Di Carlo[26] for the New York data.

The *Times* basically ignores this imprecision, and takes the estimates at face value.

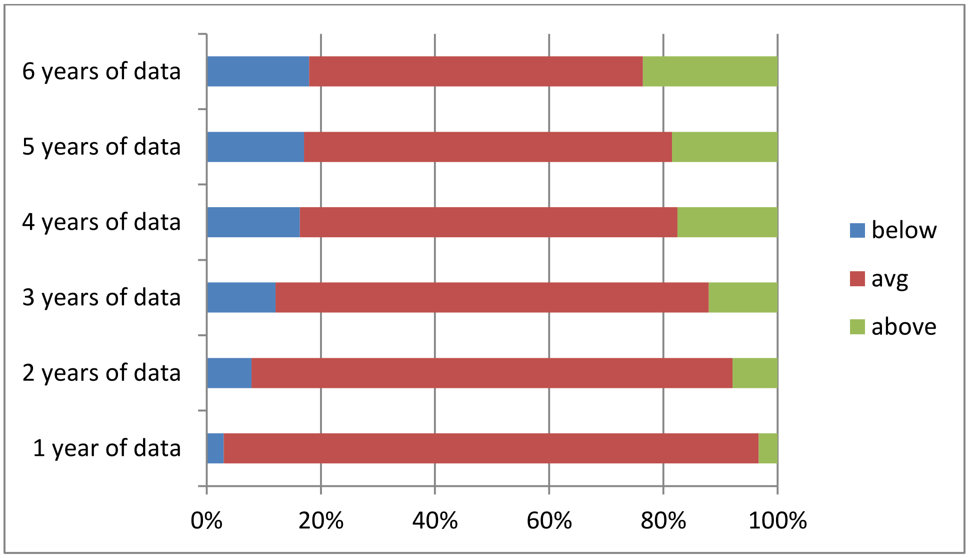

Figure 4. The proportion of teachers having 90% intervals lying below the average, including the average value, and lying entirely above the average value for ELA.

In fairness, the *Los Angeles Times*' Teacher Ratings do include information about the standard deviation of the teacher-linked effects by indicating the symmetric 90% probability interval for the effect, but the database presents far more prominently the mean and the categorical rating (e.g., "very effective"). A reader trying to assess a teacher's "performance," or to compare two teacher-linked effects, is most likely to use the means and especially categorical ratings, as understanding the 90% intervals is more complicated, and requires some statistical background.

Variability Over Time

Given that the majority of the teachers in the data are represented by data for three or fewer years, and 26% are represented by just one year, the question of the stability of effects calculated for shorter periods is relevant—estimates based on fewer years of data represent smaller samples, which yield less precise estimates.

This can be illustrated by modeling two separate persistent teacher-linked effects for teachers who are in fact present for the full six years. By splitting their records into two shorter periods, and estimating the persistent teacher-linked effects for the shorter periods separately within the 6-year model, one can assess directly the consistency of persistent effects estimated on the basis of shorter periods of data for the same teachers. The technical appendix provides details of the process.

The results show substantial differences. For math, for two three year periods, the median distance between the earlier and later values is about 0.11 SD units while for ELA it is about 0.07 SD units. For math, 20% of the changes are 0.20 SD units or more, regardless of whether the pairs of effects are based on 1, 2, or 3 years of data in each period. For ELA, this value is about 0.14 SD units. For reference, the widths of the intervals between the 20th, 40th, 60th and 80th percentiles range from 0.06 to 0.14 SD units.

Figure 5 presents these changes in terms of the effectiveness categories used by the *Times*. They indicate the change between the first and the second rating for teachers rated in two separate 3-year periods. For both math and ELA, the majority of teachers change categories. For ELA, 24% change by 2 or more categories. For math, 17% change by 2 or more categories.

A common argument is that value-added models are better equipped to identify the very highest- and lowest-performing teachers, and so it also makes sense to consider more

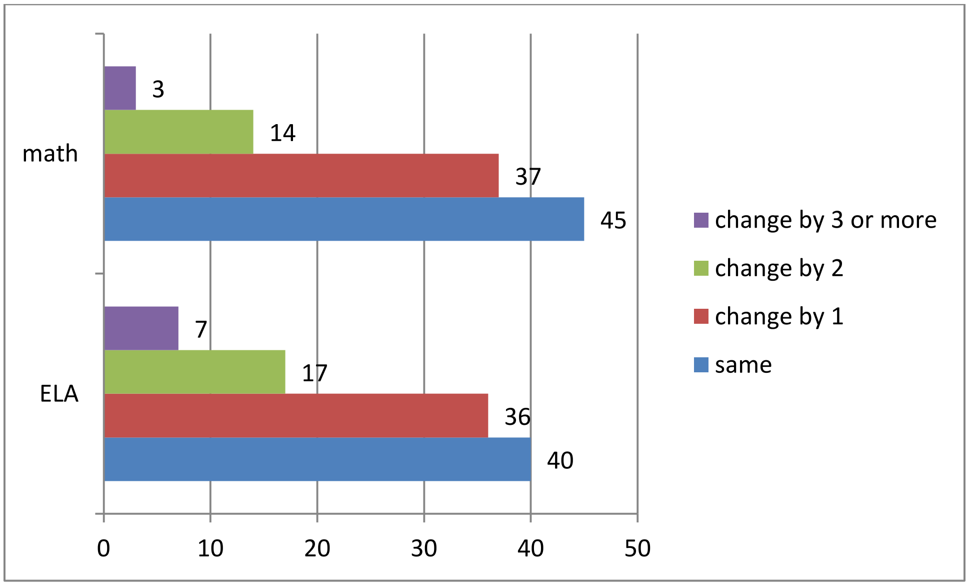

Figure 5. Percentages of rating changes for teacher-linked effects estimated for the same teacher in two separate three-year periods.

extreme teacher-linked effects. In math, 48% of the teachers with effects in the top 5% overall in the first three-year period are in the top 5% in second, while 54% of the teachers with effects in the lowest 5% over all in the first three year period are in the bottom 5% in the second three-year period. These percentages are 47% and 41% for reading.

Overall, then, the means of the teacher-linked effects for many individual teachers differ substantially between the two periods. However, when the variances of the effects are taken into account, the differences between the effects estimated from the first period of data and the effects estimated from the second data period collectively are consistent with the model assumption that the persistent effect does not change over time, except in one case. For math effects estimated by three years of data, the effects for the first three years and for the second three years for each teacher are not consistent with the model assumption that the persistent teacher-linked effect is constant over time.[27] This raises the question of whether the assumption of a constant persistent teacher-linked effect is reasonable over extended periods: though the effects for one and two years are consistent, the teacher-linked effects for three years in math, the case with the most precise effect estimates, show significant differences between the two periods. Of note, for ELA, the effects for single years are consistent with all the effects being 0 ($p>.5$). This is also true for the Model 4 ELA results for teachers actually represented by one year in the study (see Technical Appendix).

Variability in Prediction

Another way to assess the stability of the teacher-linked effects is to carry out prediction exercises—in a sense, roughly simulating what would happen if we used the estimates to try to predict the future. The availability of seven years of data enables examination of the predictive power of the teacher-linked effect based on six years of data for the test results in a new year (the seventh). This predictive power is expected to be low, due to the large size of classroom (single year) effects (e.g., peer effects not controlled for) relative to the *persistent* teacher-linked effects. (Recall that the standard deviation of the classroom effect for math is 71% of the size of the standard deviation of the persistent teacher-linked effects, while the corresponding percentage for ELA is 90%.) In fact, for math, the correlation of the persistent teacher-linked effect calculated from 2004-2009 data with the teacher-linked effect estimated from 2010 data (without an annual fluctuation) is 0.60.

To make a more direct comparison to the 2010 data we can calculate the value added in 2010 as the classroom average of the difference between the students' scores in 2010 and the students' predicted scores based on the previous year's scores and the demographic information using the weights estimated for Model 4 (which includes prior years). Since the value added for 2010 is calculated using a different method than the persistent teacher-linked effect, the effectiveness ratings for 2010 are assigned based on the 20th, 40th, 60th and 80th percentiles for the 2010 value-added results.[28] Figure 6 presents the changes in effectiveness categories between the teacher-linked effects estimated from 2004-2009 data and the value-added results for 2010.

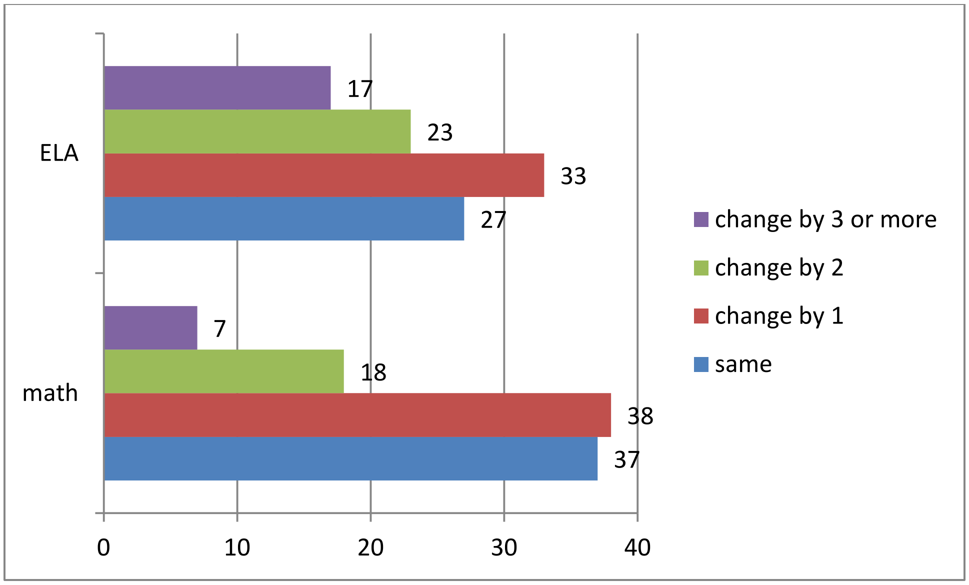

Figure 6. Percentages of rating changes between teacher-linked effects estimated for 2004-2009 and value added in 2010.

There is more consistency between the persistent teacher-linked effects and the classroom value-added effects for the extreme ratings than for the middle three ratings. For math, 53% of those in the lowest category in teacher-linked effects have classroom effects in the lowest category. In ELA, this is 43%. For math, 55% of those in the highest category in teacher-linked effects have classroom effects in the highest category. In ELA, this is 48%.

This weak predictive power in math does not appear to be due to the missing ELD data in 2010 (discussed above). There is little qualitative change in predicting test results in 2009 from 2004-2008 persistent effects. The results are not qualitatively different for persistent teacher-linked effects based on 2 through 5 years of data. The predictions from persistent teacher-linked effects based on one year are somewhat poorer. The ELA results are poorer for the prediction of 2010 from 2004-2009, with a correlation of .32, than for the prediction of 2009 from 2004-2008, with a correlation of .44, possibly due to the imputation of the ELD data, and the larger weights given to the ELD categories for ELA.

Even when multiple years of data are employed, the lack of predictive power limits the usefulness of the estimates to a parent considering whether an assigned class will improve a child's test-based performance.

Bias

The goal for much of the effort in constructing value-added models is the estimation of teacher-linked effects that can reasonably be thought to be *caused by the teachers*. Even if the estimates are stable and precise, they might still reflect the influence of factors other than teachers.

The problem of isolating the causal effect of the teacher on a student from that of the student's peers, the school, and influences outside school is profound. Any systematic differences among student assignments that influence CST scores and are not captured by the fixed effects will be absorbed by the teacher-linked effect, though these differences are not caused by the teacher. This includes efforts of other teachers in team taught classrooms, as well as efforts of specialists or other factors.

Examples of systematic differences in student assignments, within and between schools, are easy to find. Schools draw from different geographic areas, and so may have student populations that differ systematically. Results of No Child Left Behind announcements may influence some parents' choice of school. Teachers within a school may be more effective instructing students with certain characteristics. For example, in larger schools, students may be grouped by their educational needs. Differences in educational resources among schools may also have effects that are then attributed to the teachers.

The importance of bias differs depending on the source of the bias and the application of the information. For teachers and administrators, bias due to variation among schools raises issues of fairness and disincentives for teachers to work in schools with conditions that bias teacher-linked effects downward (or are perceived to do so). Bias due to variations in resources among the schools may not matter much to parents who use the ratings, provided the teacher remains in one school: the school's resources will be available to children in that teacher's class. Bias due to systematic differences in the assignment of children to teachers reduces the utility of the teacher-linked effect to parents. In other words, to whatever degree the ratings reflect non-teacher factors, parents choosing teachers based on those ratings may have bad information, even if they interpret the estimates properly (which is no guarantee). Moreover, if the teacher's typical student population changes in response to the rating, then the bias, and also the rating, may change.[29]

Potential for bias

Jesse Rothstein[30] proposed a test for value-added models that can show that a model is vulnerable to systematic differences not captured by the fixed effects. The intuition behind the method is that if something that is known not to have an effect, in the causal sense, has a significant effect in the model, then the model cannot be assumed to be unbiased.

Specifically, Rothstein's falsification test asks whether a student's future teachers influence their current performance. This is, of course, impossible, as there is no causal effect of future teachers on current performance. Thus, if these effects are statistically significant, then students are being sorted into classes, intentionally or unintentionally, by the extent to which their scores the previous year exceeded the scores predicted on the basis of the other variables in the model. That is, student assignment is not random, even after controlling for the other variables in the model, and it is non-random in a way that contributed to teacher-linked effect the previous year.

To estimate these counterfactual effects of the future on the present, one produces an altered data set in which the teacher assignments for a given year are replaced with the assignments for the following year. For example, a student's data for the fourth grade, including the prior test scores from third grade, is modeled as a function of the student's fifth-grade teacher. When the model is fitted, it calculates effects of the teachers on their students' scores in the year before the students were in the teachers' classes.

For all eight models studied, this test produces statistically significant results ($p<.001$). The additional precaution of controlling for the student's actual teacher assignment in the test year prevents the future teacher's identity from being useful by contributing information about the identity of the actual teacher. For reasons of computational intensiveness, these models were fit without annual fluctuation terms. In all eight models, with controls for the actual teacher and with all the additional student data used in the model, the next year's teacher is a significant predictor of the test score in the current year ($p<.001$). For model 4, the model standard deviation of the future teacher-linked effect for math, controlling for the current teacher, is 0.09 SD units (though the scale is affected by the lack of annual effects). For ELA, the corresponding teacher-linked effect is 0.08 SD units

This result suggests the presence of bias. If the amount by which a student's score exceeds the score predicted on the basis of lagged scores and demographic variables in one year is correlated with the amount by which the student's score is expected to exceed the predicted score the following year, the non-random assignment produces bias in the teacher-linked effects.

Bias Estimate

A recent working paper by Chetty, Friedman and Rockoff[31] pointed out, in the context of another study, that, while this test shows the potential for bias, it does not indicate the size. Chetty et al demonstrate that teachers with effects in the 95th percentile in that study essentially carry their effects with them when they move schools, moving the new school's grade level mean scores up by an amount consistent with the teacher-linked effect and the proportion of students taught by the teacher. This is presented as quasi-experimental evidence that the effect is not due to the non-random assignment discussed above.

To examine whether the effects for a teacher change if the teacher changes schools in the context of the LAUSD data, one can compare a teacher's estimated effectiveness at a first school to that same teacher's effectiveness at a second school. The data include about 800 teachers who taught at two different schools during the study period. To achieve the goal of estimating the effects for these teachers at the two schools separately without changing the model, these teachers are divided into five groups of about 160. For each group, the entire 6 year model is re-estimated, but with the teacher-linked effects for the teachers in the group estimated separately at the two schools (essentially, each teacher is modeled as

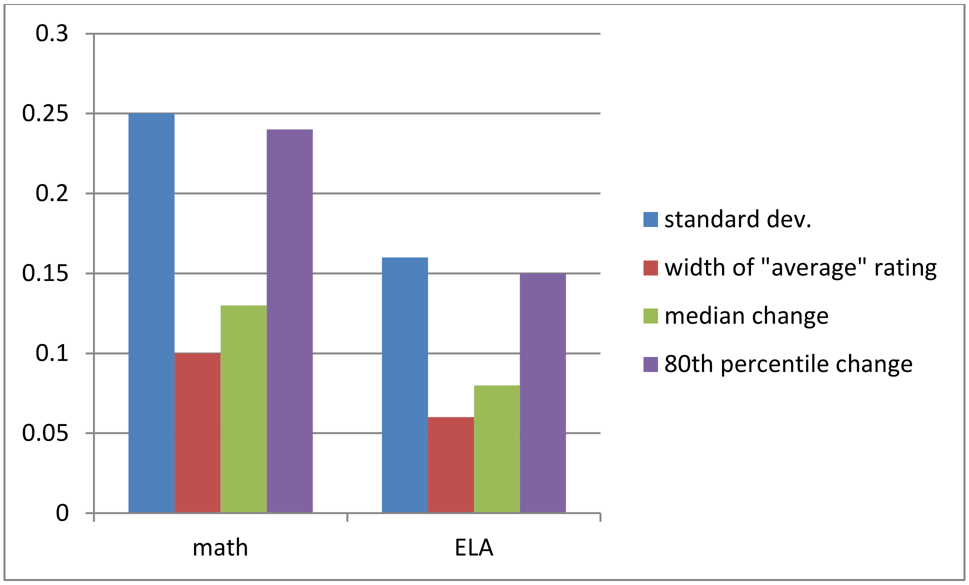

Figure 7. The size of the difference in teacher-linked effect between schools for teachers teaching at exactly two schools.

two separate teachers). This preserves the other teacher-linked effects virtually unchanged. The paired teacher-linked effects for the teachers in each group are collected from these models.

The median and the 80th percentile of the disparity in the estimates at the different schools are shown in Figure 7. For context, the standard deviation of the teacher-linked

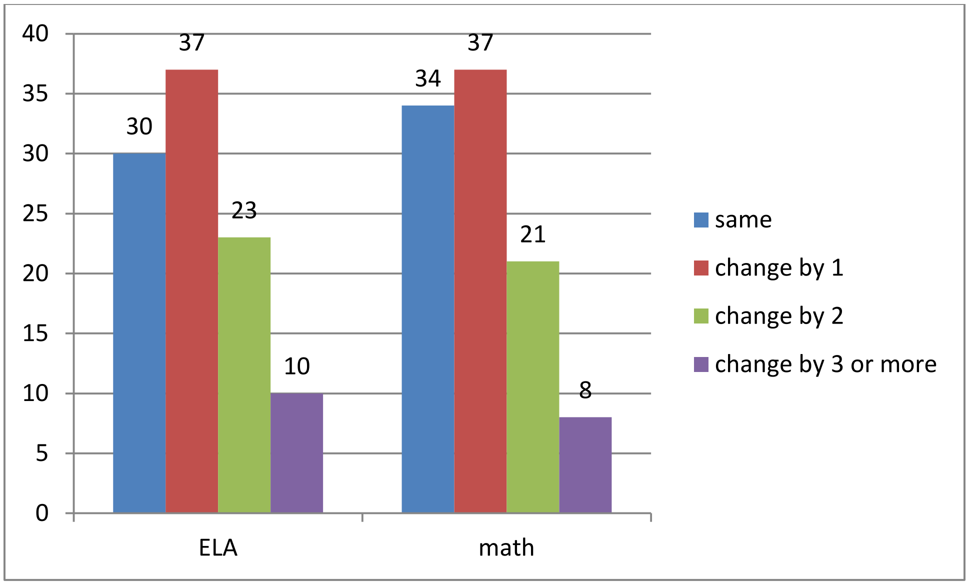

Figure 8. The percentages of teachers changing categories when changing schools.

effect and the width of the range of the “average” category are also shown. For both ELA and math, the median change is larger than the “span” of the “average” category, and the 80th percentile of the change is close to the standard deviation for the subject.

Changes in teacher-linked effect of this magnitude affect the ratings substantially. Figure 8 shows the proportions of teachers having the same rating at both schools, changing by one category, changing by two categories, and changing by three or four categories.

While these changes are substantial, they may be partially due to the changes observed earlier—those occurring when teacher-linked effects for the same teacher are estimated using data from separate time periods. Put differently, when teachers change schools, their estimated effects may change for the same reasons they do for non-switchers—imprecision.

If, on the other hand, the effects at the two schools—for teachers changing schools—vary systematically depending on properties of the schools, this is evidence of bias in the teacher-linked effect produced by differing characteristics of the schools. For example, if the teacher effect estimate is higher at the school with the higher teacher-linked effects overall, and lower at the school with generally lower teacher effects, there is an indication that some feature of the schools, such as student demographics not addressed in the model, or availability of additional instructional resources, biases teacher-linked effects.

To investigate this, for each teacher changing schools, the mean teacher-linked effect for the teacher’s first school and second school is calculated, excluding the teacher. For each teacher, the change in the teacher’s teacher-linked effect between the first and second school is calculated, and the models check whether there is an association between this difference and the mean of the teacher-linked effects at the first school and at the second school. In other words, might teachers who switch schools get different estimates because of something about the schools to (and from) which they move? If so, then perhaps the influence of the school is biasing *all* teachers’ estimates, but such bias is not detected because they remain in the same place.

A regression of the change in teacher-linked effect on the two school means for math results in significant ($p<.001$) coefficients for the first and second school means. This allows one to make a rough estimate, by multiplying the minimum and maximum values of the school means of teacher-linked effects by the regression coefficients, that the teacher-linked effects for teachers changing schools include contributions from the schools in the range (-0.08 to 0.09) SD units This range does not include potential bias from systematically differing assignment of students to teachers within schools, and does not address the range for other teachers. For ELA, the coefficients are less significant ($p<.05$). The corresponding range is (-0.04, 0.04) SD units. It should be noted that the contribution of the school means, though statistically significant, particularly for math, and on a meaningful scale, leaves much of the change of effect unexplained ($R^2=.03$ for math, $R^2=.02$ for ELA). This is not surprising, since, as shown above, substantial changes in teacher-linked effect in separate time periods are also seen for teachers who do not change schools. The purpose of this analysis is not to explain all of the change, but to show and quantify an association between the change and the teacher-linked effect means at the two schools.

Sensitivity to Model Changes

Buddin and the *Los Angeles Times*' Teacher Ratings make an effort to address the issue of bias by publishing the results of four models with increasing sets of variables for the student, classroom, and grade. Readers may view the teacher-linked effect means for each teacher for each of the four models. The sensitivity of the results to the choice of model is modest though noticeable. For ELA, the ratings for 4% of the teachers change by 2 or more categories going from model 1 to model 4. For math, 0.4% of the teachers, 50 teachers, have rating differing by 2 or more categories.

Other alternative models based on these data produce noticeably but not dramatically different results from Buddin's model 4. For example, restricting to grades four and five and using two years of prior test scores in Model 4 rather than one changes 20% of the ELA teacher-linked effects by 0.03 SD units or more, with a maximum of 0.15 SD units, and changes 20% of the math effects by 0.04 SD units or more, with a maximum change of 0.22 SD units. Though noticeable, most of these differences are small relative to the variability, and do not produce statistically significant changes.

Large sensitivity can be taken as an indication that bias is a threat, but modest sensitivity cannot be taken as evidence of lack of bias. Important information may be missing from the data. For example, income and race/ethnicity are represented in the data only by the rough indicators of Title I status, and grade-level shares of racial and ethnic groups, rather than individual racial or ethnic identity.

Correlations between the teacher-linked effects and the student data may also introduce bias. There are modest but significant correlations between the estimated teacher-linked effects and the student data. For example, correlation between the teacher-linked effect for math and the class mean of the prior math score is .12 ($p<.001$), while for ELA, the correlation between the teacher-linked effect and the prior class mean score is .11 ($p<.001$). The model will not, generally, correctly apportion weights among correlated variables.

Discussion

Many concerns about the high stakes use of teacher-linked effects from VAM go beyond issues of variability and bias. A quick review of some of these concerns is presented here. Generally, these concerns were originally raised in response to potential uses of value-added effects in teacher evaluation and in personnel decisions such as hiring and compensation. These concerns should be revisited following publications of teacher—linked effects. One reason for review is that public access to teacher-linked effects may create public pressure on teachers to improve ratings or public pressure on administrators to act on ratings. A second reason for review is to examine how these concerns relate to the ways that the reading public, parents in particular, may use the teacher-linked effects. The final reason for the review is to avoid creating the impression, by focusing on technical issues, that these other concerns have been resolved.

In the context of the use of value-added effects in teacher evaluation, one shortcoming of teacher-linked effects is that the effects do not provide any guidance regarding how the current results were achieved or what a teacher could do to improve.[32] While this lack is of interest to parents and other readers of the white paper and the *LA Times* coverage, it may not be as salient among readers who use the Teacher Ratings. It does seem to be behind some of the frustration felt by teachers reading the ratings, who do not find that the ratings accord with their experience.

Another concern is that the material covered by the tests used in the VAM is only a part of the teachers' instructional responsibility. The use of high stakes tests in teacher evaluation

> *The seemingly simple and precise effectiveness categories used in the Teacher Ratings are potentially misleading.*

presents the risk that teachers may focus on teaching students to do well on the tests, to the detriment of other subject matter. This other subject matter may also be topics of greater depth in the tested subjects that are not addressed by the test. There are indications that scores on some of the standardized tests generated under No Child Left Behind have relatively low correlations to the scores on tests designed to assess higher-order conceptual understanding.[33]

Entire subjects such as social studies and science are often not addressed by the standardized tests used in value-added models. Values for the tested subjects cannot be assumed to carry over to the untested subjects. Even values for the two tested subjects can be quite different for the same teacher. In Model 4 for LAUSD, the correlation of teacher-linked effects for math and ELA is .76, low enough that 16% of the teachers have differences of 2 or more categories between the subjects. For 27 teachers the ratings are at the opposite extremes.

In teacher evaluation, these limitations mean that value-added effects do not address many of the responsibilities that fall to teachers. Parents trying to use the teacher-linked effects to choose a teacher are receiving imprecise information about a limited range of the teachers' contributions to student learning.

A teacher's long-term effect on a student's test scores in subsequent years has been seen to differ from the effect in the year in which the student had the teacher.[34] In the LAUSD data, there is evidence that the longer-term teacher-linked effect is only loosely correlated with the effect in the first year, and that the size of the effect is greatly reduced after one year. In the contexts of teacher evaluation, personnel decisions, and parental choice of teachers, emphasis on the short-term gain estimated by the teacher-linked effect may detract from the long-term goal of a well-educated child.

The information accompanying the *LA Times*' Teacher Rankings notes that these rankings are relative to other teachers in the data. High or low ratings do not tell the reader how the teacher compares with other teachers nationally or to standards of professionalism. Some question the usefulness of competitive rankings for teachers in light of evidence that teams of teachers working collaboratively are effective in improving student learning.[35] The immediate consequence of the relative nature of the scores is that readers are left wondering whether a higher teacher-linked effect is an indication of excellence in the classroom or a minor variation between teachers. For a parent, the question is whether the higher effect is worth moving the child to another school, with the attendant disruption, or not. The longer-term concern for readers is that emphasis on these competitive measures may detract from useful reform rather than promoting it.

That teachers with dissimilar student populations are rated on a single scale presents a challenge to the rating system. Certainly the typical classrooms for teachers in this study vary widely.[36] The skills needed from teachers in very different classrooms may be very different. Depending on the scales of the tests, the effort that goes into an effect of .2, say, in different classrooms may be very different. For reform programs, inequities of effort for similar evaluations may discourage teachers from teaching in some types of classroom. For parents, the compression of many circumstances into one rating system complicates identification of highly rated teachers successful with children like their own.

These limitations on the use of teacher-linked effects, together with the technical limitations due to large variances, large annual fluctuations, and uncertain causal relation between the effect and the teacher, reduce the utility of individual teacher-linked effects to parents and general readers.

Recommendations

The 2011 white paper and the 2011 *LA Times*' Teacher Ratings do provide readers with some information about the variability of teacher-linked effects and the sensitivity of results to model specification. This is an improvement in the rating over the 2010 version. However, the inclusion of teachers for whom very little data is available exacerbates the problem of large variances. These large variances of the effects and the problem of separating teacher-linked effects from effects associated with other non-teacher factors complicate the extraction of valid information from the teacher-linked effects, especially given the manner in which the *Times* presents the data. The seemingly simple and precise effectiveness categories used in the Teacher Ratings are potentially misleading.

- Teachers in the studied data have incoming classes with very different characteristics. A high teacher-linked effect must be understood as an estimate of test score improvement applying only to the range of students typically taught by that teacher.

- The model results indicate that teacher-linked effects estimates must take annual fluctuation beyond sampling error into account.

- The teacher-linked effect means and the effectiveness categories are not reliable for comparison or prediction. The large variability must be taken into account. This limits the detail available in comparisons among teacher-linked effects, particularly for effects estimated from three or fewer years of data.

- Large annual fluctuations make even effects calculated over longer periods very approximate predictors of the effect size in the future. Parents should not rely on the published effect being reproduced in any given year.

- There is strong evidence that the teacher-linked effects include contributions to student learning not due to the teachers. These contributions are meaningful on the scale of the categories used in the *Los Angeles Times* effect report. Comparison of teacher-linked effects must be understood as comparing teachers and their work environments, not just teachers.

Technical Appendix

The Models

The model form stated in the paper to be the basis for analysis is given here for reference. Explanation will follow.

$$T_{ijt} = T_{ijt-1}\lambda + x_{ijt}\beta + \varphi_j + \varepsilon_{ijt}$$

The results tabulated in the paper are for this model:

$$T_{ijt} = T_{ijt-1}\lambda + x_{ijt}\beta + \varphi_j + \psi_{jt} + \varepsilon_{ijt}$$

The latter includes the term ψ_{jt} to estimate annual fluctuation. The exposition in the FAQ sheet for the *LA Times* ratings publication does not specify which model was used to produce the published teacher-linked effects, though given the large annual fluctuations found in the 2011 white paper, use of a model without the ψ_{jt} term would result in overstated differences in effects and underestimated variances. The exposition in the review is based on the latter model. The models are fitted as linear mixed effects models in the original and in this review.

The definitions of the variables are the same in both models. Here, T_{ijt} is the test score for student i in year t, with the j indicating the student's teacher in year t. The term T_{ijt-1} is the test score or scores for student i in year t-1, that is, the student's scores the previous year, with the j again indicating the student's teacher in year t. The x_{ijt} term holds known characteristics of student i in year t, with teacher j. The λ and the β are weights for the test scores and student characteristics. Collectively, T_{ijt-1} and x_{ijt} are the fixed effects.

The remaining terms adjust for the difference between the actual score in year t and the score predicted on the basis of T_{ijt-1} and x_{ijt}. The term φ_j, the so-called teacher effect, or persistent teacher-linked effect, is the same for all students with teacher j. The term ψ_{jt}, the teacher-year, or classroom effect, is applied to all students with teacher j in year t. The final term, ε_{ijt}, varies by student and year. In the fitted models discussed in the white paper, these three values are viewed as random effects. They are thought of as random draws from three normal distributions with zero average and variances that must be estimated from the data along with the values of λ and β. The estimated values of the variances together with the data determine how the difference between T_{ijt} and $T_{ijt-1}\lambda + x_{ijt}\beta$ will be apportioned among φ_j, ψ_{jt}, and ε_{ijt}, that is, among the teacher, the classroom, and the student in that year.

This approach to estimating adjustment terms associated with different levels of the phenomenon under study is fairly standard, and is discussed in Raudenbush and Bryk[37] and McCulloch and Searle.[38]

The models in the review are fitted using the lmer function in the lme4 package in R.

Effect Comparison for Different Periods

The effects for the same teacher in two different periods were obtained from the records of teachers present in the same school for all six years. For each number of years, y= 1,2,3, the last 2y years for these teachers were used. The first y of these became the first estimation period, and the last y made up the second estimation period. These teachers were split into groups of about 250. For each group, the teacher ids for the first and second periods were tagged with the period, separating each teacher in the group into a pair of teachers for the purposes of the model estimation. The remaining data for the teachers in the group was discarded. Then Model 4 was run on the paired records and the data for the other teachers. The teacher-linked effects for the paired teachers for each group were collected and used to compare the persistent teacher-linked effects estimated for the same teacher from non-overlapping data.

Due to the limitation of the group size to 250, the teacher-linked effects for teachers not in the group are virtually unchanged between the original and the modified model.

Significance Calculations

The consistency of two sets of teacher-linked effects for the same teacher was calculated from the effect distributions, the model assumption that the “true” persistent effects are draws from the effect distributions, and the model assumption that the two persistent effects are equal.

For a single pair of effect distributions, say a Gaussian distribution with mean μ_1 and variance σ_1^2 , $N\ \mu_1, \sigma_1^2$ and another, say $N\ \mu_2, \sigma_2^2$, the probability of obtaining 0, or a value further from the $\mu_1 - \mu_2$ than 0 for the difference between a draw from $N\ \mu_1, \sigma_1^2$ and a draw from $N\ \mu_2, \sigma_2^2$, can be calculated from $N\ \mu_1 - \mu_2, \sigma_1^2 - \sigma_2^2$. The calculation is equivalent to calculating the probability of obtaining a value as far from 0 as $\mu_1 - \mu_2$ given a normal distribution $N\ 0, \sigma_1^2 - \sigma_2^2$. This, in turn, can be calculated based on a χ^2 distribution with one degree of freedom and the statistic $\frac{\mu_1-\mu_2}{\sigma_1^2-\sigma_2^2}^2$.

Summing all such statistics for the set of pairs of effects and using the χ^2 distribution with corresponding degrees of freedom gives a probability calculation that accounts for multiple testing.

For the case of math effects calculated from two different three year time periods, the collective probability of obtaining $\mathbf{0}$ or a value further from the vector of effect differences than $\mathbf{0}$ was less than .001. This shows that the pairs of effects estimated make the assumption of equal persistent effects for each teacher for the two time periods unlikely to be true.

For the effects in ELA estimated from a single year, the probability of obtaining all **0**'s or values as far from the vector of effect means as **0** was greater than .5. This indicates that the effects estimated are consistent with all the effects being 0.

The significance of additional random effects is based on the conservative χ^2 test on the difference of the deviances described in Bates.[39]

Prediction Calculations

For the purpose of testing the predictive power of the teacher-linked effects, the 2010 effects were estimated in two ways. For the executive summary, the 2010 effects were estimated using model 4, restricting to 2010 and omitting the annual effect. This method was used for simplicity of exposition. The effect estimate for 2010 for the main part of the review is described in the review. The results are similar. Effectiveness categories for 2010 are based on the 20th, 40th,60th and 80th percentiles of the 2010 estimates to compensate for the different scales for the 6-year estimates and the 1 year estimates.

Issues with Regression of Grade Effects on Teacher-linked Effects

Regression of grade-wise demographically adjusted gains on grade-wise mean teacher-linked effects is not sensitive to bias in the teacher-linked effects under circumstances that may be common in the LAUSD data. It will not detect that some portion of a departing teacher's teacher-linked effect was produced by differential assignment of students to the teachers if the remaining staff has the same assignment pattern before and after the staff change. Further, if a staff change is due to a teacher leaving the system, and that teacher's effect was calculated entirely on the basis of classes taught in that school, the departure will not capture the portion of the departing teacher's teacher-linked effect due to differential assignment of students to schools. In data set for 2004-2009, over 70% of the teachers are not present in the data set for all 6 years but do not change schools, while under 8% in fact change schools within the data set.

Effects Over Time

The student gain over two grades was modeled using the teacher ids for each grade, and the lagged scores and mean lagged scores, student gender, ELD level, parental education, the Title I indicator, and the indicator for kindergarten in LAUSD for the beginning of the two year period.

The result for ELA was that the standard deviation assigned by the model to the second year teacher was 0.19, while the standard deviation assigned to the earlier teacher was 0.10. The correlation of the teacher-linked effects for the teachers in the first year position with the teacher-linked effects for the same teachers in model 4 is .45. The correlation of the teacher-linked effects for the teachers in the second year position with the teacher-linked effects for the same teachers in model 4 is .77.

For math, the standard deviation assigned by the model to the second year teacher was 0.30, while the standard deviation assigned to the earlier teacher was 0.12. The correlation of the teacher-linked effects for the teachers in the first year position with the teacher-linked effects for the same teachers in model 4 is .47. The correlation of the teacher-linked effects for the teachers in the second year position with the teacher-linked effects for the same teachers in model 4 is .87.

This suggests that a more thorough model for the teacher-linked effects of previous years' teachers on the students' test scores would show that the effect, eff_{old}, associated with having had a Teacher Smith in, say, grade 3, on the student's performance in grade 4, is different from the effect associated with Teacher Smith on the student's performance in grade 3, $eff_{current}$. In general, eff_{old} may be smaller in magnitude than $eff_{current}$. Further, the ordering of teachers by eff_{old} values may differ substantially from the ordering of teachers by $eff_{current}$ values.

Notes and References

1 Los Angeles Teacher Ratings (n.d.). *Los Angeles Times*, retrieved April 10, 2012 from http://projects.latimes.com/value-added/faq/.

2 Briggs, D. C. & 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. Boulder, CO: National Education Policy Center. Retrieved June 11, 2012, from http://nepc.colorado.edu/publication/due-diligence.

3 This paper can be found at http://documents.latimes.com/buddin-white-paper-20100908/.

4 This paper can be found at http://documents.latimes.com/buddin-white-paper-20110507.

5 Wainer, H. (2011). Value-added models to evaluate teachers: A cry for help, *Chance, 24*(1). Retrieved April 10, 2012, from http://chance.amstat.org/2011/02/value-added-models/.

6 Rothstein,R., Ladd, H.F., Ravitch,D., Baker,E.L., Barton,P.E., Darling-Hammond, L.,Haertel,E., Linn, R.L., Shavelson, R.J. & Shepard , L.A (2010). *Problems with the use of student test scores to evaluate teachers*, Briefing Paper #278. Washington, DC: Economic Policy Institute.

7 Los Angeles Teacher Ratings (n.d.) *Los Angeles Times*, retrieved April 10, 2012 from http://projects.latimes.com/value-added/faq/.

8 Technically, due to the use of random teacher effects, the ranges are determined by the distribution of the random teacher effect conditional on the data for that teacher. Interpretation as “error margin” is reasonable for gaining intuition.

9 For the original article, see Felch, J., Song, J., & Smith, D. (2010, August 14). Who’s teaching L.A.’s kids? *Los Angeles Times*. Retrieved April 10, 2012, from http://www.latimes.com/news/local/la-me-teachers-value-20100815,0,2695044.story.

10 Santos, F. & Otterman, S. (2012). City teacher data reports are released. *The New York Times/SchoolBook* Retrieved April 10, 2012, from http://www.nytimes.com/schoolbook/2012/02/24/teacher-data-reports-are-released.

11 NYC Public School Teacher Evaluations (2012, February 28). *New York Post*. Retrieved April 10, 2012, from http://www.nypost.com/p/news/local/ratings_of_public_school_teachers_agH4xQDbuen0uBUpEoSuDM.

12 Gates, B. (2012, February 22). Shame is not the solution. *The New York Times*, A27. Retrieved April 10, 2012, from http://www.nytimes.com/2012/02/23/opinion/for-teachers-shame-is-no-solution.html.

13 See-through Ratings (2012, February 25) *New York Post*. Retrieved April 10, 2012, from http://www.nypost.com/p/news/opinion/editorials/see_through_ratings_N1ankCsHn7B9usMwujsCzH.

14 Hinchey, P. H. (2010). *Getting Teacher Assessment Right: What Policymakers Can Learn from Research*. Boulder, CO: National Education Policy Center. Retrieved April 10, 2012, from http://nepc.colorado.edu/publication/getting-teacher-assessment-right.

15 Rothstein,R., Ladd, H.F., Ravitch,D., Baker,E.L., Barton,P.E., Darling-Hammond, L.,Haertel,E., Linn, R.L., Shavelson, R.J. & Shepard , L.A (2010). *Problems with the use of student test scores to evaluate teachers*, Briefing Paper #278. Washington, DC: Economic Policy Institute.

16 Rothstein,R., Ladd, H.F., Ravitch,D., Baker,E.L., Barton,P.E., Darling-Hammond, L.,Haertel,E., Linn, R.L., Shavelson, R.J. & Shepard , L.A (2010). *Problems with the use of student test scores to evaluate teachers*, Briefing Paper #278. Washington, DC: Economic Policy Institute.

17 Rothstein,R., Ladd, H.F., Ravitch,D., Baker,E.L., Barton,P.E., Darling-Hammond, L.,Haertel,E., Linn, R.L., Shavelson, R.J. & Shepard , L.A (2010). *Problems with the use of student test scores to evaluate teachers*, Briefing Paper #278. Washington, DC: Economic Policy Institute.

18 These papers can be found at
http://documents.latimes.com/buddin-white-paper-20100908/
and at http://documents.latimes.com/buddin-white-paper-20110507.

19 Briggs, D. C. & 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. Boulder, CO: National Education Policy Center. Retrieved June 11, 2012, from http://nepc.colorado.edu/publication/due-diligence.

20 Los Angeles Teacher Ratings (n.d.) *Los Angeles Times*, retrieved April 10, 2012 from http://projects.latimes.com/value-added/faq/.

21 Rothstein, J. (2011). *Review of "Learning About Teaching: Initial Findings from the Measures of Effective Teaching Project."* Boulder, CO: National Education Policy Center. Retrieved April 11, 2012, from http://nepc.colorado.edu/thinktank/review-learning-about-teaching.

22 Attempts to learn from Buddin and the *LA Times* about the treatment of these missing data were unsuccessful. The imputation of the missing data in this review was made by predicting students' next year's ELD level from the current year's data using a basic logistic regression model. The classification model was developed on the complete data for 2003-2009. Applying the classification method to the data for 2009 produced levels for use with the 2010 data. The resulting model fits are consistent with the results reported by Buddin for the period 2005-2010. Using the estimation technique to replace the ELD levels for 2009 with imputed levels, fitting the models for 2004-2009 on the modified data, then comparing with fits to the actual data for 2004-2009 provides a check on the performance of the imputation method.

The performance is, in general, very good. For math in model 4, there is almost a perfect correlation between the mean teacher-linked effect produced from the actual data and the effects produced from the imputed data. All the imputed means fall within the 90% interval for the means from the actual data. All the categories of the means for the imputed data are within one category of the means from the actual data, and 99% are in the same category. This leaves 131 teachers changing category between the two models. In model 4 for ELA, the comparison gives similar results, though here 279 teachers change category, and a couple of individual teacher means change more than one category. (This does raise the question of what level of variability in assigning value categories to individuals is acceptable.)

23 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.

24 Due to the use of random effect for the teachers, this is not technically correct, but provides a good intuitive guide.

25 Buddin, R. (2011, May). *Measuring teacher and school effectiveness at improving student achievement in Los Angeles elementary schools*, 24. Retrieved June 11, 2012, from http://documents.latimes.com/buddin-white-paper-20110507/.

26 Di Carlo, M. (2012, February 27). Reign Of Error: The Publication Of Teacher Data Reports In New York City. *Shanker Blog*. Retrieved April 10, 2012, from http://shankerblog.org/?p=5189#more-5189.

27 Please see the discussion in the technical appendix under "Significance Calculations."

28 The corresponding results for 2009 predicted from effects calculated from 2004-2008 are shown in Figure 9. These show somewhat better agreement between the two periods. This may be due to the absence of imputed data in 2009, as opposed to 2010.

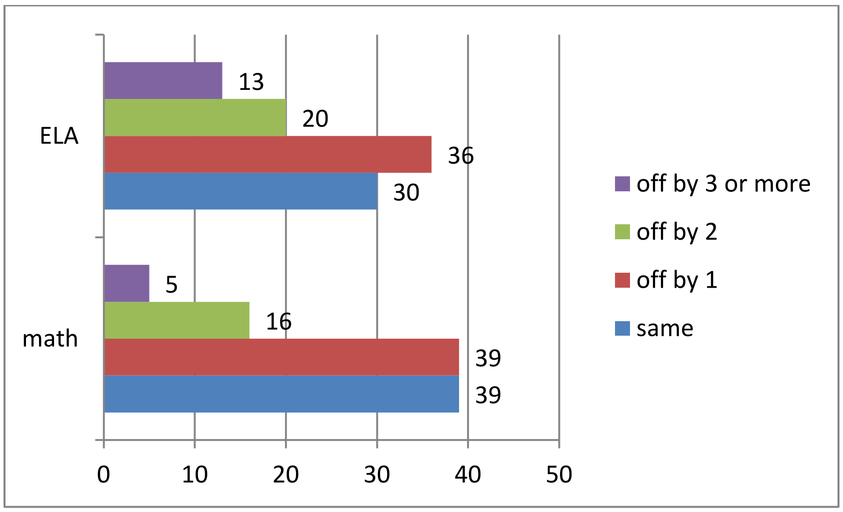

Figure 9. Percentages of rating changes between teacher-linked effects estimated for 2004-2008 and value added in 2009.

Since the value added for 2010 is calculated using a different method than the persistent teacher-linked effect, the effectiveness ratings for 2010 are assigned based on the 20th, 40th, 60th and 80th percentiles for the 2010 value-added results.

29 Di Carlo, M. (2010). *Teacher Value Added Scores: Publish and Perish*. Shanker Blog. Retrieved May 25, 2012, from : http://shankerblog.org/?p=1093.

30 Rothstein, J. (2010). Teacher Quality in Educational Production: Tracking, Decay, and Student Achievement. *Quarterly Journal of Economics, 125*(1), 175-214.

31 Chetty, R., Friedman, J.N., & Rockoff, J.E. (2011 December). The Long-Term Impacts of Teachers: Teacher Value-Added and Student Outcomes in Adulthood, NBER Working Paper No. 17699.

32 Hinchey, P. H. (2010). *Getting Teacher Assessment Right: What Policymakers Can Learn from Research*. Boulder, CO: National Education Policy Center. Retrieved [April 10, 2012] from http://nepc.colorado.edu/publication/getting-teacher-assessment-right.

33 Rothstein, J. (2011). Review of "Learning About Teaching: Initial Findings from the Measures of Effective Teaching Project." Boulder, CO: National Education Policy Center. Retrieved April 11, 2012, from http://nepc.colorado.edu/thinktank/review-learning-about-teaching.

34 Kane, T. J. & Staiger, D. (2008). *Estimating teacher impacts on student achievement: An experimental evaluation*. NBER working paper.

Chetty, R., Friedman, J.N., & Rockoff, J.E. (2011 December). *The Long-Term Impacts of Teachers: Teacher Value-Added and Student Outcomes in Adulthood*, NBER Working Paper No. 17699. Cambridge, MA: National Bureau of Economic Research.

35 Rothstein,R., Ladd, H.F., Ravitch,D., Baker,E.L., Barton,P.E., Darling-Hammond, L.,Haertel,E., Linn, R.L., Shavelson, R.J. & Shepard , L.A (2010). *Problems with the use of student test scores to evaluate teachers*, Briefing Paper #278. Washington, DC: Economic Policy Institute.

36 For example, 5% of the teachers are assigned groups of students whose mean score on the previous year's math test was less than 240 on a test with range 150-600, while 5% of the teachers were assigned groups of students whose mean score was greater than 418. For reference, for 4th grade in 2009, scores below 245 correspond to ratings of "far below proficient." while scores greater than 401 are rated "advanced." The ELA disparities are similar.

37 Raudenbush, S.W. & Bryk, A.S. (2002). *Hierarchical Linear Models, Applications and Data Analysis Methods*, Thousand Oaks: Sage Publications.

38 McCulloch, C.E. & Searle, S.R. (2001). *Generalized, Linear, and Mixed Models*. New York: John Wiley & Sons, Inc.

39 Bates, D. (2010). *lme4, Mixed Effects Modeling with R*, 44. Retrieved May 14, 2012, from http://lme4.r-forge.r-project.org/book/.