NEPC

NATIONAL EDUCATION POLICY CENTER

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*

Derek Briggs and Ben Domingue

University of Colorado at Boulder

February 2011

National Education Policy Center

School of Education, University of Colorado at Boulder
Boulder, CO 80309-0249
Telephone: 303-735-5290
Fax: 303-492-7090

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

Suggested Citation:
Briggs, D. & Domingue, B. (2011). *Due Diligence and the Evaluation of Teachers: A review of the value-added analysis underlying the effectiveness rankings of Los Angeles Unified School District teachers by the* Los Angeles Times. Boulder, CO: National Education Policy Center. Retrieved [date] from http://nepc.colorado.edu/publication/due-diligence.

Due Diligence and the Evaluation of Teachers

Derek Briggs and Ben Domingue, University of Colorado at Boulder

Executive Summary

On August 14, 2010, the *Los Angeles Times* published the results of a statistical analysis of student test data to provide information about elementary schools and teachers in the Los Angeles Unified School District (LAUSD). The analysis, covering the period from 2003 to 2009, was put forward as an evaluation of the effects of schools and their teachers on the performance of students taking the reading and math portions of the California Standardized Test.

The first step of the analysis presented in the *L.A. Times* was to predict student test scores for students on the basis of five factors: test performance in the previous year, gender, English language proficiency, eligibility for Title I services, and whether they began schooling in the LAUSD after kindergarten. These predicted scores were then subtracted from the scores that students actually obtained, with the difference being attributed to each student's teacher. If this difference was positive, this was considered to be evidence that a teacher had produced a positive effect on a student's learning. If negative, a negative effect was presumed. This process, known as value-added modeling, is increasingly being used to make strong causal judgments about teacher effectiveness, often with high-stakes consequences attached to those judgments.

The value-added analysis of elementary school teachers in the LAUSD was conducted independently by Richard Buddin, a senior economist at the RAND Corporation. As part of his analysis, Buddin produced a white paper entitled "How Effective Are Los Angeles Elementary Teachers and Schools?" We, in this new report, provide a critical review of the analysis and conclusions reached by Buddin. We conducted this review in two ways. First, we evaluated whether the evidence presented in Buddin's white paper supports the use of value-added estimates to classify teachers as effective or ineffective. This part of our report directly investigates the strength of his analysis. Second, we attempted to replicate Buddin's empirical findings through an independent re-analysis of the same LAUSD data. A hallmark of a sound analysis is that it can be independently replicated.

This new report also scrutinizes a premise of Buddin's analysis that was left unexamined: did he successfully isolate the effects of teachers on their students' achievement? Simply finding that the model yields different outcomes for different teachers does not tell us whether those outcomes are measuring what's important (teacher effectiveness) or something else, such as whether students have learning resources outside of school. Fortunately, there are good ways that a researcher can test whether such results are true or are biased. This can be done through a series of targeted statistical analyses within what we characterize as an overall "sensitivity analysis" to the robustness of Buddin's value-added model. One would expect inclusion of such a sensitivity analysis as part of any researcher's due diligence whenever a value-added model is being proposed as a principal means of evaluating teachers.

Buddin posed two specific research questions in his white paper related to the evaluation of teachers using value-added models:

1. How much does quality vary from teacher to teacher?
2. What teacher qualifications or background characteristics are associated with success in the classroom as measured by the value-added estimates?

Regarding the first question, Buddin concludes that there is in fact significant variability in LAUSD teacher quality as demonstrated by student performance on standardized tests in reading and math. To make this case, he first uses value-added modeling to estimate the effect of each teacher on student achievement. He then examines the distribution of these estimates for teachers in each test subject (e.g., mathematics and reading). For reading performance, Buddin reports a difference between high- and low-performing teachers that amounts to 0.18 student-level test score standard deviations in reading; in math it amounts to 0.27 standard deviations. These are practically significant differences.

Regarding the second question, Buddin finds that available measures of teacher qualifications or backgrounds—years of experience, advanced degrees, possession of a full teaching credential, race and gender—have only a weak association with estimates of teacher effectiveness. On this basis, he concludes that school districts looking to improve teacher quality would be well served to develop "policies that place importance on output measures of teacher performance" such as value-added estimates, rather than input measures that emphasize traditional teacher qualifications.

In replicating Buddin's approach, we were able to agree with his finding concerning the size of the measured reading and math teacher effects. These are approximately 0.20 student-level test score standard deviations in reading, and about 0.30 in math. Our results, in fact, were slightly larger than Buddin's. Our other findings, however, raise serious questions about Buddin's analysis and conclusions. In particular, we found evidence that conflicted with Buddin's finding that traditional teacher qualifications have no association with student outcomes. In our reanalysis of the data we found significant and meaningful associations between our value-added estimates of teachers' effectiveness and their experience and educational background.

We then conducted a sensitivity analysis in three stages. In our first stage we looked for empirical evidence that students and teachers are sorted into classrooms non-randomly on the basis of variables that are not being controlled for in Buddin's value-added model. To do this, we investigated whether a student's teacher in the future could have an effect on a student's test performance in the past—something that is logically impossible and a sign that the model is flawed (has been misspecified). We found strong evidence that this is the case, especially for reading outcomes. If students are non-randomly assigned to teachers in ways that systemically advantage some teachers and disadvantage others (e.g., stronger students tending to be in certain teachers' classrooms), then these advantages and disadvantages will show up whether one looks at past teachers, present teachers, or future teachers. That is, the model's outputs result, at least in part, from this bias, in addition to the teacher effectiveness the model is hoping to capture. Because our sensitivity test did show this sort of backwards prediction, we can conclude that estimates of teacher effectiveness in LAUSD are a biased proxy for teacher quality.

The second stage of the sensitivity analysis was designed to illustrate the magnitude of this bias. To do this, we specified an alternate value-added model that, in addition to the variables Buddin used in his approach, controlled for (1) a longer history of a student's test performance, (2) peer influence, and (3) school-level factors. We then compared the results—the inferences about teacher effectiveness—from this arguably stronger alternate model to those derived from the one specified by Buddin that was subsequently used by the *L.A. Times* to rate teachers. Since the *Times* model had five different levels of teacher effectiveness, we also placed teachers into these levels on the basis of effect estimates from the alternate model. If the *Times* model were perfectly accurate, there would be no difference in results between the two models. Our sensitivity analysis indicates that the effects estimated for LAUSD teachers can be quite sensitive to choices concerning the underlying statistical model. For reading outcomes, our findings included the following:

- **Only 46.4% of teachers would retain the same effectiveness rating under both models, 8.1% of those teachers identified as effective under our alternative model are identified as "more" or "most" effective in the *L.A. Times* specification, and 12.6% of those identified as "less" or "least" effective under the alternative model are identified as relatively effective by the *L.A. Times* model.**

For math outcomes, our findings included the following:

- **Only 60.8% of teachers would retain the same effectiveness rating, 1.4% of those teachers identified as effective under the alternative model are identified as ineffective in the *L.A. Times* model, and 2.7% would go from a rating of ineffective under the alternative model to effective under the *L.A. Times* model.**

The impact of using a different model is considerably stronger for reading outcomes, which indicates that elementary school age students in Los Angeles are more distinctively sorted into classrooms with regard to reading (as opposed to math) skills. But depending on how the measures are being used, even the lesser level of different outcomes for math could be of concern.

Finally, in the third and last stage of our analysis we examined the precision of Buddin's teacher effect estimates—whether the approach can be used to reliably distinguish between teachers given different value-added ratings. We began by computing a 95% confidence interval, which attempts to take potential "sampling error" into account by providing the range that will capture the true value-added for that teacher 95 of 100 times. Once the specific value-added estimate for each teacher is bounded by a confidence interval, we find that between 43% and 52% of teachers cannot be distinguished from a teacher of "average" effectiveness. Because the *L.A. Times* did not use this more conservative approach to distinguish teachers when rating them as "effective" or "ineffective", it is likely that there are a significant number of false positives (teachers rated as effective who are really average), and false negatives (teachers rated as ineffective who are really average) in the *L.A. Times'* rating system. Using the Times' approach of including only teachers with 60 or more students, there was likely a misclassification of approximately 22% (for reading) and 14% (for math).

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*

Introduction[1]

On August 14, 2010, prior to the start of the 2010-11 academic school year, the *Los Angeles Times* published results from a statistical analysis of elementary schools and teachers in the Los Angeles Unified School District (LAUSD).[2] The analysis was used to evaluate the effects of schools and their teachers on the performance of students taking the reading and math portions of the California Standardized Test between 2003 and 2009. To accomplish this for any given year, the test scores that were predicted for students were compared with the scores that were actually obtained. The predicted scores took account of their prior grade test performance, gender, English language proficiency, eligibility for Title 1 services, and whether they joined the LAUSD after kindergarten. The difference between the actual and predicted score, known as a *residual*, was then attributed to the teacher or school with which students were associated. If the residual was positive, it was considered evidence that a teacher or school had produced a positive effect on a student's learning. If negative, a negative effect was presumed.

The process loosely described above, so-called *value-added assessment* or *value-added modeling*, has become the latest lightening rod in the policy and practice of educational accountability. The method has been championed as a significant improvement over preexisting approaches (e.g., those used by states to comply with the federal No Child Left Behind law) that essentially compare schools solely on the basis of their students' achievement levels at the end of a school year.[3] Few would argue that it is not an improvement. However, value-added models also lead to strong causal interpretations about what can be inferred from a single statistic. And because the method is being applied not just to schools, but also to rate a school's teachers, these causal interpretations can strike a very personal chord.

In Los Angeles, teachers were classified into one of five levels of "effectiveness" for their teaching in reading, math and a composite of the two. The decision by the *L.A. Times* to make these results publicly available at a dedicated web site, and to publish an extensive front page story that contrasted—by name—teachers who had been rated by their level of effectiveness was promptly criticized by many as a public "shaming" of teachers.[4] In contrast, U.S. Secretary of Education Arne Duncan argued that teachers and schools should have "nothing to hide" and that members of the public (particularly parents) have a right to the information derived from value-added assessments.[5] There is reason to believe that what has occurred in Los Angeles could be a harbinger for other cities and school districts, as the use of value-added assessments to evaluate teachers becomes more common. In fact, at the time of this writing the *New York*

Post and several other New York media outlets were involved in litigation as part of an attempt to publish value-added ratings of New York City teachers.

The purpose of the present report is to evaluate the validity of the ratings themselves, not to weigh in on the wisdom of the decision by the *L.A. Times* to publish teacher effectiveness ratings. The value-added analysis of elementary school teachers in the LAUSD was conducted by Richard Buddin, a senior economist at the RAND Corporation.[6] As part of his analysis, Buddin produced a white paper[7] entitled "How Effective are Los Angeles Elementary Teachers and Schools?" Our first objective is to provide a critical review of the analysis and conclusions reached by Buddin. We conduct this review by evaluating whether the evidence presented in Buddin's white paper supports the high-stakes use of value-added estimates to classify teachers as effective or ineffective. We also attempt to replicate Buddin's empirical findings through an independent re-analysis of the same LAUSD data. Our second objective is to scrutinize a premise of Buddin's analysis that was unexamined in his white paper: that he has successfully isolated the effects of teachers on their students' achievement. To this end we present the results from the kind of "sensitivity analysis" that one should expect as due diligence any time a value-added model is being proposed as a principal means of evaluating teachers. We highlight especially those cases where the sensitivity analysis leads to substantively different inferences than those suggested on the basis of Buddin's white paper.

In what follows we will focus only on Buddin's analyses that relate to inferences about teacher effectiveness rather than school effectiveness. A quick note on terminology: a value-added model can be viewed as a subset of a broader class of growth models in which the explicit purpose is to make causal inferences about some educational treatment or intervention. In this sense, while we would be uncomfortable about labeling the aggregated residuals from a growth model as estimates of teacher "effects," it is entirely consistent with the purpose of a value-added model, so we will intentionally invoke causal language in referring to teacher effects throughout.[8] What is primarily at issue in our reanalysis is whether these effects are being estimated without bias—that is, whether they are systematically higher or lower for certain kinds of teachers in certain kinds of classrooms in the LAUSD.

Findings and Conclusions Of Buddin's White Paper

Buddin poses two specific research questions[9] at the outset of his white paper:

1. How much does value-added vary from teacher to teacher?
2. What teacher qualifications or background characteristics are associated with success in the classroom as measured by the value-added estimates[10]?

He finds that there is indeed significant variability in teacher value-added with respect to student performance on tests of reading[11] and math achievement. To make this case, he begins by estimating the effect of each teacher on student academic achievement (using a model that we describe in the next section). He then compares, for each test subject, the achievement difference that would be predicted of students with teachers that are one standard deviation apart on the effectiveness distribution. For reading performance, this amounts to 0.18 of a

standard deviation of student-level scores; in math this amounts to 0.27. For the second research question, Buddin finds that available measures of teacher qualifications or backgrounds—years of experience, advanced degrees, possession of a full teaching credential, race and gender—have only a weak association with his estimates of teacher effectiveness.

These findings lead Buddin to conclude, among other things, that school districts looking to improve teacher quality would be well served to develop "policies that place importance on output measures of teacher performance" (p. 18), such as value-added estimates, rather than input measures that emphasize traditional teacher qualifications. He also argues in favor of merit pay systems that would "realign teaching incentives by directly linking teacher pay with classroom performance" (p. 18). Notably, all of Buddin's conclusions presume that a statistical model can be used to validly and reliably estimate the effects of teachers on student achievement.

The Report's Rationale for Its Findings and Conclusions

Buddin's findings derive from the specification of a statistical model to estimate the "value" that individual teachers "add" to the academic achievement of their students. Because this is so critical to his analysis, and because we will be presenting the results from a replication of this model, we devote considerable attention here to a conceptually-oriented presentation of it. (For a more technically-oriented presentation, we refer the reader to the appendix of this report.)

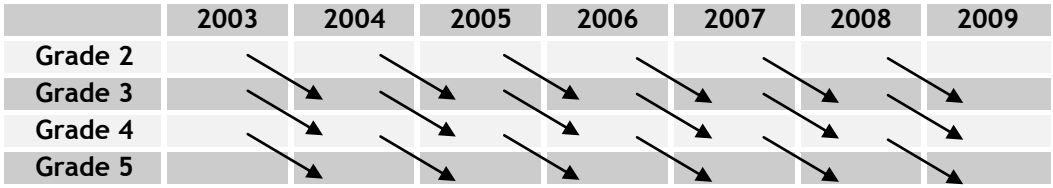

Figure 1. Longitudinal Student Cohorts Used in LAUSD Analysis

The data made available to the *L.A. Times* by the LAUSD have a longitudinal structure that spans the school years from 2002-03 through 2008-09. Figure 1 is meant to help the reader appreciate the number of student cohorts by grade that this represents. Each arrow in Figure 1 represents a distinct cohort of students enrolled in the third, fourth, or fifth grade in a given year who would have also taken reading and math tests in the previous grade. So, if a teacher has taught in the third, fourth, or fifth grade over this time span, this dataset would provide test score information for as many as six different student cohorts. This rather large number of student cohorts is relatively rare as a basis for estimating teacher effects with a statistical model and represents a strength of the LAUSD data.

Buddin uses these data as the basis for his value-added model of student achievement. Consider a student taking a test in a given grade (3, 4 or 5) and year (2004-2009). The student's performance on this test is modeled[12] as

$$CurrentYrScore = a*PriorYrScore + b*FEMALE + c*ELL + d*TITLE1 + e*JoinPostK + f_1*TEACHER_1 + f_2*TEACHER_2 + \ldots + f_J*TEACHER_J + \text{"black box"}.$$

For each test subject (reading or math), the current year test score of the student (*CurrentYrScore*) is modeled as a function of the prior year score (*PriorYrScore*). Both of these variables are standardized within each grade and year so that each variable has an average of 0 and a standard deviation of 1. This means, for example, that a student with a current year test score that is positive has performed above average in a normative sense, while a student with a score that is negative has performed below average.[13] Current year test performance is also modeled as a function of the variables *FEMALE*, *ELL*, *TITLE1*, and *JoinPostK*, which represent indicator (i.e., "dummy") variables that take on a value of "1" if a student is female, an English Language Learner, eligible for Title I services,[14] or joined an LAUSD school after kindergarten, and a value of "0" otherwise. The most important variables in the model are indicator variables for LAUSD elementary school teachers: $TEACHER_1$, $TEACHER_2$, … , $TEACHER_J$. For each student, one of these variables will take on a value of "1" to represent the teacher to whom the student has been assigned in the current year and grade, while the rest are set to "0" (and are thus effectively removed from the analysis as regards that particular student). The letters *a* through *f* represent parameters (or coefficients) of the model, where the values of *a* through *d* indicate the unique contribution of the variables described above on student achievement. Buddin's primary interest is to make inferences on a teacher-by-teacher basis using the numerical value of the parameter f_j. This parameter represents the increment in a student's current year test score that is attributable to the teacher to whom he or she was assigned. The subscript *j* is used to index a specific teacher and can equal anywhere from 1 (e.g., "teacher 1") to "J" (e.g., "teacher 7809") for any given teacher in the sample under analysis. In this model, the larger the value of f_j, the larger the value that a specific teacher has added to the student's achievement. Finally, the term labeled "black box" represents a numerical value that for each student, is drawn at random from a distribution with the same mean (0), variance (a constant value), and shape (normal, i.e., bell-shaped).

The term "value-added" as applied to the model above is intended to have the same meaning as the term "causal effect"—that is, to speak of estimating the value-added by a teacher is to speak of estimating the causal effect of that teacher. But once stripped of the Greek symbols and statistical jargon, what we have left is a remarkably simple model that we will refer to as the "LAVAM" (Los Angeles Value-Added Model). It is a model which, in essence, claims that once we take into account five pieces of information about a student, the student's assignment to any teacher in any grade and year can be regarded as occurring at random. If that claim is accurate, the remaining differences can be said to be the value added or subtracted by that particular teacher.

Defending this causal interpretation is a tall order. Are there no other variables beyond the ones included in the model that contribute to a student's current year test score? What about parent education levels, school attendance, and involvement in a special education or language emersion program, just to name a few? What if a teacher's causal effect is itself caused by one or more of these variables that either were not (or could not) be included in the model?[15] What if particular teachers are more or less effective for certain kinds of student? And why should we believe that the "black box" portion of the model is independent from one student to the next? Any parent with two or more children would question such an assumption. We also know that elementary-school students work and play together in small groups within classrooms, and this will introduce systemic differences in the data that may undermine what is being assumed in the

equation above. The findings and conclusions in Buddin's white paper all presume that he has addressed these sorts of questions or that they simply are not important.

The Report's Use of Research Literature

Buddin's white paper is similar in narrative and rhetorical structure to a study he had previously published in the *Journal of Urban Economics* in 2009 with co-author Gema Zamarro. Buddin's analysis conducted for the *L.A. Times* differs from the journal article primarily with respect to the time span of the data (the journal study involved panel data from 2000 to 2004) and the nature of the test (from the California Achievement Test to the California Standardized Test), as well as some key details in the specification of his value-added model.

While Buddin references a number of important empirical studies that have estimated teacher effects using value-added models, he does not cite certain studies and reports that would call into question the choices made in the LAVAM specification.[16] Perhaps the most important omissions are two recent publications by UC Berkeley economist Jesse Rothstein. Rothstein introduced a statistical test that could be readily conducted to evaluate whether teacher value-added estimates are unbiased. Using longitudinal data from North Carolina, Rothstein was able to strongly reject this hypothesis for a model very similar to the LAVAM. That is, he showed the estimates to be substantially biased. In a recent working paper, Cory Koedel and Julian Betts applied the Rothstein test to data they had previously analyzed from the San Diego Unified School District.[17] They came to conclusions similar to those of Rothstein, but suggested that the bias Rothstein's approach had uncovered could be mitigated through the combination of averaging over multiple cohorts of students (as Buddin does with the LAUSD data) and by restricting the norm group used to interpret teacher effects to only those teachers who had taught the same students (i.e., including student "fixed effects," something Buddin does not do[18]).

The general form of the educational "production function"[19] Buddin introduces in his white paper is a fairly standard starting point in the research on value-added models that has been conducted by economists. However, the way that Buddin has implemented the model with the LAUSD data is quite different from other empirical implementations because he includes far fewer "control" variables. Consider, for example, a study Buddin cites by Tom Kane and Douglas Staiger that also used data from LAUSD schools.[20] In their analysis, Kane and Staiger specified three models with different sets of control variables. The specification that is closest to that of the LAVAM included six additional student-level variables that were not part of Buddin's model: indicators of race/ethnicity, migrant status, homeless status, participation in gifted and talented programs or special education, and participation in the free/reduced-price lunch program. Kane and Staiger also included classroom-level versions of all these student-level variables.

Similarly, the value-added models specified by economists recently[21] for data from New York City include a laundry list of student and classroom-level variables that go far beyond those included by Buddin in the LAVAM. These variables include whether a student attended summer school, how often a student was absent, and how often a student had been suspended. These were included as indicators at the student level and as percentages at the classroom level. Table 1 formally contrasts the "control" variables included by Buddin in the LAVAM with those that

have been included in other prominent value-added applications that derive from the same basic educational production function. This does not necessarily mean that these more complex model specifications are right while the one specified by Buddin is wrong. However, since they differ significantly in terms of omitted variables, the results from the Kane & Staiger study provide no justification for the specification of the LAVAM.

Table 1. Differences in Control Variables Included in Large-Scale Value-Added Model Implementations.

	Buddin 2010 Los Angeles	Kane & Staiger 2008 Los Angeles	Wisconsin VA Research Center 2010 New York City
Student-Level Control Variables Included[1]	Prior Year Test Score (subject specific), Gender, Title 1, ELL, Joined District after Kindergarten	Prior Year Test Score (subject specific), Gender, ELL, Title 1, Race/Ethnicity, Migrant, Homeless, Gifted and Talented Program, Special Education Program, Free/Reduced Price Lunch Eligible	Prior Year Test Score (both math and reading), Gender, Race/Ethnicity, ELL, Former ELL, Disability, Free Lunch, Reduced Lunch, Summer School, Absences, Suspensions, Retained in Grade before Pretest Year, Same School Across Years, New to City in Pretest Year
Classroom-Level Variables Included[2]	None	Prior Year Test Score, Gender, Title 1, ELL, Race/Ethnicity, Migrant, Homeless, Gifted and Talented Program, Special Education Program, Free/Reduced Price Lunch Eligible Price Lunch Status	Prior Year Test Score (both math and reading), Class Size, Gender, Race/Ethnicity, ELL, Former ELL, Disability, Free Lunch, Reduced Lunch, Summer School, Absences, Suspensions, Retained in Grade before Pretest Year, Same School Across Years, New to City in Pretest or Post Test Year

[1] All variables are dichotomized as dummy variables with the exception of prior year test score.

[2] All variables are averages of student-level variables that are interpretable as proportions, with the exception of prior year test score and class size.

Review of the Report's Methods: Reanalysis of LAUSD Data

There were two stages to our re-analysis of the LAUSD data. The purpose of the first stage was to see if we could replicate Buddin's analysis and thereby come to the same findings about the variability of teacher effect estimates and the weak association between teacher qualifications and backgrounds and these estimates. The purpose of the second stage of our analysis was to critically examine the premise that the estimates from the LAVAM can be validly interpreted as a teacher's causal effect. To this end we perform a sensitivity analysis in which we asked the following questions:

1. Is there evidence that supports the interpretation that Buddin's estimates of "value-added" are not biased by the sorting of students and teachers on variables not included in the model?
2. How sensitive are the rankings of teachers to another defensible specification of the underlying model?
3. How precisely are teachers being classified as effective or ineffective?

Replicating Buddin's Analysis

Buddin estimates the parameters (i.e., the values for *a-f*) of the LAVAM using linear regression. Note that because there are up to six cohorts of students available per teacher, a teacher's effect estimate is the average of the teacher's effect for each cohort he or she has taught over the 2004 to 2009 time period (weighted by the number of students per cohort). After adjusting for "sampling error,"[22] Buddin arrives at a corrected effect estimate for each teacher. He then computes a standard deviation across the LAUSD for these estimates in both reading and math. Next, he takes the effect estimates for each teacher from this initial regression, and uses them as the outcome variable in a second regression with teacher-level predictor variables for years of experience, education, credential status, race and gender. Finally he examines the significance of these variables' association with teacher effectiveness. (For a technically oriented presentation of these steps and how we went about reproducing them, see our description in the appendix of this report.)

Our replication of Buddin's principal findings led to mixed results. While we were not able to exactly replicate the parameter estimates from Buddin's student-level regressions (see Appendix Table A-1 for a comparison), the standard deviations we computed for teacher effectiveness distributions in reading and math were in the same ballpark (though slightly larger).[23] For reading outcomes, our adjusted estimate was 0.231 student-level standard deviations compared with Buddin's 0.181. For math outcomes our adjusted estimate was 0.297 compared with Buddin's 0.268. These results are supportive of the finding that if, in fact, we are validly estimating teacher effects on student achievement with the LAVAM, then there is significant variability in these effects, and the variability is larger in math than it is in reading.

On the other hand, our results from replicating the teacher-level regressions are not consistent with those shown by Buddin, either in terms of our parameter estimates or in our interpretation of their practical significance. Table 2 compares the regression coefficients, R^2 and sample sizes from our teacher-level regressions with those reported by Buddin. There are some important discrepancies between the two sets of results, and we have highlighted the more notable differences with the rows set in bold italic. Like Buddin, we find that inexperienced teachers (those in their first two years on the job) are, on average, the least effective, and the association is stronger in reading than in math. However, our estimates for the magnitudes of these associations in reading and math (-0.11 and -0.07) are much larger than those reported by Buddin (-0.05 and -0.02).

We find this to be the case for a number of other traditional qualification variables as well. While we agree with his finding that there is no statistically significant association between credential

Table 2. Replication of LAVAM: Buddin's Table 5 "ELA and Math Teacher Effects and Teacher Characteristics"

	Reading			Math		
	Buddin	Briggs & Domingue	Effect Size[1]	Buddin	Briggs & Domingue	Effect Size[1]
Experience < 3 years	***-.05****	***-0.11****	***-0.48***	***-0.02***	***-0.07****	***-0.21***
	(0.01)	(0.01)		(0.01)	(0.02)	
Experience 3-5 years	***-.01****	***-0.06***	***-0.26***	***0.02****	***-0.02****	***-0.06***
	(0.006)	(0.01)*		(0.01)	(0.01)	
Experience 6-9 years	***0.00***	***-0.04****	***-0.17***	0.02*	0	0
	(0.01)	(0.01)		(0.01)	(0.01)	
Bachelor's + 30 semester hours	0.00	0	0.00	***0.00***	***0.03****	***0.09***
	(0.01)	(0.01)		(0.01)	(0.01)	
Master's	***0.01***	***0.03****	***0.13***	***0.00***	***0.04****	***0.12***
	(0.01)	(0.01)		(0.01)	(0.02)	
Master's + 30 semester hours	0.01	0.02	0.09	***0.01***	***0.06****	***0.18***
	(0.01)	(0.01)		(0.01)	(0.01)	
Doctorate	-0.03	0.02	0.09	-0.04	0	0
	(0.02)	(0.03)		(0.03)	(0.04)	
Full Teaching Credential	***0.00***	***0.05***	***0.22***	0.01	0.04	0.12
	(0.01)	(0.03)		(0.02)	(0.04)	
Black/African American	***-0.05****	***-0.10****	***-0.43***	***-0.07***	***-0.11****	***-0.34***
	(0.01)	(0.01)		(0.01)	(0.01)	
Hispanic	***-0.01***	***-0.06****	***-0.26***	-0.01	0	0
	(0.00)	(0.01)		(0.01)	(0.01)	
Asian/Pacific Islander	0.03*	0.01	0.04	0.07*	0.07*	0.21
	(0.01)	(0.01)		(0.01)	(0.01)	
Female	0.04*	0.07*	0.30	0.02*	0.03*	0.09
	(0.00)	(0.01)		(0.01)	(0.01)	
Grade 4	-0.01	-0.01	-0.04	-0.01	-0.03	-0.09
	(0.01)	(0.01)		(0.01)	(0.01)	
Grade 5	-0.01*	0	0.00	-0.01	-0.01	-0.03
	(0.01)	(0.01)		(0.01)	(0.01)	
Constant	-0.02	0.02		-0.03	-0.06	
	(0.01)	(003)		(0.02)	(0.04)	
R-squared	0.027	0.059		0.020	0.030	
Number of Teachers	8719	7809		8719	7888	
SD of Teacher Effects from Student-level Regression (unadjusted)	.210	.268		.297	.326	
SD of Teacher Effects from Student-level Regression (adjusted)	.181	.231		.268	.297	

Notes: *Statistically significant at 5% level. Standard errors are in parentheses. The omitted categories are White non-Hispanics, Male, BA only, no full teaching credential, experience of 10 or more years, and grade 3. The dependent variables are teacher effect estimates from LAVAM unadjusted for sampling error. Estimation done through use of Feasible Generalized Least Squares.

1 Effect size = Regression Coefficient Estimated by Briggs & Domingue divided by adjusted SD of Teacher Effects

status and teacher effects, it is hard to read much into this since the available credential indicator variable makes no distinction as to the type or quality of the credential, and only 10% of the teachers in the sample lack a teaching credential. Finally, we find evidence of statistically significant associations with effectiveness by teacher race and gender.[24]

When working with large sample sizes, it is especially important to draw a distinction between regression results that are statistically significant and those that are *practically* significant. Relative to a standard deviation for student-level test scores, which is 1.0, the regression coefficient estimates reported in Table 2 appear small. Even the results from our regressions, though they tend to be larger in magnitude than those reported by Buddin, are never larger in absolute value than 0.11. Along these lines, Buddin argues that even for the variables where a statistically significant association exists, the size of the regression coefficients are small enough to be considered practically insignificant:

> Teacher experience has little effect on ELA scores beyond the first couple years of teaching—teachers with less than 3 years of experience gave teacher effects 0.05 standard deviations lower than comparable other teachers with 10 or more years of experience. Students with new teachers score 0.02 standard deviations lower in math than with teachers with 10 or more years of experience, but the effect is not statistically different from zero. These effect sizes mean that students with the most experienced teachers would average 1 or 2 percentile points higher than a student with a new teacher. These effects are small relative to the benchmarks established by Hill et al. (2008). (Buddin, 2010, p. 10-11.)

The problem with this interpretation is the frame of reference, because the units of analysis for the regression results presented in Table 2 are not students, but teachers. Recall that the teacher effects estimated by Buddin in his first-stage regression had adjusted standard deviations of 0.18 and 0.27 for reading and math outcomes respectively. *These* standard deviations (obviously much smaller than 1) are the relevant frame of reference against which the regression coefficients should be evaluated. Therefore, the finding that a teacher with fewer than three years of experience has an estimated effect on reading achievement that is 0.05 *student-level test score standard deviations* lower than a teacher with 10 or more years of experience should be more appropriately interpreted as an *effect size at the teacher-level* of –0.5 / 0.18 = –0.28. The last columns for each subject area presented in Table 2 rescale the regression coefficients we found from our replication analysis, after dividing these estimates by the adjusted teacher effect standard deviations for reading and math (0.231 and 0.297). The impact of this is to highlight that a number of teacher variables that show up as statistically significant are also quite practically significant once they have been expressed in the appropriate effect-size metric. Hence we can conclude that some traditional teacher qualifications, such as experience and educational background, appear to matter as much to being an effective teacher as having an effective teacher matters to being a high-achieving student.[25]

How can it be that we arrived at substantively different numbers in our replication of Buddin's analysis if we were using the same data source and the same model specifications? One possibility is that the difference in our results can be explained by a difference in the sample of teachers and students who were included in our respective regressions. Not all students and teachers present in grades 3 through 5 between the years 2004 and 2009 in the LAUSD were

included in his analysis. Buddin's white paper provides little information about decisions made with regard to sample restrictions. However, we learned from Buddin (personal communication) that he excluded students for a particular grade/year combination if they were missing either reading or math test scores, a teacher identifier, a school identifier, or a prior year test score. In addition, teachers in schools with 100 or fewer test-takers and in classrooms with 15 or fewer students were excluded.[26] When we imposed these same sample restrictions before replicating the first stage student-level regressions, we were left with a total of 733,193 and 743,764 students with valid test scores in reading and math, respectively, and a total of 10,810 and 10,887 teachers (respectively) linked to these students. These numbers are smaller than the 836,310 students and 11,503 teachers (the same for both reading and math outcomes) reported by Buddin in his Table 4. For Buddin's second stage teacher-level regressions, an additional restriction was imposed limiting the analysis to only those teachers with at least 30 students over the six-year time span of the data. This restriction, along with missing teacher-level predictor variables, reduced our sample size to 7,809 in reading and 7,888 in math, numbers once again smaller than the 8,719 reported by Buddin.[27] Of course, if the inclusion or exclusion of a subset of teachers and students can lead to the differences in the regression results we have reported here, this represents an important preliminary sensitivity analysis in itself.

Sensitivity Analysis

Are these good estimates of teacher "quality"?

When a student performs better than we would predict on a standardized test, can we attribute this residual to his or her teacher? The answer to this question depends on our ability to control for all other variables that contribute to student test performance. If one or more of these variables have been omitted from the model, and if the variables are correlated with how students and teachers have been assigned to one another, we should be understandably hesitant to make a causal attribution. One way to formally evaluate this concern is to conduct the empirical test alluded to earlier that has been introduced by Jesse Rothstein. Rothstein's "falsification" test rests on the logically compelling premise that a student's teacher in the future cannot have an effect on his or her performance in the past. If this counterfactual state of affairs appears to be the case, then it suggests that students are being sorted to future teachers (or vice-versa) on the basis of variables that predict student test performance that are not being controlled for in the model. Put differently, while a model such as the LAVAM explicitly controls for differences in a student's prior year test scores, it implicitly assumes that teachers are not more or less likely to get students who had performed above or below expectation in their prior grade. This is something that is easy enough to test by making a small change to the LAVAM's specification of teacher indicator variables; namely we can exchange a student's indicator variable for the teacher he or she has in the current grade (e.g., grade 4) with an indicator variable for the teacher the student will have in the next grade (e.g., grade 5).[28] After doing this and estimating our regression coefficients, we test the formal statistical hypothesis that future teachers have no effect on student achievement (i.e., all f_j's are equal to 0). If we are able to reject this hypothesis and instead we observe an estimated distribution of future teacher effects with considerable variability (a significant standard deviation for the f_j's), this raises a red flag,

warning us against drawing the conclusion that the LAVAM is producing unbiased estimates of the causal effects of teachers on their students.

Table 3. Performing Rothstein's Falsification Test on the LAVAM

	Reading Outcomes in			Math Outcomes in		
	Grade 3	Grade 4	Grade 5	Grade 3	Grade 4	Grade 5
SD of Effects for Teachers in						
Grade 4	***.223***	.221		***.211***	.298	
Grade 5		***.180***	.204		***.227***	.306
Control Variables						
Title 1 Status	X	X	X	X	X	X
Gender	X	X	X	X	X	X
English Language Learner	X	X	X	X	X	X
Joined after Kindergarten	X	X	X	X	X	X
Year	X	X	X	X	X	X
Grade 2 Test Score	X			X		
Grade 3 Test Score		X			X	
Grade 4 Test Score			X			X

Note: The p-values for F-tests that teacher effects = 0 are< .001 for all outcomes.

We performed this test on the LAVAM specification by examining whether it appears that grade 4 and 5 teachers were having "effects" on the performance of grade 3 and 4 students, respectively. The results are summarized in Table 3; the null hypothesis of no effect for both reading and math outcomes was rejected decisively in each grade ($p < .001$). For each test subject there are three main columns corresponding to LAVAM specifications in which the student test score outcomes of interest are in grades 3, 4 and 5. The key rows of interest are those that indicate the standard deviation of the grade 4 and grade 5 teacher effect estimates (adjusted to account for "sampling error") associated with each of the three columns. For the row representing grade 4 teachers, the cell corresponding to a grade 3 column represents a "counterfactual" (set in bold italic type)—the performance of teachers' current grade 4 students in grade 3. The cell corresponding to the grade 4 column represents the performance that is actually observed by grade 4 students after they have been assigned to grade 4 teachers. A similar interpretation holds for the row with grade 5 teachers; the cell associated with the grade 4 column represent the counterfactual, the cell associated with the grade 5 columns represent the test performance that is actually been observed.

Our results indicate that the variability of the counterfactual effects are substantial. For reading outcomes, the counterfactual effects are 101% and 88% of the grade 4 and grade 5 observed effects standard deviations for grade 4 and 5 teachers, respectively. For math, the proportion is lower, but still quite large—71% and 74%. These results would support the logically impossible conclusion that the impact of teachers on student achievement in the past is almost as large (and in one case larger) than the impact on student achievement in the present. These results provide strong evidence that students are being sorted into grade 4 and grade 5 classrooms on the basis of variables that have not been included in the LAVAM, and this sorting appears to be considerably stronger with respect to variables associated with reading achievement than it is

for math achievement.[29] This is empirical evidence that the LAVAM estimates of teacher value-added are biased. What is not as clear is the practical impact of this bias. We address this in the next section.

Are teacher effectiveness classifications sensitive to the choice of variables included in the value-added model?

The LAVAM appears to be producing biased estimates of teacher effects because it omits variables that are associated both with student test performance and how students and teachers are assigned to one another. For variables that are simply not available in the LAUSD data (e.g., parental involvement, free and reduced-price lunch eligibility, etc), one can only speculate about the impact of the exclusion of these variables on effect estimates. However, for some variables that *were* available but purposefully excluded by Buddin in his specification of the LAVAM, we can evaluate the empirical impact of their exclusion. In doing so we will be restricting our attention to roughly 3,300 teachers who taught in grade 5 between 2005 and 2009 (see Figure 2) with 115,418 and 112,159 students who had previously been tested in math and reading respectively in grades 2, 3 and 4.

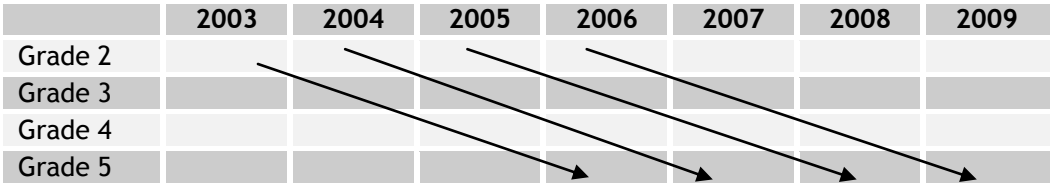

Figure 2. Subset of Longitudinal Data Used to Estimate Bias Due to Observable Omitted Variables

We consider the sensitivity of the LAVAM specification for grade 5 teachers to the exclusion of the following three sets of variables:

- Subject-specific student-level test scores in grades 2 and 3.
- The mean of grade 4 test scores in each grade 5 classroom.
- An indicator for a school's location within California's School Similarity Rank.[30]

We focus on these particular variables because they have been widely discussed as plausible confounders in the research literature.[31] The first is an explicit attempt to mitigate the sorting bias described by Rothstein, among others, by controlling for a longer history of student achievement; the second represents an attempt to control for influence of a student's peers in a given classroom; finally, the third represents an attempt to control for school-level demographics and characteristics that could otherwise be erroneously attributed to a school's teachers. We refer to the model that results when these variables have been added to those already included in the LAVAM as the "altVAM" model and proceed as though it provides "true" estimates of teacher effects. Given this, we can evaluate the bias in the LAVAM by examining how closely it approximates the "truth" represented by the altVAM.[32] We quantify this in four different ways.

- First, we compute the correlation between the teacher effects estimated from each model and the average of prior grade achievement across all the classes that were the basis for a given teacher's estimated effect. Because the altVAM model controls for this latter variable explicitly, we expect this correlation to be 0. The further this correlation departs from 0 under the LAVAM, the harder it would be to argue that teachers with higher-achieving incoming students are no more likely to be classified as effective relative to teachers with incoming students who are lower achieving.
- Second, we compute a bias term[33] for each estimated teacher effect by taking the difference between the teacher effects estimated under the LAVAM and altVAM, respectively. We then express the standard deviation of these bias terms across teachers as a proportion of the standard deviation of the "true" teacher effects under the altVAM.
- Third, we compute the correlation between the LAVAM and altVAM teacher effect estimates (after each has been adjusted to account for "sampling error").
- Finally, we examine the shift in teacher classifications by test subject when going from the altVAM to LAVAM. Recall that the *L.A. Times* classified teachers according to their quintile (i.e., their fifth) of the teacher "effectiveness" distribution. Depending on the quintile within which they fell (i.e., first 20% of distribution, second 20%, etc.), teachers were given one of five classifications: least effective, less effective, average, more effective, and most effective. We examine cross tabulations by model specification to see the extent to which the omission of the three sets of variables above leads to substantial changes in teacher classifications.

Table 4. Quantifying Bias in LAVAM Estimates of Teacher Effects Relative to altVAM

	Reading		Math	
	LAVAM	altVAM	LAVAM	altVAM
Correlation with mean prior achievement of classroom	0.50	-0.08	0.27	-0.06
SD of Teacher Effects	0.21	0.16	0.31	0.28
SD of Bias	0.14		0.12	
Bias as % of altVAM SD	88%		43%	
Correlation of LAVAM and altVAM Teacher Effects	0.76		0.92	

Note: The same sample of teachers and students is being used for both LAVAM and altVAM.

Table 4 summarizes the results from the first three of the four approaches described above. As expected, the correlation of altVAM teacher effect estimates with the average prior achievement of a teacher's students is very close to 0 (actually slightly negative) for both reading and math outcomes. In contrast, the correlation under the LAVAM is considerably higher than 0 for math outcomes (0.27), and much higher for reading outcomes (0.50). The different impact of this bias by tested subject is also evident when the bias in teacher effects is summarized as a proportion of the "true" variability in teacher effects found in the altVAM. Here the standard deviation of the bias for math outcomes is a substantial 43% of the altVAM teacher effect standard deviation. For reading outcomes, the variability in bias is 88% of the variability in the true effects. Finally,

we note that while the intercorrelation of teacher effects across the two models is strong in both math and reading, it is considerably stronger for math (0.92) than for reading (0.79).

Table 5. Changes in Teacher Quintile Ranking Going from altVAM to LAVAM: Reading (Percentages)

		Teacher Effect Quintile Ranking from altVAM				
		1	2	3	4	5 (best)
Teacher Effect Quintile Ranking from LAVAM	1	64.4	25.6	9.2	0.8	0.0
	2	18.5	39.8	26.4	13.5	1.8
	3	9.1	17.5	35.3	40.0	7.3
	4	6.7	10.8	18.2	32.9	31.4
	5 (best)	1.4	6.4	10.9	21.9	59.5

Note: Total Number of Teachers = 3,298, Teachers per column = 660, 659, 660, 659, 600. Values in cell represent column percentages. Columns sum to 100%.

Table 6. Changes in Teacher Quintile Ranking Going from altVAM to LAVAM: Math (Percentages)

		Teacher Effect Quintile Ranking from altVAM				
		1	2	3	4	5 (best)
Teacher Effect Quintile Ranking from LAVAM	1	76.3	21.6	2.1	0.0	0.0
	2	19.2	50.1	27.9	2.9	0.0
	3	4.1	23.4	45.6	25.8	1.2
	4	0.5	4.8	22.6	52.8	19.3
	5 (best)	0.0	0.2	1.8	18.6	79.5

Note: Total Number of Teacher = 3,315, Teachers per column = 663. Columns sum to 100%.

The key question here is whether we observe a significant shift in the classifications of teachers as "effective" or "ineffective" when moving from the altVAM to LAVAM specification. As is evident from the cross tabulations presented in Tables 5 and 6, this appears to be the case for both test score outcomes. Overall only 46.4% and 60.8% of teachers maintain the same quintile rankings for reading and math outcomes, respectively. (This calculation is not shown in the tables above but is based on averages across the cells.) Note that classifications are a great deal more fluid within the middle three quintiles of the effectiveness distributions. But even when we only consider teachers in the top and bottom quintiles under the altVAM reading outcome specification, we find that only 64.4% and 59.5%, respectively, of these teachers maintain their position. The results are more consistent for math outcomes, where 79.5% of teachers in the top and 76.3% in the bottom quintiles would maintain the same position. Finally, we find that 8.1% of teachers classified as "more" or "most" effective for reading outcomes using the altVAM would shift to being classified as "less" or "least" effective using the LAVAM, while 12.6% would shift in the opposite direction (from ineffective to effective classifications; again, these calculations are not directly shown in the tables above). For math outcomes, these sorts of dramatic shifts would be much rarer—1.4% and 2.7%.

Given these results it would be hard to argue that teachers or the stakeholders evaluating them would be indifferent to the choice of model used to produce these ratings, especially in the case

of reading outcomes. And these results are likely to understate the magnitude of bias, since the altVAM specification itself omits many theoretically important variables, such as family poverty and student motivation, to name just a few.

Why did Buddin exclude the additional variables in our altVAM specification? One possible reason for excluding additional measures of prior grade achievement is that this reduces the number of teachers and students that can be included in the analysis. In the altVAM, because three prior years of test scores for each student are needed, it is impossible to estimate effects for teachers when students are in grades 3 and 4. School-level control variables in the form of our School Similarity Rank indicators might be excluded either because there is much less variability in student performance across schools than there is within, or because of a desire to compare teachers with a normative group that spans the entire district. Finally, a decision might be made to exclude a classroom-level variable for prior achievement, because this essentially creates different performance expectations for students simply because they happened to land in a classroom in which their peers are relatively higher or lower achieving.[34] In other words, the question of what variables to include in a value-added model is not straightforward.[35] There may well be strong pragmatic or ethical reasons for excluding certain variables even if this decision adds bias to the teacher effects being estimated. Our criticism of the LAVAM is therefore not so much that it omits important variables (though this certainly is a reasonable criticism), but that Buddin's white paper does not acknowledge the equivocal nature of these decisions and their potential impact on inferences about teacher effectiveness.

How "precise" are these teacher effect estimates?

Throughout this review and reanalysis, we have typically referred to teacher effects as *estimates*. Up to this point, however, little has been said about the precision of these estimates. For example, if a confidence interval[36] were to be formed around the estimate of a teacher effect, would the interval be narrow or wide? Before we proceed, going back now to the complete longitudinal data set used for the LAVAM student-level regression (Figure 1), we first point out that answering this question provides no insights as to whether the model is producing a valid estimate of a teacher's causal effect. The classic analogy here is to shooting an arrow at a target—by precision, we simply ask whether we are able to strike roughly the same place on the target, whether or not the place that we strike is the bull's-eye or somewhere on the edge. The inferential thought experiment at work here is as follows: if we were to randomly and repeatedly sample different cohorts of students to be assigned to the teacher in question over the given time period, how much would the estimate of the teacher's effect vary across distinct sets of student cohorts?

The answer is presented visually in the so-called "caterpillar plots" in Figure 3. For reading and math outcomes, the estimated effect of each teacher (relative to the vertical axis) is bounded above and below the line, which represents a 95% confidence interval. The length of each interval is inversely proportional to the total number of students that was the basis for that teachers' effect estimate. Note that Buddin had up to six student cohorts to work with for each teacher, which is an unusually large number relative to other empirical analyses that have been done using value-added models. The teachers with the longest confidence intervals (i.e, the least amount of precision in their estimated effects) tend to be those who taught fewer than six

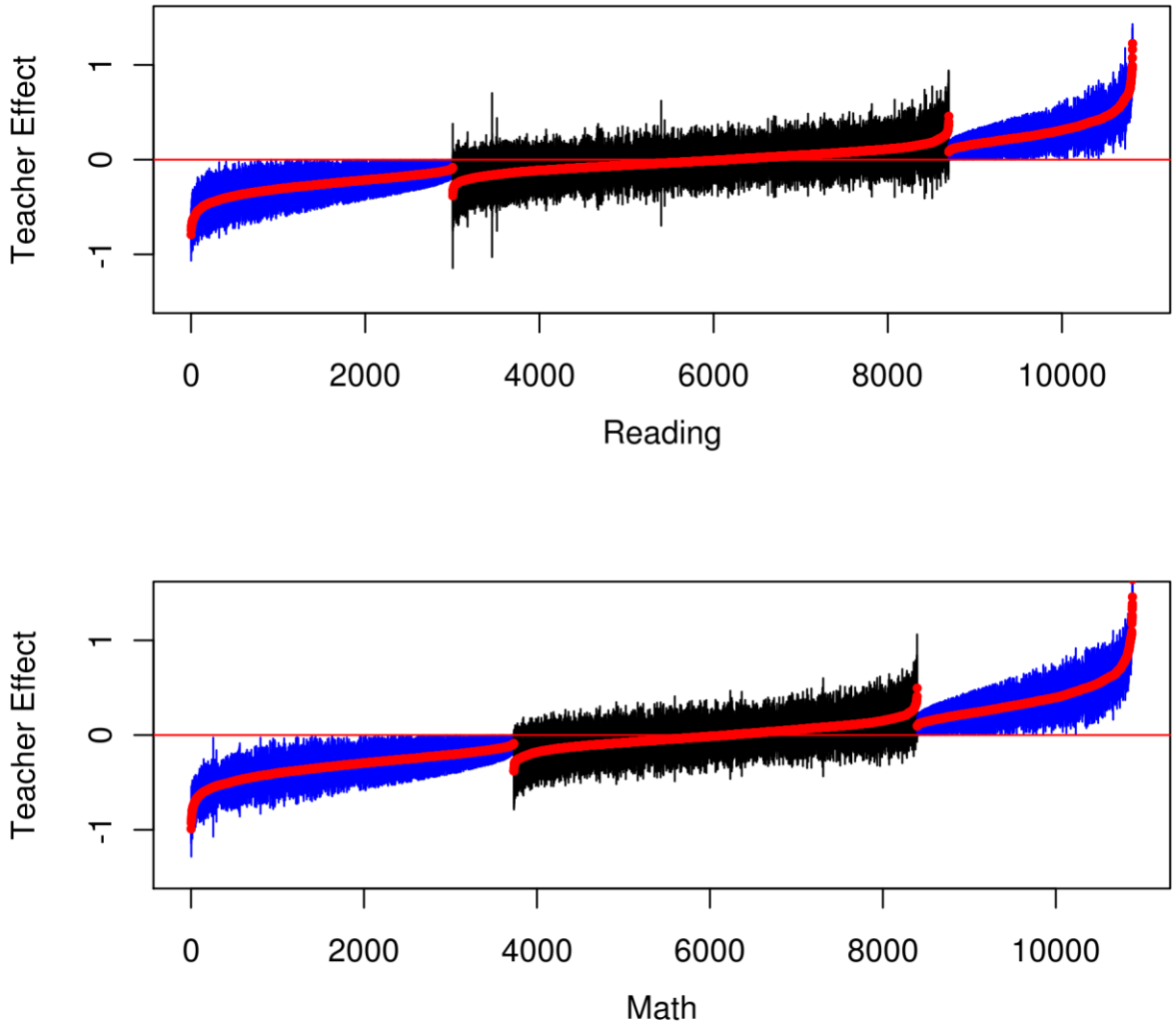

Figure 3. The Precision of Teacher Effects Estimated using the LAVAM

cohorts between 2004 and 2009 (e.g., because they had stopped teaching in LAUSD, joined LAUSD after 2004, etc.). In Figure 3 teachers are ordered into one of three groups along the horizontal axis according to whether or not their associated interval falls below the "average" effect (0), crosses it, or falls above it.[37] These results indicate that for reading outcomes, we could classify only 27.8% and 19.5% of teachers in our data (J=10,810) as "ineffective" or "effective". The remaining 52.7% of teachers are not significantly different from average. For math outcomes, though higher proportions of teachers can be classified as significantly below or above the average in their effects (34.3% and 22.9%), the sampling variability in these estimates

is still substantial enough that roughly 43% of teachers (J=10,887) cannot be distinguished from the average.

An interesting contrast is to compare the classifications of teachers into three levels of effectiveness when using a 95% confidence interval as above (average, effective, and ineffective) with the classifications that result when the teacher effectiveness distribution is broken into quintiles. How many teachers classified as effective or ineffective under the latter classification scheme would be considered only average under the former, more conservative approach? We consider a teacher with an effect in the bottom two quintiles of the distribution but with a confidence interval that overlaps 0 to be a "false negative," and the opposite extreme to be an example of a "false positive." For classifications based on reading outcomes, we find 1,325 false negative and 2,215 false positives—12.3% and 20.5% of the total sample of teachers for whom these effects could be estimated. For classifications based on math outcomes, we find 693 false negatives and 1,864 false positives—6.4% and 17.2% of the total sample of teachers for whom these effects could be estimated.[38]

Did Buddin conduct any sensitivity analyses?

The kinds of analyses presented above were not part of Buddin's white paper, and the tenor of his narrative appears to reflect great confidence regarding the validity and reliability of the teacher effects being estimated. He implicitly addresses some minimal concerns about validity by presenting the results from a teacher-level regression of his effect estimates on a variety of classroom composition variables, such as proportion of students eligible for free and reduced-price lunch services, proportion of students who are English language learners, proportion of students with parents graduating from college, and so on. These results are shown in Buddin's Table 7, and they indicate that his classroom composition variables have, at best, a very weak association with his estimates of teacher effects. Hence he argues

> These small effect sizes suggest that the value added measure is doing a good job of controlling for the mix of students assigned to individual teachers. While class composition varies considerably across LAUSD, the proportions of students with different demographic and socioeconomic factors have little effect on value added rankings of teacher effectiveness (p. 15).

What is easy to miss, because it is only reported in a footnote not associated with Table 7, is that these results are *not* based on an analysis of the data we have been examining here (i.e., spanning the years 2003 through 2009), but derive from an analysis of an earlier LAUSD data set with students and teachers spanning the years 2000 through 2004. As it turns out, regressions identical to the ones presented in Buddin's Table 7 could not be conducted with the present data because information about variables such as race/ethnicity, disability status, and free and reduced-price lunch status were not provided. Because these results are impossible for us to replicate, they are less convincing. The use of data from an earlier time period to evaluate the validity of results from a later time period is also questionable.[39] Perhaps most importantly, even if we were able to replicate these findings, they provide for a weak diagnosis of bias that would not contradict the results from the sensitivity analysis presented above.

Validity of the Report's Findings and Conclusions

The biggest problem with Buddin's white paper is that he never asks or answers the hard and important questions about the validity of the teacher effects he presumes to be estimating. The results from our analyses indicate that:

1. There is strong evidence that students are being sorted into elementary school classrooms in Los Angeles as a function of variables not controlled for in the LAVAM.

2. The estimates of teacher effects are sensitive to the specification of the value-added model. When the LAVAM is compared relative to an alternative model (altVAM) with additional sets of variables that attempt to control for (i) a longer history of a student's test performance, (ii) peer influence, and (iii) school-level factors, only between 46% and 61% of teachers maintain the same categorization of effectiveness.

3. The estimates of teacher effects appear to be considerably more biased by the sorting of students and teachers for reading test outcomes than they are for math test outcomes.

4. When a 95% confidence interval is placed around the teacher effects estimated by the LAVAM, between 43% and 52% of teachers cannot be distinguished from a teacher of "average" effectiveness.

Of Buddin's empirical results, we are only able to agree with his finding with regard to the magnitude of a standard deviation for his reading and math teacher effectiveness distributions. This seems to range between about 0.2 student-level test score standard deviations in reading, and about 0.3 in math. In contrast, we disagree with Buddin's finding that traditional teacher qualifications have no "effect" on student outcomes. In the first place, as Buddin himself acknowledges, his study was not designed to address this question in a causal manner—his measures of a teacher's educational background and credential status are crude, and his evidence is purely correlational. But beyond this, his analysis of this issue should be framed with teachers as the units of analysis, not students. When this is done, we find important associations between teacher effectiveness and both teacher experience and educational background that are not trivial.

Usefulness for Policy and Practice

The analyses presented in Buddin's white paper are the principal justification behind the teacher ratings that were published by the *L.A. Times*. Along with these ratings, the *Times* web site includes a section of "frequently asked questions" (FAQ) about value-added analysis (http://projects.latimes.com/value-added/faq/). One of the FAQs is especially relevant given the findings from our sensitivity analysis:

> *Is a teacher's or school's score affected by low-achieving students, English-language learners or other students with challenges?*
>
> Generally not. By comparing each child's results with his or her past performance, value-added largely controls for such differences, leveling the playing field among teachers and schools. Research using L.A. Unified data has found that teachers with a high percentage of

> students who are gifted students or English-language learners have no meaningful advantage or disadvantage under the value-added approach. The same applies to teachers with high numbers of students who are rich or poor.

Our findings do not support the assertion that a teacher's "scores" are unaffected by low-achieving students. And, as we have noted, it is not possible to verify the findings—based on Buddin's analysis of prior data from 2000 to 2004—that there is no "meaningful" relationship between value-added estimates and classroom demographic variables such as gifted and talented status, special needs, ELL status and poverty levels. So while Buddin's analysis has clearly proven itself to be useful from the perspective of the *L.A. Times*, this utility is misleading in that it casts the *Times'* teacher ratings in a far more authoritative and "scientific" light than is merited.

In the research literature on value-added modeling, there is currently healthy debate about whether (and to what extent) the approach will lead to evaluative conclusions comparable to what would be achieved if students and teachers could be randomly assigned to one another.[40] In a value-added model, a teacher is considered an educational treatment or intervention in the same sense as a new reading curriculum, and the goal is to figure out which teacher "works" and which teachers do not. Yet instead of the restricted (though still extremely challenging) task of comparing outcomes for students assigned to a new reading curriculum (treatment) relative to an old one (control), the value-added analyst has the unenviable task of comparing outcomes for students assigned to one teacher (treatment) relative to *hundreds or thousands of other teachers* (controls). There is a great irony here that following a decade in which there has been a great push to increase the scientific rigor of educational evaluations of "what works" through an increasing emphasis on the importance of experimental designs, much of this seems on the verge of being thrown out the window in pursuit of teacher evaluations with non-experimental designs that would not be eligible for review were they to be subject to the standards of the What Works Clearinghouse (http://ies.ed.gov/ncee/wwc/).

To the credit of the *L.A. Times*, teachers have at least been given the opportunity to post responses to their ratings. These largely unfiltered written responses can be found alongside a given teacher's rating and as a collection in chronological order at http://projects.latimes.com/value-added/responses/page/1/. As one might expect, many of the responses are emotionally charged, especially those that immediately followed the publication of the ratings. The teachers who chose to respond in writing are unlikely to constitute a representative sample of LAUSD teachers. Nonetheless, it is interesting to peruse their comments, because in many cases the teachers are able to anticipate—even without necessarily understanding the details of the value-added model used to rate them—plausible reasons why they may have been rated unfairly: the failure to take into account a change in school administration, student attendance rates, team-teaching practices, etc. Some of the responses are both thoughtful and prescient, as this example from a teacher named Daniel Taylor illustrates:

> The years that I taught 3rd grade were for Bilingual Waiver classes. Many of the students in those classes were taught primarily in Spanish, along with instruction in ESL. The students had various degrees of English language proficiency, and since Spanish was the dominant

> language for all of those students, the expectations were not that high for the CST scores for those students, since standardized tests were largely discounted as an assessment at that time. Also, it was expected that these bilingual students would perform much better on the Spanish counterpart of the CST -- the "Aprenda" test. These scores were not reported in the LA Times database, nor were these tests even mentioned in the LA Times report… I'm not sure of the value of publicly labeling teachers as less or least effective at raising test scores, since the parents don't generally get to choose the teacher, any more than the teachers get to choose which students will be in their classroom. It is also worth noting that the gifted students and the ones with serious behavior problems have not been evenly distributed. This kind of public rating will most likely serve to reinforce those kinds of placements. And yes, these types of students (especially the latter) do affect the learning environment and performance of classrooms as a whole. I'd bet that there are many "less effective" teachers who have seen their students make significant progress in writing and other areas that aren't necessarily measured on the CST tests. But the message we're getting from the LA Times public rating scheme is that these test scores are paramount. Teachers might feel compelled to do whatever it takes, by any means necessary, to get on the upper half of that Value-added Normal Curve. But the normal distribution of scores requires that half of the teachers fall below the 50th percentile or Statistical Mean Average. This means that when the Value-added Ratings of some teachers go up, it follows that others will be going down.

At the end of his white paper, Buddin concludes that value-added models should be used to evaluate teachers because "these measures would provide useful feedback for teachers on their performance and for administrators in comparing teacher effectiveness." Notwithstanding that his analysis provides no evidence in support of this assertion, the teacher response above touches on the interesting question of how the "feedback" from a value-added model would be expected to lead to long-term, system-wide improvements in teacher quality—and how one would know if this had happened. As Mr. Taylor correctly appreciates, value-added models provide purely normative information in the way that teachers are compared with one another. So long as there is variability in value-added estimates, there will always be a distribution of "effective" and "ineffective" teachers. Is the idea that as teacher quality rises over time, the variability in value-added will decrease to the point that it becomes practically insignificant? And what is to happen with "low-quality" teachers? Are they to be given intensive training or will they be replaced? If it is the former, what is the training that would be implemented, and who would pay for it? If the latter, where is the reservoir of high-quality teachers waiting in the wings? And how are the majority of teachers in grades or subjects for which there are no preexisting standardized tests to be evaluated?

There is a danger that debates around the use of test scores to evaluate schools and teachers will involve two extreme, but mistaken, positions. First it is argued that unless value-added models can be shown to lead to perfect classifications of effective and ineffective teachers (however defined), they should not be incorporated into the high-stakes decisions likely to accompany teacher (and school) evaluations. This argument is a false one because, as a number of researchers have pointed out, relative to a status quo for teacher evaluations that most people regard as unacceptable, classifications driven at least in part by value-added estimates do not need to be perfect in order to constitute an improvement.[41]

The second mistaken position is that any critique of the inferences being ascribed to a value-added model constitutes an endorsement of the status quo. One may well agree that test scores need to be given a normative context before they can be interpreted for evaluative purposes, but not agree on what that context should be, or whether a particular value-added model has produced it. The use of standardized test scores to evaluate teachers involves making difficult choices in which there are invariably some tradeoffs between decisions that might be optimal from the perspective of estimating an unbiased causal effect, but not optimal from the perspective of crafting an educational accountability policy with a coherent theory of action. The obligation for those with expertise in statistical and econometric methods is to be explicit and transparent about these choices, so that policymakers and administrators have the information they need to weigh the costs and benefits, and so that all stakeholders have an entry point to the policy debate.

The Buddin white paper presents a picture that implies a "have your cake and eat it too" scenario: that from a technical standpoint we know how to validly isolate the causal effect of a teacher, and from a policy standpoint we know how to create an incentive structure that winnows away the ineffective teachers while rewarding the effective ones enough to encourage new ones to enter the field. This picture is an illusion. Causal inference may well be the holy grail of quantitative research in the social sciences, but it should not be proclaimed lightly. When the causal language of teacher "effects" or "effectiveness" is casually applied to the estimates from a value-added model simply because it conditions on a prior year test score, it trivializes the entire enterprise. And instead of promoting discussion among parents, teachers and school administrators about what students are and are not learning in their classrooms, it seems much more likely to shut them down.

Appendix: Technical Details

Building R Dataframes from LAUSD Data Files

On August 25, 2010 we submitted a formal request to the LAUSD, under the auspices of California's Public Records Act, for the same data that had been used by Richard Buddin for his value-added analysis of Los Angeles teachers. Though this request was granted as of September 8, 2010, we did not actually receive the data, until October 26, 2010. The datafile we received was 143GB in size and contains seven directories/folders named "attendance" (data on school attendance), "CELDT" (data on students taking English Language Proficiency Assessments), "CST" (data on students taking the California Standardized Tests), "Elementary Marks" (data linking elementary grade students to teachers and schools), "PAIF" (data on characteristics of teachers), "Program" (data on students eligible for Title 1 services) and "Secondary Marks" (data linking elementary grade students to teachers and schools). Each directory contained text files with raw data ordered chronologically from the 2002-03 school year to the 2008-09 school year.

In replicating Buddin's first stage student-level fixed effect regression we primarily made use of the CST and Elementary Marks files to build a dataframe containing vectors for year by grade test scores (current and lagged) in reading and math. As noted in our report, there were up to six cohorts of students that could be used as a basis for estimating a teacher's value-added. A student was excluded from a cohort if he or she was missing a test score in either the current or prior year, if the student's ID was duplicated in the data file, or if the student was missing a teacher or school identifier. A teacher (and by extension, that teacher's students) was excluded from a specific cohort if

- the teacher was in a school with 100 or fewer test-takers, or
- a classroom with 15 or fewer students.

We extracted our "control" variables from the following sources:

Variable Name	Source	% in data used for math outcome analysis over grades and years (N = 737,403)
Female	Taken directly from existing variable "GENDER_CODE" in the "Elementary Marks" data files	49.63%
ELL	Derived from "CELDT" data files. If a student was part of this file in any given grade/year combination the variable ELL took on a value of 1.	48.86%
Title 1	Derived from "Program" data files. If a student was part of this file in any given grade/year combination the variable Title1 took on a value of 1.	89.34%
JoinAfterK	Taken directly from the existing variable "STD_GRADE_FIRST_ENROLL" in the Elementary Marks" data files.	28.95%

A couple of things are worth noting here. First, we are unable to compare the descriptive statistics for the variables above with those for the same variables used in Buddin's analysis because he does not report them. Second, there are two student-level variables and one classroom-level variable that have typically been used in previous large-scale value-added modeling applications that were available in the LAUSD data: student attendance rates, parent educational levels, and class size. In fact, in Buddin & Zammaro's 2009 paper class size is found to have a significant effect on test score gains. Buddin does not explain why these were excluded from the LAVAM specification.

In replicating Buddin's second stage teacher-level regressions, we extracted teacher-level variables from the PAIF datafiles, and imposed the additional restriction that only teachers with at least 30 students cumulatively would be included.

Fixed Effects Regression Using Buddin's LAVAM Specification

Using Buddin's notation, the LAVAM model takes the form (see his Equation 3, p. 5)

$$T_{it} = T_{it-1}\lambda + x_{it}\beta + u_i\eta + \psi_j + \varepsilon_{it}$$

where

- the subscripts i, j and t represent students, teachers and years respectively,
- T_{it} and T_{it-1} represent current and lagged test scores,
- x_{it} represents a vector of time varying student characteristics,
- u_i represents a vector of time invariant student characteristics,
- ψ_j represents a teacher effect, and
- ε_{it} represents a composite of "individual and teacher time variant unobserved characteristics."

No time varying student characteristics are actually included in the model, but there are dummy variables for the grade and year (without interactions) included in x_{it}. The time invariant student characteristics u_i include the control variables described in the previous section.

Instrumental Variables

A novel aspect of Buddin's analysis is his use of an instrumental variable (IV) approach to correct for the attenuation in his regression coefficients due to measurement error in the lagged test score variable T_{it-1}. He accomplishes this by using reading test scores as instruments for math test scores and vice-versa. This is a plausible approach so long as (a) math and reading test scores are strongly correlated, and (b) the measurement error in math scores is uncorrelated with the measurement error in reading scores. The first condition can be tested empirically, and we find that the correlation is in fact quite strong—about 0.8. The second condition cannot be tested empirically; it hinges on the assumption that whatever unknown factor(s) cause a student to score higher or lower than his or her true score, these factors change at random from one testing occasion to another. We are not so sure this assumption is plausible, especially when the time span between taking the CST in math and reading is very short. If, for example, a student

scores poorly because of a family emergency in the preceding week, this source of "measurement error" would be likely to have an impact on both reading and math performance in the same way. Nonetheless, we implemented the IV approach in the same way described by Buddin, replacing, for example, a lagged math test score with the predicted value of the regression of this score on the lagged reading test score and student-level control variables. Doing so significantly increases the magnitude of the estimated regression coefficient on the lagged test score: from 0.74 for reading and 0.70 for math to 0.83 for both.

Estimation of Fixed Effects

A complication in the estimation of the teacher fixed effects in the LAVAM is the high-dimensional nature of the matrix that needs to be inverted (e.g., there are more than 10,000

Table A-1. Comparison of Fixed Effect Regression Parameter Estimates

	Reading		Math	
	Buddin	B&D	Buddin	B&D
Lagged Reading	0.876	0.824		
Lagged Math			0.871	0.824
Grade 4	0.013	-0.006	0.006	0.005
Grade 5	0.026	-0.014	0.017	-0.005
Title 1	-0.032	-0.034	-0.053	-0.129
Female	0.027	0.035	0.021	0.023
English Language Learner	-0.027	-0.003	-0.012	-0.026
Joined after Kindergarten	0.026	0.028	0.02	0.028
Test year 2005	0.014	-0.005	0.016	-0.001
Test year 2006	0.005	-0.005	0.005	0.001
Test year 2007	0.006	-0.007	0.004	-0.009
Test year 2008	0.007	-0.003	0.007	-0.003
Test year 2009	-0.003	-0.006	0.002	-0.005
Constant	0.018	0.042	0.034	0.133
Teacher Effect SD	0.21	0.268	0.297	0.326
R-Squared	0.685	0.611	0.596	0.584
Student Years	836310	733193	836310	743764
Number of Teachers	11503	10810	11503	10887
Shrunken SD	0.181	0.231	0.268	0.297

Notes: The omitted reference categories are grade 3, not in a Title I school, male, not an ELL, joined LAUSD in kindergarten, and test year 2004. The dependent variables are student ELA and math test scores standardized by grade and year.

teacher dummy variables in the regression). Standard regression estimation procedures such as "lm" in R will not suffice. In his white paper, Buddin gives no indication of how he went about estimating these fixed effects. We followed his reference to his own 2009 paper for insight. In this Buddin references an approach developed by Abowd, Creecy & Kramarz (2002) for the software STATA. We located this working paper and the relevant STATA code, but the approach only works with two cross-classified sets of dummy variables (e.g., student and teacher dummy variables). Because Buddin specifically says that he has removed student fixed effects from the model, it is unclear what approach he took for parameter estimation. Our solution was to implement an approach described by Guimares & Portugal (2009) after rewriting it for R. This code is available upon request.

Empirical Bayes Shrinkage

Buddin is vague about the methods he uses to correct for "measurement error" in his estimates for of teacher effects noting only that "We used Bayesian methods to shrink these estimates and correct for measurement error" (p. 9). In the context of value-added modeling, the most detailed presentation of Empirical Bayes methods for shrinkage can be found in articles by Kane & Staiger (2009) and McCaffrey et al (2009). The key move here is to estimate for each teacher, a "reliability" coefficient between 0 and 1 that gets used to "shrink" a teacher's observed effect back to the district average. The lower the reliability, the more a teacher's effect estimate gets shrunken. However, there are actually two different coefficients that could be used here as McCaffrey et al distinguish between indices for "reliability" and "stability." Somewhat confusingly, what McCaffrey et al call "stability", Kane & Staiger call "reliability". We use a simple approach to estimate a reliability coefficient (applying the McCaffrey et al terminology) for each teacher and use this as our shrinkage factor. (Note that we make no attempt in doing this to get separate estimates for the variance of the "persistent" teacher effect relative to the variance of a "non-persistent" classroom "shock".)

We have some serious reservations about the theoretical underpinnings of this shrinkage approach and suspect that it overestimates the "reliability" of teacher effects. The basic model, given a single longitudinal cohort of students, says that a teacher effect is a linear combination of two independent random variables—"true" teacher quality and "noise" due to the "idiosyncrasies" of the particular cohort of students in a teacher's classroom. There is no a priori reason to think these two terms are uncorrelated. Nor is there any reason to believe that noise at the student level is uncorrelated across students. If it were possible to account for these dependencies, the reliability coefficients for teacher effects would be likely to decrease considerably. In our view this is just the tip of the iceberg when it comes to conceptual and theoretical problems behind the notion of "sampling error" as applied to teacher effect estimates. A full explication and exploration of the iceberg is outside the scope of this report, but will surely be the topic of a subsequent one.

Teacher-level Regressions

Buddin is not clear about whether he uses the shrunken or unshrunken estimates of teacher effects as the outcome variables in his teacher-level regressions. However, Koedel & Betts (2010)

take a similar analytical approach and clearly seem to be using unshrunken estimates of teacher effects as their outcome variable. We followed this lead in our approach. (Just to check, we also ran the regressions using the shrunken estimates and saw that this had a negligible impact on our findings—our coefficient estimates remained larger than those reported by Buddin. This is not surprising because with given six cohorts of students, most teacher effect reliability coefficients are quite high, usually greater than 0.9.) Buddin does note that his regressions are estimated using Feasible Generalized Least Squares following an approach outlined in a footnote by Borjas (1987). We implemented this same approach.

Appendix References

Abowd, J., Creecy, R. & Kramarz, F. (2002). Computing person and firm effects using linked longitudinal employer-employee data. Working paper.

Borjas, G. (1987). Self-selection and the earnings of immigrants. *American Economic Review, 77*(4), 531-553.

Buddin, R. & Zamarro, G. (2009). Teacher qualifications and student achievement in urban elementary schools. *Journal of Urban Economics, 66*, 103-115.

Guimaraes, P. & Portugal, P. (2009). A simple feasible alternative procedure to estimate models with high-dimensional fixed effects. IZA Discussion Paper No. 3935.

Koedel, C., & Betts, J. (2010). Value Added to What? How a Ceiling in the Testing Instrument Influences Value-Added Estimation. *Education Finance and Policy, 5*(1), 54–81.

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.

Notes and References

1 We would like to thank Henry Braun, Gene Glass, Cory Koedel, Bill Mathis, Jesse Rothstein, and Kevin Welner for their helpful feedback on earlier versions of this report. We thank John Rogers and Gary Blasi for both their feedback and initial assistance in getting us access to the LAUSD data that had been used by the *L.A. Times*. We thank Ed Haertel for his help linking LAUSD schools to the California API and SSI database maintained by the California Department of Education.

2 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 February 2, 2011, from
http://www.latimes.com/news/local/la-me-teachers-value-20100815,0,2695044.story

3 For a good discussion of this position, see
Harris, D. N. (2009). Would Accountability Based on Teacher Value Added Be Smart Policy? An Examination of the Statistical Properties and Policy Alternatives. *Education Finance and Policy, 4*(4), 319–350.

4 See, for example, the August 17, 2010, *Education Week* blog entry by Rick Hess:
Hess, R. (2010, Aug. 17)."LAT on Teacher Value-Added: A Disheartening Replay" (blog entry). *Education Week*, edweek.org. Retrieved February 2, 2011, from
http://blogs.edweek.org/edweek/rick_hess_straight_up/2010/08/lat_on_teacher_value-added_a_disheartening_replay.html.

5 Felch, J. & Song, J. (2010, Aug. 16). U.S. schools chief endorses release of teacher data. *Los Angeles Times*. Retrieved February 2, 2011, from
http://articles.latimes.com/2010/aug/16/local/la-me-0817-teachers-react-20100817

6 While Dr. Buddin is an employee of the RAND Corporation, an organization well-known for its contributions to the research literature on value-added modeling, he was hired independently by the *Los Angeles Times* to conduct the analysis. This is notable because RAND typically has a very rigorous internal peer review before it releases an official research report, and Buddin's white paper was not subject to this same process.

7 This paper can be found at http://documents.latimes.com/buddin-white-paper-20100908/.
Note that the version of this paper that was available to us when we began our review is not the same as the version that became available on the *L.A. Times* web site as of December 15, 2010. According to the *L.A. Times*, the "earlier version of the document contained similar findings based on a slightly smaller number of students and teachers used at an earlier stage of the analysis." It is unclear to us why the earlier "stage" of the analysis used a smaller number of students and teachers, given that the underlying source of data from LAUSD has remained constant. In any case, all our subsequent comparisons with Buddin's results are based on the more recent version of the white paper, dated "August 2010."

8 For an example in which the aggregated results from a growth model are very deliberately not given the label of effects, see Damian Betebenner's work on the "Colorado Growth Model" as described in
Briggs, D. C. & Betebenner, D. (2009). *Is Growth in Student Achievement Scale Dependent?* Paper presented at the invited symposium "Measuring and Evaluating Changes in Student Achievement: A Conversation about Technical and Conceptual Issues" at the annual meeting of the National Council for Measurement in Education, San Diego, CA, April 14, 2009.

9 The title of Buddin's white paper—*How Effective are Los Angeles Teachers and Schools?*—implies that this a research question, but it is an overly ambitious one given the nature of the data, and not one that he actually addresses.

10 We have made slight changes to the wording of Buddin's research questions to be consistent with the terminology we use throughout this report. Buddin tends to use a number of terms such as "quality" or "success in the classroom" interchangeably with "value-added" or "effectiveness." The exact wording of his research questions (in regards to teachers) were as follows:

- How much does teacher quality vary from school to school and from teacher to teacher?
- What teacher qualifications or backgrounds are associated with teacher success in the classroom?

11 We use "reading" instead of "English Language Arts" throughout.

12 We have simplified, for ease of conceptual presentation, the equation for Buddin's model by removing indicator variables for grade and year. In the full model, these variables are included to control for grade- or year-specific features of the test over the 2004-2009 time period. When we replicated the model, all these variables were included.

13 Because the California Standardized Tests do not have scale scores that have been vertically linked such that a score in one grade can be compared in an absolute sense to a score in the grade before or the grade after, no information is being lost by standardizing the scores within grade and year in this manner. However, the lack of a developmental (i.e., vertically linked) score scale means that any notion of student growth or "learning gains" is mostly metaphorical. If a student has a math test score of 0.5 in grade 4 and a score of 1.5 in grade 5, the correct interpretation is that relative to her peers, her test performance in grade 5 is much better than her test performance in grade 4. Relative to a student with a grade 4 score of 0.5 and a grade 5 score of -0.5, it seems safe to say that the first student has learned more than the second. But the interpretation becomes problematic when (as an extreme example) none of the students have learned the new material being tested in grade 5. In this case the interpretation of a score of 1.5 relative to a score of -0.5 is much more equivocal. The upshot is that while standardizing test scores makes it easier to interpret regression coefficients, is also makes it harder to evaluate the validity of the outcome a test purports to measure. There are a number of important issues that could be raised along these lines with regard to choice of outcome measures when employing a value-added model (for example, the impact of possible ceiling effects on the CST), but these are outside the scope of this report.

14 Title 1 eligibility is not the same thing as eligibility to receive free or reduced price lunch (FRL) services. While any student eligible for FRL on the basis of household income is also eligible for Title I services, an entire school can qualify for Title I services. When this happens, even students who are not FRL eligible can receive Title I services. As a result the percentage of students eligible for Title I in the LAUSD is extremely high (about 90%), and will always be higher than the proportion of students eligible for FRL services.

15 This is an example of what economists refer to as the problem of *endogeneity*, because the value for the teacher effect, f_j, is itself a function of unobservable variables that exist, in composite form, in the "black box" term.

16 See
Braun, H, Chudowsky, N, & Koenig, J (eds). (2010) *Getting value out of value-added. Report of a Workshop*. Washington, DC: National Research Council, National Academies Press.

Braun, H. I. (2005). *Using student progress to evaluate teachers: A primer on value-added models*. Princeton, NJ: Educational Testing Service. Retrieved February, 27, 2008.

Koedel, C., & Betts, J. R. (2009). *Does student sorting invalidate value-added models of teacher effectiveness? An extended analysis of the Rothstein critique*. Working Paper.

McCaffrey, D. F., Lockwood, J. R, Koretz, & Hamilton, L. (2003). *Evaluating value-added models for teacher accountability*. RAND Research Report prepared for the Carnegie Corporation.

McCaffrey, D. F., Lockwood, J. R., Koretz, D., Louis, T. A., & Hamilton, L. (2004). Models for value-added modeling of teacher effects. *Journal of Educational and Behavioral Statistics, 29*(1), 67.

Rothstein, J. (2009). Student sorting and bias in value-added estimation: Selection on observables and unobservables. *Education Finance and Policy, 4*(4), 537–571.

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

Sanders, W. L., Saxton, A. M., & Horn, S. P. (1997). The Tennessee Value-Added Assessment System: A quantitative outcomes-based approach to educational assessment. In J. Millman (Ed.), *Grading teachers, grading schools: Is student achievement a valid measure?* (pp. 137-162). Thousand Oaks, CA: Corwin Press.

17 See
Koedel, C. & Betts, J. (2010).Value added to what? How a ceiling in the testing instrument influences value-added estimation. *Education Finance and Policy, 5*(1), 54–81. (While this research paper has just been published, it has been widely circulated in working paper form since 2007.)

18 Buddin argues that student fixed effects are statistically insignificant, and thus add little to the model when they are included. However, Koedel & Betts (2009) have argued that statistical significance can be a misleading criterion to use in such contexts because of limited power to detect real differences when they are actually present. Buddin also provides no specific information about how his test of significance was conducted.

19 This term has a long history in labor economics. In an industrial production function, the outputs (e.g., "widgets") being manufactured are viewed as a function of inputs such as worker productivity. In the education context the idea is to replace "widgets" with student achievement and "workers" with teachers.

20 Kane, T. J. & Staiger, D. (2008). *Estimating teacher impacts on student achievement: An experimental evaluation*. NBER working paper. Note that a model virtually identical to the second specification in Kane & Staiger's Los Angeles study is now the basis for the much larger "Measures of Effective Teaching" study being funded by the Gates Foundation in five different urban school districts.

21 The report was produced by the Wisconsin Value-Added Research Center and is available at http://schools.nyc.gov/NR/rdonlyres/A62750A4-B5F5-43C7-B9A3-F2B55CDF8949/87046/TDINYCTechnicalReportFinal072010.pdf

22 The adjustment to these estimates is intended to take into account the number of students that were the basis for a teacher's estimate—sometimes described as "sampling error." This is done by computing each teacher effect as a weighted average between (1) the average effect of all Los Angeles teachers (equal to 0 in this case), and (2) the teacher effect that was estimated in the initial regression. The weighting factor is the presumed reliability of the effect estimate—the proportion of observed variability in teacher effects that is true variability. We describe this in a bit more detail in our appendix. As we also note there, we regard the notion of reliability in the value-added context in which teacher and schools are the units of analysis as a conceptual and theoretical stretch, but to date such objections have not been raised in the value-added literature and they are outside the scope of this report.

23 We were not able to replicate the exact same coefficient estimates as those reported by Buddin for his student-level regressions. A comparison of our estimates with his can be found in Appendix Table A-1.

24 As we emphasize later, these results should not be interpreted in a causal manner. Some readers may also notice that the variance explained in these teacher-level regressions (R^2) is quite small. However, the variance in student achievement that is explained by teachers is also very small. For example, adding teacher indicator variables to the regression model of student achievement only tends to explain an additional 3% of the variance. In contrast, unsurprisingly, a student's prior test score accounts for a very large part of the variance.

25 If anything, these results probably understate the unfairness in ranking less experienced teachers in the same distribution as more experienced teachers because more experienced teachers are also more likely to have had the time to pursue an advanced degree while teaching, something that we also find to have a positive association with estimated effects. In apparent recognition of this, the model used by the Wisconsin Value-Added Research Center for New York City conditions on years of experience before it produces teaching rankings.

26 No rationale is provided for these particular thresholds.

27 Another explanation might be coding error on our part or on Buddin's, or a discrepancy in the data files that were provided to us and to Buddin by the LAUSD, given that the LAUSD data files are massive in size and came with sparse documentation. We performed all our analyses in the R statistical programming environment. Separate scripts (i.e., code) were written to build the relevant longitudinal data files, to replicate Buddin's results, and to conduct the sensitivity analysis. These are described in more detail in the appendix. In an effort to be as transparent as possible about what we have done and any mistakes we may have made, the script files that were used to generate our results are all available upon request.

28 Alternatively, one could also replace each student's current year test score (on the right hand side of the equation) with his or her test score in the previous year, and the previous year test score (on the left hand side of the equation) with his or her test score from two years ago.

29 This differential sorting by test subject makes some intuitive sense if we assume that students are much more likely to be sorted and grouped on the basis of their language skills (especially given the sizeable contingent of English Language Learners in LAUSD schools) than they are by their math skills.

30 California's "School Characteristics Index" (SCI; Technical Design Group of the Advisory Committee for the Public Schools Accountability Act of 1999, 2000) was created in response to a legislative mandate (the California Public School Accountability Act) that required any school ranking system to include a comparison to schools with similar characteristics. Under California law, each school receives a measure of absolute performance (known as the "Academic Performance Index" or API) as well as a measure meant to reflect its academic performance relative to schools facing similar challenges. Technical details regarding the construction of the Similar Schools Rank can be found on the California Department of Education website http://www.cde.ca.gov/ta/ac/ap/documents/tdgreport0400.pdf. In short, the SCI is formed by regressing school-level API scores on school demographic characteristics (e.g., mobility rates), traditional teacher quality measures (e.g., credential status), and operational characteristics (e.g., average class size) and then using the predicted values to form—for each school—a unique comparison group of 100 schools. A school's Similar Schools Rank reflects its observed API decile relative to the API scores of only those 100 schools in its unique comparison group. We include dummy variables for each decile of the Similar Schools Rank such that higher values represent schools that have in some sense overachieved even when compared to other schools facing similar challenges.

31 For importance of longer test score history, see the papers referenced earlier by Sanders, Saxton & Horn (1997) and Rothstein (2009). For importance of peer influence and school-level factors, see the report by McCaffrey et al (2003).

32 As we discuss later, we are not arguing that the altVAM specification is actually "true"—indeed it is very likely that our altVAM specification is also biased. However, it is impossible to quantify bias without having some hypothesis of

a true value of a teacher effect. The altVAM represents one such hypothesis, and this serves as a frame of reference in testing the sensitivity of the LAVAM approach to the exclusion of certain key variables that were available to Buddin.

33 Let $\hat{f}_j^{altVAM}$ and $\hat{f}_j^{LAVAM}$ represent the estimated effects of the same teacher using the altVAM and LAVAM specifications, respectively. The "bias term" we are computing is equal to $\hat{f}_j^{LAVAM} - \hat{f}_j^{altVAM}$. If we assume that the altVAM is "true", then when the bias term takes a positive value for a given teacher it means that the LAVAM has overestimated that teacher's effect on student achievement (in statistical jargon, the estimate is biased upwards). For the converse, if the value of the bias term is negative, it means that the LAVAM has underestimated a teacher's effect. To summarize all this, we take the standard deviation of each bias term across all teachers (row 3 of Table 4), and then we express this as proportion of the standard deviation of the "true" teacher effects under the altVAM to place the magnitude in context. All else being equal, a larger proportion represents a stronger flag of bias.

34 The same argument could be extended to any predictor variable in the model that is outside the control of students, parents and teachers to alter. For example, imagine we have two students with the same math test score in grade 4, but one student is white and the other black. Because black students tend to score lower on the grade 5 math test than white students, the predicted grade 5 test performance for the black student will be lower than the predicted score for the white student. So it would in essence be easier for the black student to have a positive residual, even if both the black and white student get the exact same observed test score on the grade 5 test. This is why Damian Betebenner (personal communication), who has been working directly with a number of states on growth model implementations, argues that it is unethical to include anything other than previous test performance as "control variables" when contextualizing current grade test performance for individual students.

35 Indeed McCaffrey et al (2003) and Ballou, Sanders and Wright (2004) argue that the inclusion of covariates can actually lead to an overcorrection when teachers are in fact more or less effective for students, classrooms or schools with certain characteristics that are being "controlled" statistically.

36 The inferential concept of a standard error, which is required before a confidence interval can be formed, requires a rather heroic leap of faith in the present context. What is the source of "noise" or "chance error" when one estimates the effect of a teacher by aggregating the residuals of the students in that teacher's classroom? The typical answer in the value-added modeling research literature is that there is uncertainty due to the particular cohort of students—"sampling error." Yet what is the chance process whereby each cohort of students has been "sampled"? Obviously, there is none. Any standard errors being estimated are premised on the fiction that the students in a teacher's class are an independent random sample from a "superpopulation" of possible student cohorts. If there are dependencies within these student cohorts—and surely there are among groups of students that play and work together in small groups—then the simple standard errors computed under the assumptions of "frequentist" sampling theory will be too small or too big, depending on the nature of the dependencies in the data. Hence there are some very good reasons to be skeptical of authoritative statements about the precision (or lack thereof) of value-added estimates. Nonetheless, for the purpose of this review we put this issue to the side and play the game by the house rules that have been adopted in the value-added literature.

37 While only intended to be illustrative, the approach we are taking here is likely to actually overstate the number of teachers with estimated effects significantly different from average because of the large number of multiple comparisons involved. Methods for correcting this would typically lead to even wider confidence intervals.

38 Reporter Jason Felch of the *Los Angeles Times* objected to the above precision analysis, as presented in an embargoed version of this report, pointing out that the *Times* had shared the concern and therefore had limited their analysis to only those teachers with 60 or more students. In response, we reanalyzed the data with that restriction. As expected, the number of false positives and false negatives decreased but remained substantial. The following is from the note sent to Mr. Felch:

Now your concern, as I understand it, is that these statistics, which are based on the full sample of nearly 11,000 teachers for whom value-added could be estimated, are potentially misleading because readers might assume that they generalize to the way that the LA Times used the results for its website ratings. You note that only teachers with at least 60 students were included in these ratings. Because doing this reduces the sample of teachers considerably, the results above will overstate the numbers and proportions of false positives and false negatives if they are generalized to the L.A. Times' ratings. Now while I think we used an appropriate frame of reference in what we did and were very explicit about it, I regard this as a reasonable concern. It is also something we can examine empirically, so we did.

When we impose the N>60 restriction this reduces our sample of teachers to 5124 in math and 5047 in reading. Now we can re-write the [statement from the main body of this report] with respect to this sample:

For classifications based on reading outcomes, we find 516 false negative and 612 false positives--10.2% and 12.1% of the total sample of teachers that would be included in ratings released by the L.A. Times. For classifications based on math outcomes, we find 257 false negatives and 454 false positives--5.0% and 8.9% of the total sample of teachers that would be included in ratings released by the L.A. Times.

So in summary, I agree that use of N>60 criterion by the L.A. Times serves to mitigate the issue of false positive and false negatives. But I stand by our statement in the executive summary of the report that "it is likely that there are a significant number of false positives (teachers rated as effective who are really average), and false negatives (teachers rated as ineffective who are really average) in the L.A. Times' rating system." The fundamental point we were making is that the application of a 95% confidence interval to group teachers into three categories will be more conservative than a quintile approach that groups teachers into five. This remains true with or without the N>60 restriction.

39 There are two reasons in particular to be skeptical of these earlier findings. First, the NCLB legislation was implemented in 2003, changing the nature of the stakes attached to the large-scale assessments being administered in the years that followed, and changing the incentives that might underlie the sorting of students to teachers and vice-versa. Second, the value-added model Buddin specified for the 2000-2004 data differs in its functional form relative to the LAVAM being specified for the 2003-2009 data, most notably in that it constrains the coefficient for the prior test score variable to be 1.

40 In addition to the papers already cited by Braun, 2005; Braun et al, 2010; Kane & Staiger, 2009; and Koedel & Betts, 2010, see

Briggs, D. C. & Wiley, E. (2008) Causes and effects. In *The Future of Test-Based Educational Accountability*, K. Ryan & L. Shepard (Eds). New York: Routledge.

Harris, D. N. & Sass, T. R. (2008). Teacher training, teacher quality and student achievement. Unpublished. Tallahassee, FL: Florida State University.

Reardon, S. F. & Raudenbush, S. W. (2009). Assumptions of value-added models for estimating school effects. *Education Finance and Policy, 4*(4), 492–519.

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.

41 See
The Brookings Brown Center Task Group on Teacher Quality (2010). *Evaluating teachers: the important role of value-added*. Washington, DC: The Brookings Institution. Retrieved February 2, 2011, from http://www.brookings.edu/reports/2010/1117_evaluating_teachers.aspx.