WORKING PAPER 25

MATHEMATICA Policy Research

Working PAPER

BY MATTHEW JOHNSON, STEPHEN LIPSCOMB, AND BRIAN GILL

Sensitivity of Teacher Value-Added Estimates to Student and Peer Control Variables

November 2013

Sensitivity of Teacher Value-Added Estimates to Student and Peer Control Variables

Matthew Johnson, Stephen Lipscomb, and Brian Gill

November 2013

ABSTRACT

The validity of value-added models (VAMs) of teacher effectiveness depends on the ability of the measures to isolate teachers' contributions to their students' achievement. Existing VAMs differ in key aspects of their empirical specifications, however, leaving policymakers with little clear guidance on what factors are important to include when constructing a fair model. We examine the sensitivity and precision of teacher value-added estimates obtained under model specifications that differ based on whether they include student-level background characteristics, peer-level background characteristics, and/or a double-lagged achievement score. We also test the sensitivity of teacher VAM estimates to two model variations that the literature has not previously evaluated. First, because the data available in some states or districts may only link students to teachers rather than linking students to specific classrooms, we test whether replacing classroom average peer characteristics with teacher-year level averages affects the VAM estimates. Second, we allow for variation in the relationship between current and lagged achievement scores based on student demographic characteristics. Using data from a northern state and a medium-sized, urban district in that state, we find that teacher estimates are highly correlated across model specifications. Nonetheless, differences in VAM specifications can affect the placement of teachers across performance categories. The lowest correlation we observed—0.909—implies that 26 percent of teachers who are ranked in the bottom quintile under one specification would be ranked above this quintile when a different set of control variables is used. Differences in VAM estimates are often systematically related to student characteristics: teachers in a district that serves a relatively large fraction of poor and minority students receive lower performance estimates when controls for student and peer demographic characteristics are excluded from the VAM.

I. INTRODUCTION

A growing number of school districts and states are using value-added models (VAMs) to measure teacher effectiveness.[1] The VAMs in use today vary in model specifications, although all teacher VAMs seek to facilitate a better understanding of the individual contributions of teachers to the achievement of their students. If the assignment of students to teachers were random, then neither the estimation strategy nor the choice of control variables to include in the model would substantially affect teacher effectiveness estimates (Guarino et al. 2012a). But students are not randomly assigned to teachers within or between schools, and assignment is likely related to a variety of institutional, residential, family, and student characteristics and choices (Clotfelter et al. 2005, 2006). Teacher effectiveness estimates thus may be affected, potentially to a large degree, by which factors are included as controls in the VAM.

Table 1 summarizes features of the model specifications of five VAMs that are currently used for teacher evaluation. As is standard with teacher VAMs, all models control at least for prior test scores. The table lists whether each VAM accounts additionally for student characteristics, peer (classroom average) characteristics, and/or multiple years of prior scores. Collectively, these five VAMs illustrate the extent to which model specifications vary. The SAS EVAAS model accounts for multiple years of prior scores but does not control for student or peer characteristics; the Chicago model controls for student characteristics but not for multiple years of prior scores; the DC IMPACT and Pittsburgh models control for both student and peer characteristics; and the VAM currently used in Florida controls for student characteristics, peer characteristics, and multiple years of prior scores. Within the three categories of variables in Table 1, there is variation across models in terms of which variables they include. For example, the Chicago and Pittsburgh models include controls for both race/ethnicity and free/reduced-priced lunch (FRL) status; DC IMPACT excludes race/ethnicity but includes FRL status; and the Florida model excludes both race/ethnicity and FRL status.

At first glance, it might seem natural that VAMs should include all the types of control variables listed in Table 1. As these variables relate to student achievement and correlate with how students are assigned to teachers, their inclusion could reduce bias in teacher effectiveness estimates. Even if the other variables fully account for selection bias, additional control variables that correlate with student achievement can reduce the variance of the error term in the VAM and thereby increase the precision of the teacher effectiveness estimates.[2]

[1] Some states and districts are also using Student Growth Percentile (SGP) models for teacher evaluations. Conceptually, measuring median student growth differs from measuring teacher value-added, because it does not explicitly attribute growth to the teacher. (VAMs do not necessarily claim to measure achievement growth.) In policy and practice, however, SGPs and VAMs are both used for teacher evaluations, thereby implicitly attributing the resulting estimates to the teacher. An analysis of SGP models is beyond the scope of this paper.

[2] Adding control variables that are highly correlated with teacher assignment to a model with teacher fixed effects can also reduce the precision of the VAM estimates. For example, the addition of student fixed effects can substantially reduce the precision of teacher VAM estimates (McCaffrey et al. 2009).

Table 1. Control Variables Used in Five School-District and Statewide Teacher VAMs[3]

Value-Added Model	Student Characteristics	Peer Characteristics	Multiple Years of Prior Scores
SAS EVAAS	No	No	Yes
Chicago Public Schools	Yes	No	No
DC IMPACT	Yes	Yes	No
Pittsburgh Public Schools	Yes	Yes	No
Florida	Yes	Yes	Yes

However, there can be trade-offs to including each set of variables, complicating decisions about their inclusion for researchers and policymakers. For instance, some researchers and policymakers believe that controlling for socioeconomic and demographic factors implicitly reduces the expectations for performance from poor and minority students and therefore VAMs should exclude these factors (Sanders et al. 2009).[4] A similar argument can be made about the inclusion of student peer characteristics. The usefulness of peer characteristics can also be influenced by whether sufficient variation exists to identify the coefficients on the classroom average variables separately from the teacher effects. If a model with teacher fixed effects includes only one year of data for homeroom elementary school teachers (in which teachers teach only one classroom per year), then classroom average characteristics would be collinear with teacher effects and cannot be included. To identify coefficients for classroom average characteristics in a model with teacher fixed effects, the data must include within-teacher variation, either through multiple classes or multiple years of teaching for each teacher. If within-teacher variation in classroom average characteristics is insufficient to be used in a fixed effects model, the model can still include classroom averages when it treats teacher effects as random. However, if there is systematic sorting of students to teachers based on student characteristics, a random effects model can lead to biased teacher effect estimates (Guarino et al. 2012b).

The decision to include controls for peer characteristics can also be affected by the availability of the data necessary to calculate these variables. Some state and district data systems can match students to teachers but cannot track each student's classroom. In such cases, only teacher-level averages of the peer variables can be included. This approach may be undesirable for two reasons: (1) There is less within-teacher variation when the peer characteristics are averaged over all of a teacher's students, and (2) as peer effects arise through students interacting with each other in the same classroom, teacher-level averages will be noisy proxies for the classroom average variables.

---

[3] Sources:
SAS EVAAS: http://www.sas.com/resources/asset/Response_to_Criticisms_of_SAS_EVAAS_11-13-09.pdf
Chicago Public Schools: http://www.cps.edu/Pages/valueadded.aspx
DC IMPACT: http://www.dc.gov/DCPS/Files/downloads/In-the-Classroom/IMPACT%20Guidebooks/Value-Added_Guidebook_2012-13.pdf
Pittsburgh: http://www.mathematica-mpr.com/publications/pdfs/education/value-added_pittsburgh.pdf. Note: Pittsburgh VAMs do not control for multiple years of prior year scores at the elementary and middle school levels but control for 8th-grade scores in addition to prior-year scores in high school VAMs.
Florida: www.fldoe.org/committees/pdf/PresentationValue-addedModel.pdf

[4] We are unaware of any evidence that expectations or achievement change for poor or minority students in districts that include controls for poverty and race in their VAMs. However, this is a tradeoff that states and districts consider.

Whether to include more than one year of lagged achievement scores also involves a trade-off. The inclusion of additional prior scores may reduce bias in the teacher VAM estimates and increase their precision by reducing the variance of the error term in the model. However, in most states, standardized tests are not given until 3rd grade. Therefore, including multiple years of baseline scores for students in a VAM would either eliminate the possibility of estimating value-added for an entire grade of teachers or necessitate the use of different VAMs for different grade levels, because only one year of prior test scores will be available to evaluate 4th-grade teachers. In addition, at all grade levels, some students will be missing the additional prior year of scores if they were absent on the testing day that year or if they transferred into the district/state during the previous year and the district was unable to obtain their previous test records. Researchers must either drop these students or impute the students' missing test scores. Imputation may be an undesirable option, however; students with missing scores are from a selected sample that comprises many transfer students with unobservable characteristics that may differ from those of students with non-missing prior scores.

The goal of this paper is to inform policymakers, researchers, and decision makers by examining how certain choices impact estimated teacher effects. We use data from a northern state and from a medium-sized urban district within that state to estimate teacher-level VAMs. Let us call the district in our analysis "district X". District X differs from the state average in several important demographic variables. For instance, the percentage of African-American students is almost three times the state average and the percentage of students receiving free or reduced-price lunches is nearly double the state average. District X also serves a larger percentage of students in special education programs.

We first examine the sensitivity of the teacher effect estimates to various modeling decisions about the inclusion of student background characteristics, peer characteristics, and multiple years of prior test scores. It is important to perform these sensitivity analyses both at the district and state levels, because policymakers may be interested in how teachers perform both relative to other teachers in the district and relative to other teachers in the state. For example, to the extent that the distribution of student characteristics across teachers within a district may be more homogenous relative to the distribution of student characteristics across teachers statewide, teacher estimates in certain districts may be more sensitive to adding or dropping some student background characteristics in statewide VAMs. Using data at both state and district levels allows us to consider more fully which types of student control variables might matter more for teacher VAMs. That is, the state data include more students, but the data from district X include additional variables that relate to student achievement scores. At the district level, we are able to include control variables for prior-year discipline incidents and gifted program participation that are unavailable at the state level.

We next examine whether including a more precise measure of peer average variables affects teacher VAM estimates. The data from district X provide reliable information on classroom identifiers, whereas the state data do not. We therefore can include classroom average peer characteristics in the district VAMs but only teacher-year level average peer characteristics in the state VAMs. We use this additional data to examine the sensitivity of teacher VAM estimates to the inclusion of more precise peer average variables.

Finally, we estimate models that control for student demographic characteristics in a more flexible form than is used in the rest of the VAM literature. When controlling for a variable such as FRL status, most researchers include an indicator variable that equals one if the student is receiving a free or reduced-price lunch. Adding an indicator variable to the model allows for shifts in the average predicted achievement of FRL students relative to non-FRL students, conditional on prior-year scores and other control variables. However, it is also possible that the relationship between

current and prior-year scores differs for students based on their demographic characteristics. We allow for this possibility by interacting demographic characteristic indicators with prior-year score variables. We examine the extent to which controlling for student background characteristics in this manner can affect teacher VAM estimates.

Our analyses suggest that teacher estimates are, in general, highly correlated across model specifications. The correlations we observe in the state and district data range from 0.90 to 0.99 relative to our baseline specification, which includes one year of prior scores along with student and peer background characteristic controls. The lowest correlation (0.909) is obtained by comparing the baseline model with a VAM that includes two years of lagged scores and no student or peer background characteristics. When comparing the rankings of teachers across these two models, we find that 26 percent of teachers who are ranked in the bottom quintile under one specification would be ranked above this quintile under the alternative specification. Few teachers shift by more than one quintile.

Although value-added estimates appear to be relatively stable statewide, they can shift the rankings of teachers in district X by meaningful amounts even under model variants that correlate highly with the baseline VAM in the aggregate. For instance, the statewide rank of the median teacher in district X falls four percentiles in 8th-grade math and nine percentiles in 8th-grade reading when student and peer characteristics are excluded, even though the correlation coefficients for those models statewide are 0.964 and 0.979, respectively. The ranks of district teachers in the tails of the distribution (15^{th} and 85^{th} percentile) decrease as well, though to a lesser extent. The statewide ranks of 5th-grade math and reading teachers in district X also decline when this comparison is performed, though by a smaller amount for 8th-grade teachers. These changes likely occur because students in district X are relatively disadvantaged in terms of the student and peer characteristic controls. Our findings thus indicate that choices about which VAM to adopt at the state level can impact the VAM estimates for a district that differs substantially from the state in terms of the student population it serves.

The precision of the value-added estimates is relatively similar across all the VAM specifications we examine. Using data from district X that contain reliable classroom identifiers, we find that teacher VAM estimates are more sensitive to the use of classroom average student characteristic variables in place of teacher-year level average student characteristics than they are to the inclusion of additional control variables available at the district level but not at the state level. When we allow for the relationship between current and prior test scores to differ based on student demographic characteristics, we find that there is a significant difference for non-FRL and FRL students. However, the difference is not large enough to have meaningful impacts on the teacher VAM estimates.

II. PREVIOUS LITERATURE

Multiple previous studies have explored which models and estimation strategies are preferable when using VAMs to measure teacher effectiveness (McCaffrey et al. 2004; Rothstein 2010; Koedel and Betts 2011; Guarino et al. 2012a; Ballou et al. 2012; Ehlert et al. 2012). However, the literature has not produced a consensus, so we chose a baseline model and estimation strategy that has been shown to perform well in the literature and would be appealing for school districts or states to use to evaluate teachers. Guarino et al. (2012a) show that a model that treats teacher effects as fixed and includes prior scores as control variables (rather than a gain score model) is the most robust model when used on simulated data under a variety of assumptions about the student assignment process.

Ballou et al. (2012) also show that specifications that treat teacher effects as fixed rather than random perform better in the presence of omitted variable bias. We therefore use a model that treats teacher effects as fixed and use prior-year scores as control variables.[5] We take the extra step of controlling for measurement error in the prior test scores with an errors-in-variables strategy that makes use of published test reliability coefficients (Buonaccorsi 2010). Our models do not include student fixed effects, which require large amounts of data and substantially diminish the precision of the teacher effect estimates, thus also diminishing their appeal for districts and states (McCaffrey et al. 2009). Nor do our specifications include school fixed effects, which would compare teachers to other teachers in the same school rather than to other teachers in the district or state.

Many researchers who use VAMs to estimate teacher effectiveness have examined the sensitivity of the estimates to including controls for student characteristics, multiple years of prior scores, and peer characteristics. Ballou et al. (2004) examine the sensitivity of teacher estimates to the inclusion of student demographic controls and find that the teacher estimates are highly correlated with estimates from their baseline model that excludes these factors. Ballou et al. (2012) examine the effect of including variables that are traditionally omitted from VAMs on teacher effectiveness estimates. These variables include student, school, and neighborhood characteristics that are not usually contained in district- or state-level data. They find that the value-added estimates are sensitive to the exclusion of these additional covariates, but the importance of these often-omitted covariates decreases slightly when the data include multiple years of prior test scores.

As a robustness check to their main VAM specification, Chetty et al. (2011) examine the sensitivity of their effectiveness estimates to the exclusion of student and classroom average control variables and to the inclusion of an additional year of prior scores. Aaronson et al. (2007) use a sample of 9th-grade students in Chicago and estimate multiple VAM specifications.[6] They examine the sensitivity of teacher effectiveness estimates to including student characteristics, peer characteristics, and multiple years of prior scores both in terms of the correlation of the teacher effect estimates and the standard deviation of the estimates. Both papers find that teacher estimates are more sensitive to the exclusion of student and peer characteristics than to the inclusion of additional years of prior scores, though they find that the estimates are highly correlated across model specifications.

Researchers have also examined the sensitivity of teacher VAM estimates to the inclusion of peer characteristics. Ballou et al. (2004) include peer average controls for FRL and find that teacher effect estimates are very sensitive to the inclusion of this variable. However their data do not allow them to identify individual classrooms, so they include the school-by-grade average fraction of FRL students as a proxy variable. Ballou et al. (2012) also include peer characteristics, but include teacher-level averages of these variables rather than classroom averages. To the extent that teacher-level averages are a noisy proxy for classroom averages, controlling for them may have a relatively small impact on teacher VAM estimates. Burke and Sass (2008) find that peer effects are stronger at the

[5] Not all existing VAMs treat teacher effects as fixed, however. Two of the five VAMs described in Table 1 treat teacher effects as random (the Florida and SAS EVAAS models). Analyzing the difference between fixed-effects and random-effects teacher VAMs is beyond the scope of this paper.

[6] Most papers estimating VAMs (including our paper) use student test scores from grades 3 through 8, because these grades are the most commonly tested grades in states and districts nationwide.

classroom level than at the grade level, showing that it is important to have reliable classroom information when examining the peer effects in the context of VAMs.

When exploring the sensitivity of model choice to various permutations of control variables, Lockwood et al. (2007) report average correlations ranging between 0.92 and 0.98. Papay (2011) also examines the sensitivity of teacher VAM estimates to the choice of control variables. Among models excluding school fixed effects, Papay reports correlations ranging between 0.88 and 0.99 when including or excluding student and classroom-level control variables. Both Lockwood et al (2007) and Papay (2011) also examine the sensitivity of teacher VAM estimates to the choice of outcome assessment used in the model. As we discuss in detail in the conclusion to this paper, these authors find that teacher VAM estimates are much more sensitive to the choice of outcome assessment than to the choice of control variables.

Briggs and Domingue (2011) use data from the Los Angeles Unified School District and compare teacher effectiveness estimates between VAMs that include and exclude multiple years of prior scores, peer characteristics, and school characteristics. They find lower correlations between teacher effectiveness than those reported in the rest of the literature: 0.92 in math and 0.76 in reading. Goldhaber et al. (2012) use North Carolina state data to examine differences in teacher effectiveness estimates across multiple model types and sets of control variables. The authors find that the effectiveness estimates are in general highly correlated across VAM specifications that omit student and school fixed effects. They report correlations ranging from 0.96–0.99 in math and 0.91–0.99 in reading for models including/excluding student characteristics and classroom characteristics.

Although many researchers have examined the extent to which adding control variables to teacher VAMs can affect the bias of teacher effect estimates, fewer researchers have directly examined the extent to which control variables can affect the *precision* of the estimates. McCaffrey et al. (2009) shows that the precision of teacher effect estimates is substantially diminished when student fixed effects are used in place of student control variables. Florida Department of Education (2011) shows that adding test scores from two prior years as control variables in teacher VAMs results in teacher effect estimates with lower standard errors.

Our paper adds to the growing literature in a number of ways. Most previous papers have used only district- or state-level data. We use data at both levels to determine whether student-level control variables might matter less in a district in which student characteristics are distributed more homogenously than in the state containing that district. That is, our analysis provides insights to school districts about how the estimated effectiveness of their teachers can change in statewide VAMs constructed with different model specifications. We also examine the impact of including additional control variables available at the district level on teacher VAM estimates. In addition, we look at whether controlling for classroom-level average student characteristics rather than teacher-level average student characteristics can change teacher VAM estimates. Finally, we explore the extent to which allowing the relationship between current and prior test scores to vary among student subgroups can affect teacher VAM estimates.

III. MODEL

Equation (1) describes the baseline VAM that we use in this paper:

$$Y_{i,t} = Y_{i,t-1}\lambda + X_{i,t}\beta + \bar{X}_{i,t}\gamma + D_{i,t}\delta + T_t\tau + e_{i,t} \quad (1)$$

The variable $Y_{i,t}$ represents the math or reading test score of student *i* in year *t*. We include the prior year scores in math and reading as control variables indicated by the vector $Y_{i,t-1}$. $X_{i,t}$ is a vector of student background characteristics and $\bar{X}_{i,t}$ is a vector of teacher-level average variables (in state models) or classroom-level average variables (in district models). $D_{i,t}$ is a set of teacher fixed effects, T_t is a vector of school year indicator variables, and $e_{i,t}$ is an error term.[7]

The student characteristics included in $X_{i,t}$ are listed in the second and third columns of Table 2. More student variables are available in the district VAMs, including prior-year attendance and suspension data and gifted program participation.[8] In the third and fourth columns of Table 2, we show the set of peer characteristics in the $\bar{X}_{i,t}$ vector, which include teacher-year level averages (in state models) or classroom-level averages (in district models). In addition to peer average demographic characteristics, we add the average prior achievement of students in the same subject to account for the fact that classes with higher achieving students might provide a more constructive learning environment. We include the standard deviation of prior achievement to allow for the possibility that classes with a large dispersion of achievement might be difficult to teach, because the teacher might have to target lesson plans toward the average student in the class and might not be as effective at increasing the test scores of students in the tails of the prior achievement distribution. Using district data, we can include more peer average variables due to increased availability. In the district VAMs we also add the number of students in the classroom to capture the fact that larger classes are more difficult to teach.[9] These variables represent the peer characteristics as experienced by student *i*, so that the averages are taken over all other students except student *i*.[10] If a student transfers between schools or teachers during the school year, then that student's peer average variables are a weighted average of the peer characteristics experienced by the student with each teacher. Peer characteristics are calculated at the subject level to allow for the fact that if a student was enrolled in multiple math classes during a year, then peers from each class could have affected his or her achievement.

The VAM estimate for each teacher is an average across three years of teaching from the 2008–09 school year through the 2010–11 school year. All models are run separately by subject.

[7] The constant term is omitted from Equation (1) so that all teacher fixed effects can be included in the regression and the coefficients on the teacher indicators can be interpreted as the difference in effectiveness of each teacher relative to the average teacher in the state or district.

[8] Harris and Anderson (2012) show that controlling for whether students are taking advanced-track courses and can have important impacts on teacher VAM estimates. While our data do not allow us to directly control course track, the inclusion of the indicator for gifted program participation will likely account for some of the effects of tracking.

[9] Reliable classroom identifiers are necessary to calculate the number of students in a classroom, so state models cannot include this variable.

[10] The standard deviation of lagged achievement and the number of students in the classroom include all students in the classroom in the calculation.

Measurement error in the prior-year test scores can cause attenuation bias in the estimated coefficients on prior scores. Bias induced by measurement error in prior-year test scores can significantly impact teacher effectiveness estimates (Meyer and Dokumaci 2010; Koedel et al. 2012). To control for measurement error in prior-year test scores, we use an errors-in-variables strategy to estimate the VAMs (Buonaccorsi 2010).[11] To implement the errors-in-variables regression, we use reliability coefficients available from the test publisher that are specific to each prior year, grade, and subject. The models include all students with non-missing data, though we produce estimates only for teachers that in a given year teach more than 10 students with non-missing data. We adjust for measurement error in the teacher VAM estimates through an Empirical Bayes (shrinkage) procedure (Morris 1983). This adjustment shrinks teacher estimates with higher standard errors toward the mean estimate and lowers the probability that teachers with small numbers of students will end up in the tails of the estimated effectiveness distribution.[12] We keep the sample of students the same across model specifications throughout the paper to separate the effect of changes in the set of control variables on the teacher estimates from changes in the sample.

Table 2. Student and Class Characteristics in State and District Models

	Student Characteristics (State)	Student Characteristics (District)	Peer Characteristics (State)	Peer Characteristics (District)
Free or Reduced-Price Meals	x	x	x	x
Disability	x	x	x	x
Race/Ethnicity	x	x	x	x
Gender	x	x		
English Language Learner	x	x	x	x
Age/Behind Grade Level	x	x		
Gifted Program Participation		x		x
Lagged Rate of Attendance		x		x
Lagged Fraction of Year Suspended		x		x
Average Prior Achievement in Same Subject			x	x
Standard Deviation of Lagged Achievement			x	x
Number of Students in Classroom				x

[11] Koedel et al. (2012) suggest that the controls for measurement error should account for the fact that test measurement error tends to be greater in the tails of the test score distribution. Relatively few students in our sample are in the tails of the test score distribution, however, so we implement a linear errors-in-variables model for simplicity.

[12] The results in this paper are very similar when the shrinkage adjustment is not applied.

IV. RESULTS

A. How Sensitive Are Teacher Effect Estimates to Alternative Sets of Control Variables?

1. Correlation of Teacher Effect Estimates: State Data

Our main findings suggest that teacher VAM estimates are highly correlated under different VAM specifications. We see these patterns emerge in Table 3, which reports correlation coefficients between effectiveness estimates for 5th- and 8th-grade math and reading teachers in the state under the baseline VAM (that is, by estimating equation (1)) and four alternative VAM specifications. We examine 5th and 8th grades because these grades typically distinguish elementary and middle schools, and often differentiate departmentalized (subject-specific) teachers (in 8th grade) from general elementary-school teachers responsible for multiple subjects (in 5th grade). The alternative specifications, in turn, exclude peer characteristics; exclude student and peer characteristics; add a double-lagged score; or add a double-lagged score and exclude student and peer characteristics. The correlation coefficients are based on post-shrinkage VAM estimates and use the same samples of students and teachers.

Table 3. Correlation of State Teacher VAM Estimates Relative to the Baseline Specification

	Grade 5		Grade 8	
	Math (n=6,491)	Reading (n=6,600)	Math (n=2,778)	Reading (n=3,347)
Exclude Peer Characteristics	0.974	0.982	0.970	0.982
Exclude Student and Peer Characteristics	0.974	0.981	0.964	0.979
Add Score from Year t-2	0.988	0.976	0.977	0.958
Add Score from Year t-2 and Exclude Student/Peer Characteristics	0.968	0.962	0.946	0.946

Note: Findings for each column are based on VAM estimates from 2008–2009 to 2010–2011 obtained from the same sample of students. The baseline specification includes student characteristics, teacher-year level average peer characteristics, and one year of prior test scores.

The correlation of 5th- and 8th-grade teachers' estimates relative to the baseline VAM is between 0.946 and 0.988 across subjects and VAM specifications. The lowest correlations for both math and reading are obtained by comparing the baseline model with a VAM that includes two years of lagged scores and no student or peer background characteristics. Our findings are broadly consistent with evidence presented by Goldhaber et al. (2012) and Chetty et al. (2011) in that the correlations across all specifications are relatively high.

Correlation coefficients provide an indication of the degree of similarity between two sets of VAM estimates, but they do not address the question of how many teachers would change performance categories under alternative VAM specifications, which is ultimately a more policy-relevant statistic. In Table 4, we present a transition matrix to help visualize what a correlation of 0.946, the lowest value in Table 3, implies for the movement of teacher estimates across performance categories. We separate 8th-grade reading teachers into effectiveness quintiles under the baseline VAM and assess how the teachers in each quintile place under the alternative specification that adds a double-lagged score and drops the student and peer characteristics.

Table 4. Percentage of 8th-Grade Reading Teachers in Effectiveness Quintiles Based on State Data for Baseline Model and Model that Includes Scores from Year t-2 but Excludes Student and Peer Average Characteristics

		Include Scores From t-2; Exclude Student and Peer Average Characteristics				
		1st (Lowest)	2nd	3rd	4th	5th (Highest)
Baseline Model	1st (Lowest)	81	17	1	1	0
	2nd	18	57	23	3	0
	3rd	1	23	53	22	1
	4th	0	3	22	59	16
	5th (Highest)	0	0	1	16	83

Note: Findings are based on VAM estimates for 3,347 reading teachers in grade 8 from 2008–2009 to 2010–2011. Correlation with baseline = 0.946.

Table 4 shows that most teachers would receive a VAM estimate in the same quintile under either model specification; in each quintile above fifty percent of teachers would be placed in the same performance category under both specifications. Of the teachers whose VAM estimate moves into another performance quintile, the amount of movement is rarely by more than one category, though 3 percent of teachers in the second quintile under the baseline specification would move to the fourth quintile under the alternative specification. Classification in the bottom quintile may be of greatest interest, because imposing a negative consequence on a misidentified teacher would be a particularly undesirable result. From that perspective, it may be of concern that the alternate model places in a higher quintile 19 percent of the teachers that under the baseline specification would be in the bottom quintile. It is especially noteworthy that that 2 percent of the bottom quintile teachers under the baseline model would be placed in the third or fourth quintile under the alternate specification.

2. Correlation of Teacher Effect Estimates: District Data

We now turn to results based on the district-level VAMs. We have access to data directly from the school district, so we are able to include more student and peer control variables than in the state VAMs. We add controls for participation in the district's gifted program, prior-year attendance rate, and prior-year suspensions (see Table 2). We also add to our baseline district VAM peer-level averages for these three variables, along with a control for class size. District X provided us with reliable data on classroom identifiers, which enables us to use in the set of peer control variables classroom-level averages (rather than teacher-level averages). We can therefore use the district-level VAMs to determine how including additional and more precisely calculated peer controls affects teacher value-added estimates. To increase the sample size of teachers in our analysis, we estimate a series of VAMs that include all upper elementary school teachers (grades 4 and 5) and a series of VAMs that include all middle school teachers (grades 6 through 8). Because the upper elementary teacher VAMs include grade 4, and grade 3 is the earliest-tested grade, we are unable to examine the sensitivity of estimates to the addition of a second year of prior scores.

The main results are displayed in Table 5. They are broadly similar to our state results in that the teacher value-added estimates are highly correlated across model specifications. The correlations are slightly lower than those in the state results, due both to the inclusion of more precise peer average variables and the additional control variables used in the baseline model.

Table 5. Correlation of District Teacher VAM Estimates Relative to the Baseline Specification

	Grades 4–5		Grades 6–8	
	Math (n=173)	Reading (n=188)	Math (n=164)	Reading (n=215)
Exclude Peer Characteristics	0.948	0.973	0.955	0.963
Exclude Student and Peer Characteristics	0.925	0.974	0.918	0.949
Add Score from Year t-2	n/a	n/a	0.987	0.967
Add Score from Year t-2 and Exclude Student/Peer Characteristics	n/a	n/a	0.927	0.909
Exclude Student and Peer Characteristics Unavailable in State Data	0.976	0.991	0.973	0.991

Note: Findings for each column are based on VAM estimates from 2008–2009 to 2010–2011 obtained from the same sample of students. The baseline specification includes student characteristics, classroom average peer characteristics, and one year of prior test scores.

In the last row of Table 5, we explicitly examine the extent to which the additional student and peer control variables available in the district data matter for teacher VAM estimates. We exclude the following variables listed in Table 2 at both the student and peer level that are available in the district data but not in the state data: gifted program participation, lagged suspensions, lagged attendance, and classroom number of students. The additional variables available in the district data matter more for math teacher VAM estimates than for reading teacher VAM estimates. The correlation between teacher VAM estimates from the model with the excluded variables and from the baseline model are 0.976 in upper-elementary math, 0.973 in middle school math, and 0.991 in reading for both grade ranges. These correlations are higher than the correlation of 0.82 reported in Ballou et al. (2012) when the authors examine the sensitivity of VAM estimates to the of control variables not often included in VAMs. The lower correlation reported in Ballou et al. 2012 is likely due to the fact that these authors include a number of additional control variables that we do not.[13]

In reading the largest change in teacher estimates is found in the model that both adds a second prior year of scores and excludes all other student and peer control variables. Table 6 shows the performance category transition matrix between this model and the baseline model. As before, most teachers receive a VAM estimate in the same quintile under either model specification. Of the teachers whose VAM estimate moves into another performance quintile, the movement rarely exceeds one category. The alternate model places in the second quintile 26 percent of teachers who under the baseline specification would be in the bottom quintile.

[13] Additional control variables in the alternate model specified by Ballou et al. (2012) include student days tardy, GPA, and teacher ratings as well as school-level average variables and community-level average variables based on the zip codes of student residences.

Table 6. Percentage of 6th- through 8th-Grade Reading Teachers in Effectiveness Quintiles Based on District Data for Baseline Model and Model that Includes Scores from Year t-2 but Excludes Student and Peer Average Characteristics

		Include Scores From t-2; Exclude Student and Peer Average Characteristics				
		1st (Lowest)	2nd	3rd	4th	5th (Highest)
Baseline Model	1st (Lowest)	74	26	0	0	0
	2nd	21	53	23	2	0
	3rd	5	14	51	30	0
	4th	0	7	21	51	21
	5th (Highest)	0	0	5	16	79

Note: Findings are based on VAM estimates for 215 reading teachers in grades 6–8 from 2008–2009 to 2010–2011 for a medium-sized, urban district. Correlation with baseline = 0.909.

B. How Does the Choice of Control Variables Affect the Rankings of Teachers With Large Fractions of Disadvantaged Students?

We next examine the extent to which the choice of VAM control variables can affect the estimates for teachers who teach a larger fraction of disadvantaged students. This question is important, because even if the estimates across models are similar for most teachers, the estimates for teachers who teach large fractions of low-income or minority students may still experience large changes. We explore the extent to which the ranks of teachers in district X relative to the state shift when models that exclude peer and/or student characteristics are used in place of the baseline specification. District X is an example of one in which the student population differs substantially from the state average: The percentage of African-American students is almost three times higher and the percentage of students receiving free or reduced-price lunches is nearly twice as high. District X also serves a larger percentage of students in special education programs compared to the state average.

Table 7 indicates the extent to which the percentile ranks of teachers in district X change relative to the state when peer average characteristics are excluded and when both student and peer characteristics are excluded from the baseline model. Results are displayed separately by subject for grades 5 and 8. Table 7 shows the state rank of the teacher ranked in the 15th, 50th, and 85th percentiles in the district distribution. Teachers in district X are generally ranked slightly above the median in the state in the grades and subjects we examine; the median teacher in the district is ranked at the 53rd percentile in math and reading in grade 5, at the 62nd percentile in grade 8 math, and at the 66th percentile in grade 8 reading. The rankings of the teachers in district X decline when the model excludes student and peer characteristics, with the largest decrease observed in grade 8 reading. In grade 8 reading, the median district teacher falls nine percentile ranks, dropping from 66 to 57. However, the extent to which controlling for student and peer characteristics matters varies across grades and subjects. In grade 5 reading, there is no decline in the median rank in the model that excludes student and peer variables, though the teacher at the 85th percentile in the district drops from the 83rd to the 79th percentile in the state.

Table 7. Statewide Percentile Ranks Corresponding to District Teachers at the 15th, 50th, and 85th Percentiles of the District Value-Added Distribution, 5th- and 8th-Grade Math and Reading

	Math Grade 5			Reading Grade 5		
	15th	50th	85th	15th	50th	85th
Baseline	21	53	84	16	53	83
Exclude Peer Average Characteristics	18	51	83	15	54	80
Exclude Student and Peer Average Characteristics	18	49	81	16	54	79
	Math Grade 8			Reading Grade 8		
	15th	50th	85th	15th	50th	85th
Baseline	21	62	86	17	66	90
Exclude Peer Average Characteristics	23	59	86	13	59	87
Exclude Student and Peer Average Characteristics	20	58	85	13	57	89

To provide a more complete picture of how the distribution of teachers in district X can change as a result of the exclusion of student and peer control variables, we display the distribution of VAM estimates for teachers in district X under various specifications. Figure 1 shows the distribution of estimates for grade 8 reading teachers in district X under the baseline specification and under the model that excludes both student and peer control variables. Both the middle and left tails of the distribution shift to the left, and the right tail remains relatively stable.

As noted previously, the extent to which teacher estimates in district X are sensitive to the exclusion of student and peer averages varies across grades and subjects. In Figure 2, we highlight one of the distributions that indicated a very small change in ranks for district teachers: the distribution for grade 5 reading teachers.

Figure 1. Distribution of Statewide Value-Added Scores for 8th-Grade Reading Teachers in District X for Baseline Model and Model that Excludes Student and Peer Average Characteristics

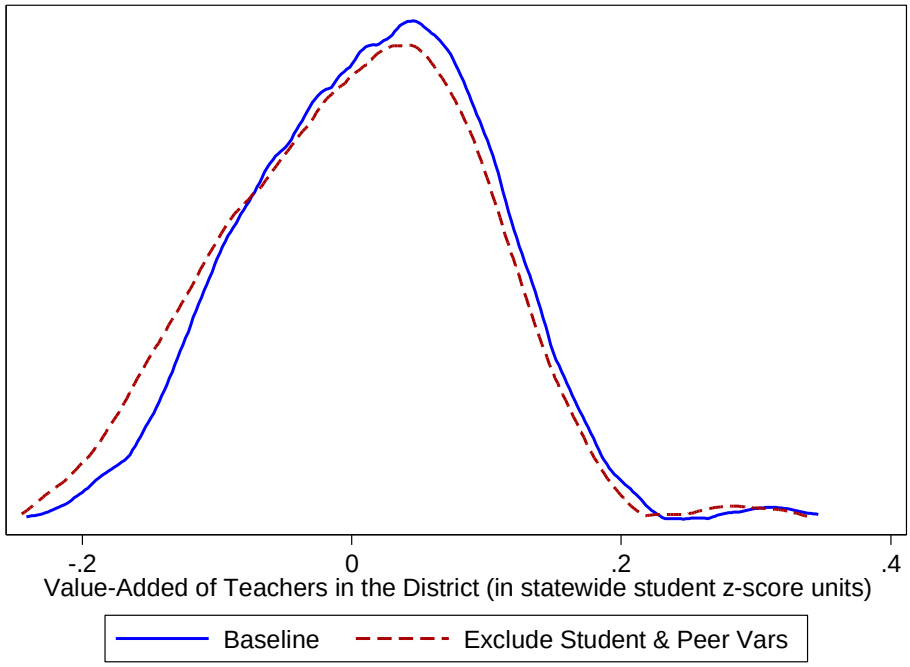
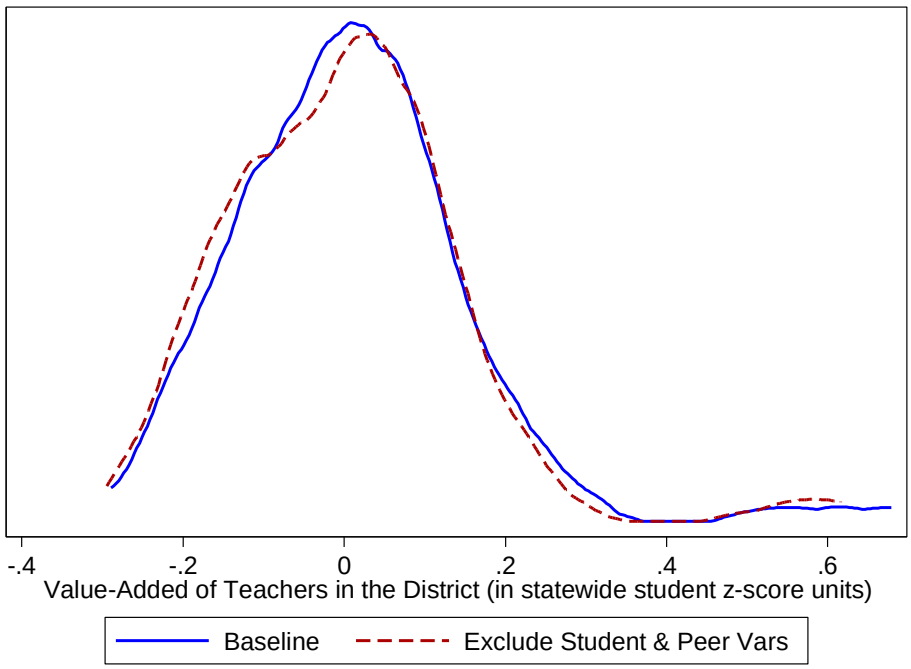


Figure 2. Distribution of Statewide Value-Added Scores for 5th-Grade Reading Teachers in District X for Baseline Model and Model that Excludes Student and Peer Average Characteristics



C. How Sensitive is the Precision of Teacher Effect Estimates to the Choice of VAM Control Variables?

The precision of teacher VAM estimates (as well as bias) may be affected by the decision of which control variables to include in the model. If the choice of control variables has a large influence on the precision of the estimates, then states and districts may want to consider potential changes in precision when deciding on a VAM specification. The effect of adding control variables to VAMs on the precision of the teacher effect estimates is theoretically ambiguous. If the additional variables are highly correlated with the outcome test scores, then the precision of the teacher effect estimates will increase, because the variance of the error term in the VAM will be reduced. On the other hand, if the additional variables are highly correlated with the teacher indicator variables, the precision of the teacher effect estimates could decrease, because there will be less independent variation in the teacher indicator variables that can be used to identify the teacher effect estimates.

We examined the changes in precision of the state teacher VAM estimates across the models analyzed in Table 3 by looking at the average standard error of the teacher VAM estimates and the percentage of estimates that were significantly different from average at the 0.05 level. The results for the grade 8 teacher VAMs are reported in Table 8.[14] The choice of control variables has little impact on the precision of the teacher VAM estimates. The largest changes occur in math but involve changes of only 0.003 in the average standard error and an increase in the number of teachers statistically distinguishable from average by 3.2 percentage points. Moreover, the direction of the change in precision is not consistent in reading and math.

Table 8. Precision of State Teacher VAM Estimates Across VAM Specifications: Grade 8

	Math		Reading	
	Average Standard Error	Percentage Significant	Average Standard Error	Percentage Significant
Baseline Model	0.055	46.4%	0.060	23.0%
Exclude Peer Characteristics	0.055	45.8%	0.061	25.4%
Exclude Student and Peer Characteristics	0.055	45.5%	0.062	25.2%
Add Score from Year t-2	0.053	49.6%	0.058	24.1%
Add Score from Year t-2 and Exclude Student/Peer Characteristics	0.052	48.5%	0.059	24.9%

Note: Findings for each column are based on VAM estimates for grade 8 from 2008–2009 to 2010–2011 obtained from the same sample of students. The baseline specification includes student characteristics, teacher-year level average peer characteristics, and one year of prior test scores. The sample size is 2,778 teachers in math and 3,347 teachers in reading.

We also examined changes in the precision of the district teacher VAM estimates across the models described in Table 5. The precision of the teacher VAM estimates was slightly more sensitive to the choice of control variables in the district data relative to the state data. The results are reported in Table 9. The average standard error decreased by as much as 0.005, and the percentage

[14] Results were similar when we analyzed the precision of the grade 5 teacher VAMs. We omitted these results to conserve space.

of teachers significantly different from average increased by as much as 4 percentage points relative to the baseline specification. However, the district results were broadly consistent with the conclusion that the choice of control variables has little effect on the precision of the teacher VAM estimates.

Table 9. Precision of District Teacher VAM Estimates Across VAM Specifications: Grades 6–8

	Math		Reading	
	Average Standard Error	Percentage Significant	Average Standard Error	Percentage Significant
Baseline Model	0.065	37.2%	0.073	24.7%
Exclude Peer Characteristics	0.063	35.4%	0.071	26.0%
Exclude Student and Peer Characteristics	0.062	35.4%	0.072	28.8%
Add Score from Year t-2	0.063	38.4%	0.070	23.7%
Add Score from Year t-2 and Exclude Student/Peer Characteristics	0.060	38.4%	0.068	28.4%
Exclude Student and Peer Characteristics Unavailable in State Data	0.065	39.0%	0.073	27.4%

Note: Findings for each column are based on VAM estimates for grades 6–8 from 2008–2009 to 2010–2011 obtained from the same sample of students. The baseline specification includes student characteristics, teacher-year level average peer characteristics, and one year of prior test scores. The sample size is 164 teachers in math and 215 teachers in reading.

D. Using Classroom Average Student Characteristics vs. Teacher Average Student Characteristics

To account for the possibility that peer composition can influence student achievement, some VAMs include control variables for peer average characteristics in the model. The most relevant group to account for is peers in the same classrooms and studying the same subjects. It therefore makes sense to include classroom average student characteristics as control variables in the model. However, some districts or states may not have data linking students to classrooms, but rather may have only data linking students to teachers. In these cases, teacher-year level average student characteristics could be substituted in place of classroom average characteristics. If there is relatively little variation across a teacher's classroom in student characteristics, then teacher-year level averages are an adequate proxy for classroom-level averages. However, if there is important variation across a teacher's classrooms in student characteristics, the use of teacher-year level averages may produce noisier and potentially biased teacher effect estimates.

We examined the sensitivity of teacher VAM estimates to the use of teacher-year level averages in place of classroom-level averages. We kept the sample and set of control variables constant and changed the way the variables were calculated (averaged at the classroom level or the teacher-year level).[15] The results are displayed in Table 10. The correlation between VAM estimates across the two versions of the model ranged from 0.933 to 0.975. These correlations are lower than those reported in the last row of Table 5, indicating that the teacher VAM estimates are more sensitive to

[15] We removed class size from the list of covariates, because it does not have an analogous teacher-level interpretation.

the use of classroom average student characteristic variables in place of teacher-year level average student characteristics than they are to the inclusion of additional control variables available at the district level but not at the state level. The VAM estimates based on teacher-year level averages had higher average standard errors than estimates based on classroom-level averages. The difference in average standard errors ranged between 0.002 and 0.009 across models. Even though the average standard errors were higher in the teacher-year average models, the percentage of teachers distinguishable from average was not uniformly lower. In some cases, the estimates changed such that there was an increase in the percentage of teachers significantly different from average in the teacher-year average models. It is possible that the estimates from the model with teacher-year level averages contain more bias, which in some cases could lead to an increase in the percentage that are significantly different from average.

Table 10. Comparison of Results from VAM with Classroom-Level Average and VAM with Teacher-Level Average Student Characteristics

	Grades 4–5		Grades 6–8	
	Math (n=173)	Reading (n=188)	Math (n=164)	Reading (n=215)
Correlation Between Teacher Effect Estimates	0.933	0.945	0.956	0.975
Average Standard Error (Classroom)	0.088	0.095	0.065	0.073
Average Standard Error (Teacher)	0.097	0.098	0.069	0.075
Percentage of Significant Estimates (Classroom)	34.1%	28.2%	39.6%	25.1%
Percentage of Significant Estimates (Teacher)	41.0%	24.5%	39.6%	21.4%

Note: The table compares results from VAMs using the same set of students and covariates. In the classroom version, student average characteristics are calculated at the classroom level, whereas in the teacher version, student average characteristics are calculated at the teacher-year level.

It may seem counterintuitive that the correlations reported in Table 10 are lower at the elementary school level than at the middle school level. In many districts, teachers in elementary grades are homeroom teachers who teach both math and reading. Homeroom teachers may teach only one class of students per year, in which case the teacher-year average student characteristics would be the same as the classroom average student characteristics. In district X, however, only approximately 40% of teachers in upper elementary school grades have VAM estimates for both math and reading. This indicates that most upper-elementary teachers in district X are departmentalized and teach multiple classrooms in a given year. Therefore, teacher-year average and classroom average student characteristics will differ for most upper elementary school teachers in our sample.

E. Does the Relationship Between Current and Prior Test Scores Differ For FRL Students?

Most VAMs that account for student demographic characteristics do so by adding indicator variables that equal one if a student is a member of a certain group. For example, the VAMs considered above that control for student characteristics include an indicator for whether the student is eligible to receive a free or reduced-priced lunch. This variable allows for mean shifts in the predicted current-year scores for FRL students, conditional on the other variables in the model. It might account for otherwise unobserved differences in the home environment of these students that could affect their achievement. The coefficient on FRL status is negative and statistically

significant in our models, indicating that FRL students are expected to score lower than non-FRL students with similar prior test scores.

It is possible that in addition to the difference in the average predicted achievement for FRL and non-FRL students, the relationship between current and prior test scores differs for these students, as well. A relationship that differs would indicate that the estimated VAM coefficient on prior test scores is biased. A bias introduced by not accounting for differences in the relationship between current and prior test scores for FRL students could potentially bias the teacher effect estimates for teachers teaching large fractions of FRL students.

We explored the extent to which the relationship between current and prior test scores differs for FRL students by adding to the VAMs interaction terms between the FRL indicator and the prior test scores. This approach allows us to estimate separate coefficients on the baseline test scores for FRL and non-FRL students. These coefficients are reported in Table 11. We performed this analysis using the state data in order to maximize the sample size of teachers and the precision of the coefficients on the baseline scores. The coefficients differ for non-FRL and FRL students, in some cases by substantial amounts. For example, the second column in the table shows that for the grade 5 reading VAM, the coefficient on baseline reading for non-FRL students is 0.708; whereas the coefficient on baseline reading for FRL students is 0.806. The coefficient on baseline math from the grade 5 reading VAM is 0.129 for non-FRL students and 0.061 for FRL students. All coefficients reported on this table are significantly different between non-FRL and FRL students at the 0.05 level. The results show that the opposite subject baseline score is more predictive of outcome scores for non-FRL students than for FRL students while the same subject baseline score is less predictive. The sums of the two coefficients vary less between non-FRL and FRL students.

Table 11. Coefficient Estimates and Standard Errors for FRL-Baseline Score Interaction Terms

	Grade 5		Grade 8	
	Math (n=6,491)	Reading (n=6,600)	Math (n=2,778)	Reading (n=3,347)
Non-FRL Student Coefficient on Prior Year Math Score (SE)	0.801 (0.002)	0.129 (0.003)	0.847 (0.002)	0.199 (0.003)
FRL Student Coefficient on Prior Year Math Score (SE)	0.823 (0.004)	0.061 (0.004)	0.896 (0.003)	0.154 (0.004)
Non-FRL Student Coefficient on Prior Year Reading Score (SE)	0.127 (0.003)	0.708 (0.003)	0.086 (0.002)	0.641 (0.003)
FRL Student Coefficient on Prior Year Reading Score (SE)	0.060 (0.004)	0.806 (0.004)	0.031 (0.003)	0.728 (0.004)

Note: Findings for each column are based on VAM estimates from 2008–2009 to 2010–2011. The cells report the coefficients and standard errors on baseline math and reading scores separately for non-FRL and FRL students.

We examined the extent to which teacher VAM estimates are sensitive to allowing for different relationships between outcome and baseline scores for non-FRL and FRL students. The results are displayed in Table 12. Neither the teacher effect estimates nor the precision of the estimates were sensitive to these differences. The correlation between teacher effect estimates from the model with FRL-prior score interaction terms and the baseline model was above 0.99 in each grade and subject we examined. The average standard error either remained the same or decreased by 0.001 in the interaction model relative to the baseline model. The percentage of teachers significantly different from average either remained the same or increased slightly in the interaction model. We found

similar results when we interacted baseline test scores with a disability indicator in place of the FRL indicator (results not reported). Even though the relationship between outcome and baseline scores differs based on student demographics, these differences were not large enough to have a meaningful impact on the teacher VAM estimates.

Table 12. Comparison of Results from Baseline VAM and Model with Interactions Between FRL and Prior Test Scores

	Grade 5		Grade 8	
	Math (n=6,491)	Reading (n=6,600)	Math (n=2,778)	Reading (n=3,347)
Correlation With Interaction VAM Estimates	0.999	0.999	0.999	0.998
Average Standard Error (Baseline)	0.078	0.083	0.055	0.060
Average Standard Error (Interaction)	0.077	0.082	0.055	0.060
Percentage of Significant Estimates (Baseline)	42.7%	27.9%	46.4%	23.0%
Percentage of Significant Estimates (Interaction)	42.7%	28.3%	46.6%	23.1%

Note: Findings for each column are based on VAM estimates from 2008–2009 to 2010–2011 obtained from the same sample of students. The baseline specification includes student characteristics, teacher-year level average peer characteristics, and one year of prior test scores. The interaction model adds interactions between an FRL indicator and prior-year test scores.

V. CONCLUSIONS

In this paper, we have examined the sensitivity of teacher effect estimates to several types of changes in the control variables used in VAMs. Using data from a northern state and a medium-sized, urban district in that state, we find that teacher estimates have a high degree of correlation across specifications. The types of VAMs we consider toggle between including and excluding student-level background characteristics, peer-level background characteristics, and one or two years of lagged achievement scores. Overall, we found that the teacher effect estimates were highly correlated across model specifications, with correlations in all cases exceeding 0.9. Yet even correlation coefficients above 0.9 do not preclude substantial amounts of potential misclassification of teachers across performance categories. For instance, we find that 26 percent of teachers rated in the bottom quintile would be placed in higher performance categories under an alternative model that is correlated at 0.909.

It is useful to place our findings about the sensitivity of teacher VAM estimates to the inclusion or exclusion of control variables in the context of other choices about VAMs. Though we did not examine this choice in our paper, of particular importance to teacher VAM estimates is the decision about which assessment to use as the outcome variable in the VAMs. In this paper, we used the state standardized tests in math and reading as outcome variables in all the VAMs we analyzed. Ideally, teacher VAMs would measure the underlying effectiveness of teachers, and the VAM estimates would be invariant to a change in the assessment used as the outcome variable in the VAMs. Researchers have shown that this is not the case, however. Lockwood et al. (2007) tested the sensitivity of teacher VAM estimates to the use of two subscales of the same mathematics achievement test as outcome variables. The highest correlation reported by the authors between teacher VAM estimates based on the two subscales was 0.46. Sass (2008) and Concoran et al. (2011) used two different tests—a high-stakes test and a low-stakes test—as outcome variables in teacher VAMs. Sass (2008) reported a correlation of 0.48 based on two math achievement tests and Concoran et al. (2011) reported a maximum correlation of 0.62 across models and assessments.

Lipscomb et al. (2010) examined English teacher VAMs using various standardized tests and district curriculum-based assessments as outcome measures and found correlations of 0.61 or lower between teacher estimates. Papay (2011) analyzed the sensitivity of teacher VAM estimates to the use of different reading assessments and found correlations of 0.54 or lower across different models and assessments. The low correlations reported by these authors suggest that the choice of outcome assessment has a much larger effect on teacher VAM estimates than the choice of control variables. It is therefore important that assessments used for teacher VAMs accurately measure the content that states and districts want students to be learning and want teachers to be held accountable for.

REFERENCES

Aaronson, D., L. Barrow, and W. Sander. "Teachers and Student Achievement in the Chicago Public High Schools." *Journal of Labor Economics*, vol. 25, no. 1, 2007, pp. 95–135.

Ballou, D., C.G. Mokher, and L. Cavalluzzo. "Using Value-Added Assessment for Personnel Decisions: How Omitted Variables and Model Specification Influence Teachers' Outcomes." 2012, Manuscript.

Ballou, D., W. Sanders, and P. Wright. "Controlling for Student Background in Value-Added Assessments of Teachers." *Journal of Educational and Behavioral Statistics*, vol. 29, no. 1, 2004, pp. 37–65.

Brigs, D. and B. Domingue. "Due Diligence and the Evaluation of Teachers: A Review of the Value-Added Analysis Underlying the Effectiveness Rankings of Los Angeles Unified School District Teachers by the *Los Angeles Times*." National Education Policy Center, 2011.

Buonaccorsi, J. P. *Measurement Error: Models, Methods, and Applications*. Boca Raton, FL: Chapman and Hall/CRC, 2010.

Burke M.A. and T.R. Sass. "Classroom Peer Effects and Student Achievement." *CALDER Working Paper No. 18*, 2008.

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

Clotfelter, C.T., H.F. Ladd, and J. Vigdor. "Who teaches whom? Race and the distribution of novice teachers." *Economics of Education Review*, vol. 24, no. 4, 2005, pp. 377–392.

Clotfelter, C.T., H.F. Ladd, and J. Vigdor. "Teacher-Student Matching and the Assessment of Teacher Effectiveness." *The Journal of Human Resources*, vol. 41, no. 4, 2006, pp. 778–820.

Concoran, S. P., Jennings, J. L., and A. A. Beveridge. "Teacher Effectiveness on High- and Low-Stakes Tests." 2011. Manuscript.

Ehlert, M., C. Koedel, E. Parsons, and M. Podgursky. "Selecting Growth Measures for School and Teacher Evaluations." *CALDER Working Paper No. 80*, 2012.

Florida Department of Education. "Recommendations of the Florida Student Growth Implementation Committee." 2011, Manuscript.

Goldhaber, D., J. Walch, and B. Gabele. "Does the Model Matter? Exploring the Relationship Between Different Student Achievement-Based Teacher Assessments." *Calder Working Paper No. 2012-6*, 2012.

Guarino, C.M., M.D. Reckase, and J.M. Wooldridge. "Can Value-Added Measures of Teacher Performance Be Trusted?" Education Policy Center at Michigan State University Working Paper 31. 2012a.

Guarino, C.M., M. Maxfield, M.D. Reckase, P. Thompson, and J.M. Wooldridge. "An Evaluation of Empirical Bayes' Estimation of Value-Added Teacher Performance Measures." Education Policy Center at Michigan State University Working Paper 31. 2012b.

Harris, D., and A.A. Anderson. "Bias of Public Sector Worker Performance Monitoring: Theory and Empirical Evidence from Middle School Teachers." 2012, Manuscript.

Koedel, C., and J. R. Betts, "Does Student Sorting Invalidate Value-Added Models of Teacher Effectiveness? An Extended Analysis of the Rothstein Critique." *Education Finance and Policy*, vol. 6, no. 1, 2011, pp. 18–42.

Koedel, C., R. Leatherman, and E. Parsons, "Test Measurement Error and Inference from Value-Added Models." 2012, Manuscript.

Lipscomb, S., B. Gill, M. Johnson, and K. Booker. "Estimating Teacher and School Effectiveness in Pittsburgh: Value-Added Modeling and Results." Report to the Pittsburgh Public Schools. Cambridge, MA: Mathematica Policy Research. 2010.

Lockwood, J.R., D.F. McCaffrey, L.S. Hamilton, B. Stecher, V. Le, and J.F. Martinez, "The Sensitivity of Value-Added Teacher Effect Estimates to Different Mathematics Achievement Measures." *Journal of Educational Measurement*, vol. 44, no. 1, 2007 pp. 47–67.

Meyer, R.H., and E. Dokumaci, "Value-added models and the next generation of assessments." Austin, TX: Educational Testing Service, Center for K–12 Assessment and Performance Management, 2010.

McCaffrey, D.F., J.R. Lockwood, D. Koretz, T.A. Louis, and L. Hamilton. "Models for Value-Added Modeling of Teacher Effects." *Journal of Educational and Behavioral Statistics*, vol. 29, no. 1, 2004, pp. 67–101.

McCaffrey, D.F., T.R. Sass, J.R. Lockwood, and K. Mihaly. "The Intertemporal Variability of Teacher Effect Estimates." *Education Finance and Policy,* vol. 4, no. 4, 2009, pp. 572–606.

Morris, C.N. "Parametric Empirical Bayes Inference: Theory and Applications." *Journal of the American Statistical Association*, vol. 78, no. 381, 1983, pp. 47–55.

Papay, J.P. "Different Tests, Different Answers: The Stability of Teacher Value-Added Estimates Across Outcome Measures." *American Educational Research Journal*, vol. 48, no. 1, 2011, pp. 163–193.

Rothstein, J. "Teacher Quality in Educational Production: Tracking, Decay, and Student Achievement." *Quarterly Journal of Economics*, vol. 125, no. 1, 2010, pp. 175–214.

Sanders, W.L., S.P. Wright, J.C. Rivers, and J.G. Leandro. "A Response to Criticisms of SAS® EVAAS." SAS White Paper, 2009.

Sass, T.R. "The Stability of Value-Added Measures of Teacher Quality and Implications for Teacher Compensation Policy." National Center for Analysis of Longitudinal Data in Education Research Brief, no. 4, 2008.

For more information, contact Matthew Johnson, researcher, at mjohnson@mathematica-mpr.com, Stephen Lipscomb, senior researcher, at slipscomb@mathematica-mpr.com, or Brian Gill, senior fellow, at bgill@mathematica-mpr.com.

Authors' Note:

We are grateful to Duncan Chaplin, Kevin Booker, Cory Koedel, Robert Santillano, and to participants at the 2012 Association for Education Finance and Policy conference, the 2012 the Association for Public Policy Analysis and Management conference, and the 2013 Using Student Test Scores to Measure Teacher Performance: The State of the Art in Research and Practice conference for their helpful comments and suggestions. Clare Wolfendale provided excellent programming support. The text reflects the views and analyses of the authors alone and does not necessarily reflect views of Mathematica Policy Research. All errors are the responsibility of the authors.

www.mathematica-mpr.com

Improving public well-being by conducting high-quality, objective research and surveys

PRINCETON, NJ - ANN ARBOR, MI - CAMBRIDGE, MA - CHICAGO, IL - OAKLAND, CA - WASHINGTON, DC

MATHEMATICA
Policy Research