Does the Model Matter? Exploring the Relationship Between Different Student Achievement-Based Teacher Assessments

Dan Goldhaber & Joe Walch
Center for Education Data & Research
University of Washington Bothell

Brian Gabele
Seattle Public Schools

August 10, 2012

The research presented here utilizes confidential data from the North Carolina Education Research Data Center (NCERDC) at Duke University, directed by Clara Muschkin and supported by the Spencer Foundation. The authors wish to acknowledge the North Carolina Department of Public Instruction for its role in collecting this information. We also wish to thank Cory Koedel and James Cowan for helpful comments. The views expressed in this paper do not necessarily reflect those of the institutions to which the authors are affiliated. Any and all errors are the sole responsibility of the authors.

I. Student Achievement as a Teacher Performance Measure

Policymakers appear increasingly inclined to utilize measures of student achievement, often state assessment results, to inform high-stakes teacher personnel decisions. This has been spurred on by the federal government's Teacher Incentive Fund (TIF) and Race to the Top (RttT) grant programs, each of which urge states and localities to tie teacher performance to compensation, renewal, and tenure.[1] There are good reasons for this: current teacher evaluation systems in most school districts appear to be far from rigorous.[2] Those currently utilized in the vast majority of school systems in the country find nearly all teachers to be ranked in the top performance category. A recent study (Weisburg et al., 2009), for instance, showed that more than 99 percent of teachers in districts using binary ratings were rated satisfactory.[3] As Secretary of Education Arne Duncan put it, "Today in our country, 99 percent of our teachers are above average" (Gabriel, 2010).

This "Lake Wobegon Effect" flies in the face of considerable empirical evidence that teachers differ substantially from each other in effectiveness.[4] It also does not reflect the assessment of the teacher workforce by administrators or other teachers (Jacob and Lefgren, 2008; Tucker, 1997; Weisburg et al., 2009). And evaluation systems that fail to recognize the true differences amongst teachers greatly hamper the ability of policymakers to make well-informed decisions about the key education resource over which they have control.

[1] The 2010 application for the TIF states "an applicant must demonstrate, in its application, that it will
[2] For reviews, see Goldhaber (2010) and Toch and Rothman (2008).
[3] There was slightly more spread in evaluations in districts using a broader range of ratings, but it was still about 95 percent of teachers who received one of the top two ratings in these districts.
[4] See, for instance, Aaronson et al. (2007), Goldhaber et al. (1999), and Rivkin et al. (2005).

In theory, human resource policies could, and perhaps should, depend on performance evaluations. But today it is nearly impossible to act on differences between teachers when documented records show them all to be the same. There are multiple avenues through which teacher performance measures could be used to increase student achievement. Over time, teacher performance indicators could conceivably be used for the identification of teacher characteristics that are aligned with effectiveness; this could assist districts' screening and hiring processes (Rockoff and Speroni, 2011). They could help identify those teachers who are struggling and would benefit from professional development or mentoring (Darling-Hammond et al., 2012). And they could be used in higher stakes ways: for instance to financially reward effectiveness under pay for performance systems (Podgursky and Springer, 2007), or for selective retention purposes such as helping determine tenure (Goldhaber and Hansen, 2010b) and layoffs (Boyd et al., 2010; Goldhaber and Theobald, 2011).

But, while conceptually simple, the idea of using student outcomes as a teacher performance measure is complex to implement for a variety of reasons, not least of which is the fact that there is no universally agreed upon statistical methodology for translating student achievement measures into teacher performance. Moreover, it is likely that there may be a tradeoff between accuracy and transparency when it comes to different measures. For example, in the case of teacher performance pay, it is possible that a straightforward, easy to understand performance measure that teachers trust, and are therefore more likely to accept at the bargaining table, will lead to increased teacher effort. On the other hand, one might put more emphasis on accuracy for measures that will be used to help inform retention decisions.

A number of evaluation systems being implemented are designed explicitly to use student growth on standardized tests as a metric for teacher job performance.[5] This is true for some districts with TIF grants (Proctor et al., 2011), and for all states with RttT grants.[6] But while all RttT states are using student growth as an input into teacher evaluations, our own review of plans from first and second round winners of RttT, summarized in **Table 1** below, shows that there is cross-state variation in the models that are to be employed: six states are committed to using a value-added model (VAM), two states are using a student growth model, and three states are using a VAM in conjunction with a growth model.[7]

[Table 1 about here]

The fact that states will be using different methods to translate student test scores into a (component) measure of teacher performance raises the question of whether the choice of model matters. There is only a sparse literature on the extent to which differences in methodology used to translate student test achievement into teacher performance affects the ranking of teachers.[8] In this paper, we extend this literature using statewide data from North Carolina to evaluate different methodologies for translating student achievement results into teacher performance. In particular, we focus on the extent to which there are differences in teacher effect estimates generated from different

[5] We use the terms teacher job performance and teacher effectiveness interchangeably.
[6] For more information on the TIF, see http://cecr.ed.gov/TIFgrantees/. For more information on Race to the Top, see a recent review of state RttT applications (Learning Point Associates 2010), which shows that all states that applied (for the first round of RttT) were consistent with the emphasis of the grant competition by proposing to use student achievement as a "significant" portion of teacher evaluations; however, there was no consensus on how they planned to measure growth.
[7] Note that some states use VAM and SGP measures interchangeably. We discuss the differences between the two measures in Section II.
[8] Several large firms—e.g. SAS, Value Added Research Center at University of Wisconsin, Mathematica, and Battelle for Kids—offer competing, though not necessarily fundamentally different services for this translation process.

modeling approaches and to what extent classroom level characteristics predict these differences.

Our findings are consistent with research that finds models including student background and classroom characteristics are highly correlated with simpler specifications that only include a single-subject lagged test score, while value-added models estimated with school or student fixed effects have a lower correlation. Interestingly, teacher effectiveness estimates based on median student growth percentiles are highly correlated with estimates from VAMs that include only a lagged test score and those that also include lagged scores and student background characteristics, despite the fact that the two methods for estimating teacher effectiveness are, at least conceptually, quite different. However, even when the correlations between job performance estimates generated by different models are quite high, differences in the composition of students in teachers' classrooms can have sizable effects on the differences in their effectiveness estimates.

II. Using Student Test Performance as a Gauge of Teacher Effectiveness

There are two broad categories of models that are used to associate student growth with teacher effectiveness: student growth percentile (SGP) models (also commonly known as the "Colorado Growth Model") and value-added models (VAMs). The primary distinctions between the two approaches are the estimation methodology and the use of control variables included in each model.

SGPs are generated using a nonparametric quantile regression model in which student achievement in one period is assumed to be a function of one or more years of

prior achievement (Koenker, 2005).[9] SGPs are descriptive measures of growth, designed to provide an easily-communicated metric to teachers and stakeholders. They are not, however, intended to be aggregated to the teacher level for the purpose of making causal inferences about the reasons for classroom-level differences in student growth (Betebenner, 2007). An argument for not making strong causal claims is that SGPs do not explicitly account for differences in student background by including covariates, such as the receipt of free or reduced price lunch, in the model. But, while they do not explicitly control for student background, they may *implicitly* account for differences in student backgrounds by utilizing a functional form that compares students that are quite similar to each other in terms of baseline achievement.

Value-added models have long been used by economists focusing on assessing the effects of schooling attributes (class size, teacher credentials, etc.) on student achievement (Hanushek, 1979, 1986). More recently they have been used in an attempt to identify the contributions of individual teachers in the learning process (Ballou et al., 2004; Gordon et al., 2006; McCaffrey et al., 2004). In VAMs, teacher performance estimates are generally derived in one step,[10] and unlike SGPs, performance estimates from VAMs are often intended to be treated as causal because they are estimated based on models that often include student covariates. The academic literature, however, is divided concerning the extent to which different VAM specifications can be used to

[9] Quantile regression is similar to OLS but instead of fitting the conditional mean of the dependent variable **Y** on the levels of the independent variables **X**, it fits the conditional quantiles of **Y** on **X**. Additionally, the calculation of SGPs employs polynomial splines (specifically B-splines or 'basis-splines') basis functions which employ smooth, non-linear regression lines that model non-linearity, heteroscedasticity and skewness of test score data over time. Critics of using B-splines in this way claim that their use adds a level of obscurity in the computation of SGPs since they fit data without parametric assumptions, which can increase the fit to the sample at the expense of the fit to the population (Castellano, 2011).

[10] However, other common estimation approaches include two stage Random Effects or Hierarchical Linear Modeling. For more information see Kane and Staiger (2008).

generate unbiased estimates of the effectiveness of individual teachers using nonexperimental data (Ballou et al., 2004; Chetty et al., 2011; Goldhaber and Chaplin, 2012; Kane and Staiger, 2008; Koedel and Betts, 2011; McCaffrey et al., 2004; Rothstein, 2010).[11]

The question of whether VAM or SGP estimates of teacher performance are in fact unbiased measures of true teacher performance is clearly important, but our focus in this paper is on the extent to which different model estimates are comparable.[12] A small body of literature compares estimates of teacher performance generated from different VAM specifications.[13] In general, teacher effects from models including and excluding classroom-level variables tend to be highly correlated with one another ($r > .9$) (Ballou et al., 2004; Harris and Sass, 2006; Lockwood et al, 2007), while models including and excluding school fixed effects yield estimates with correlations close to 0.5 (Harris and Sass, 2006). Research has also found that VAM estimates are more sensitive to the outcome measure than with model specification (Lockwood et al., 2007; Papay, 2010). There is less agreement between traditional models that use a prior-year test score and those that use same-year, cross-subject achievement scores to control for student ability at the high school level (Goldhaber et al., 2011).

While the above research does suggest relatively high levels of overall agreement for various VAMs, even relatively strong correlations can result in cross-specification differences in teacher classifications that could be problematic from a policy perspective. For instance, one of the criticisms of the well-publicized public release of teacher

[11] For more on the theoretical assumptions underlying typical VAM models see Harris et al. (2010), Rothstein (2010), and Todd and Wolpin (2003).
[12] We are unaware of any research that assesses the validity of SGP-based measures of individual teacher effects.
[13] For an example of VAM comparisons at the school level see Atteberry (2012).

effectiveness estimates connected to teacher names in the *Los Angeles Times* is that the estimates of effectiveness were sensitive to model specification. Briggs and Domingue (2011), for instance, compare results of teachers' ratings from a specification they argue is consistent with that used to estimate effectiveness for the *Times*, to specifications that include additional control variables.[14] The correlation between the two specifications is reported to be 0.92 in math and 0.79 in reading, but when teachers are classified into effectiveness quintiles, only about 60 percent of teachers retained the same effectiveness category in math, while in reading only 46 percent of teachers shared the same ratings of effectiveness (Briggs and Domingue, 2011). These findings are, of course, not surprising to statisticians, but the results of value-added analysis are viewed by teachers' unions as "inconsistent and inconclusive" (L.A.Times, 2010).

Another concern for policymakers who wish to tie teacher performance to student test scores is the stability of teacher effect estimates over time. Models may generate unbiased estimates of teacher effectiveness but still be unstable from one year to the next. Using a large dataset of elementary and middle school math tests in Florida, McCaffrey et al. (2009) estimate several VAMs and find year-to-year correlations generally ranging from 0.2 to 0.5 in elementary schools and 0.3 to 0.6 in middle schools. Comparing teacher rankings, they find that about a third of teachers ranked in the top quintile are again in the top quintile the next year. Goldhaber and Hansen (forthcoming) perform similar analyses using statewide elementary school data from North Carolina over a ten-year period. They find year-to-year correlations of teacher effectiveness estimates

[14] Additional variables at the student-level included race, gender, grade (4^{th} or 5^{th}), prior year test score, alternate subject prior year test score and student characteristics; teacher-level variables included years of experience, education, credential status. Specifics about the methodology used to estimate the value-added reported by the *LA Times* are described in Buddin (2010).

ranging from 0.3 to 0.6, and conclude that these magnitudes are not inconsistent with the stability of performance measures from occupations other than teaching.

Despite the increasing popularity of SGPs, there is less formal literature comparing SGPs to other models that produce teacher effect estimates. Goldschmidt et al. (2012) focus on differences between SGPs and VAMs at the school level and find school effects estimated with SGPs to be correlated most highly with effects generated by simple panel growth models and covariate adjusted VAMs with school random effects, with correlations between 0.23 and 0.91.[15] They also find a high level of agreement between SGPs and VAMs with random effects when schools are divided into quintiles of effectiveness. In elementary schools the two measures place schools in the same performance quintile 55 percent of the time and within one quintile 95 percent of the time. At the school level, effectiveness estimates based on SGPs are less stable over time (with correlations ranging between 0.32 and 0.46) than estimates generated by VAMs with school fixed effects (with correlations ranging between 0.71 and 0.81).

Ehlert et al. (2012) also investigate school-level differences between various models and find estimates from school level median SGPs, single-stage fixed effects VAMs, and two-stage VAMs to have correlations of about 0.85. Despite this overall level of agreement, they find meaningful differences between the estimates from different models for the most advantaged and disadvantaged schools. For example, only 4 percent of the schools in the top quartile according to the SGP model are high poverty schools, while 15 percent of the schools in the top quartile according to the two-stage VAMs are

[15] The Simple Panel Growth Model is described as a simple longitudinal mixed effects model, where the school effects are defined as the deviation of that school's trajectory from the average trajectory. The covariate adjusted VAMs with random school effects predict student achievement using multiple lagged math and reading scores and student background variables as covariates, along with school random effects.

considered high poverty. Furthermore, for schools that fall into the top quartile of effectiveness according to the two-stage method but fall outside the top quartile according to the SGP model, the aggregate percent free/reduced price lunch is about 70 percent. This percentage is only 33 percent for schools in the top quartile according to the SGP model but outside the top quartile according to the two-stage VAM.

There are few studies that compare the estimates of teacher effectiveness using median SGPs and VAMs at the teacher level. While not the focus of their research, Goldhaber and Walch (forthcoming) note correlations between a covariate adjusted VAM and median SGPs of roughly 0.6 in reading and 0.8 in math. The issue that concerns many educators and policymakers when considering which model to adopt is whether teachers charged with educating certain students might be disadvantaged by the measure. Wright (2010) addresses this issue in the context of comparing SGPs to the EVAAS model.[16] Wright finds greater negative correlations with median SGPs and classroom poverty level than the equivalent EVAAS model.

III. Data and Analytic Approach

A. North Carolina Data

The data we utilize for our research comparing teacher performance estimates generated from different models are managed by Duke University's North Carolina Education Research Data Center (NCERDC) and are comprised of administrative records of all teachers and students in North Carolina. This dataset includes student standardized

[16] The EVAAS model is a multivariate, longitudinal mixed model that uses up to 5 years of scores in multiple subjects and excludes student background characteristics.

test scores in math and reading, student background information, and teacher employment records and credentials from school years 1995–1996 through 2008–09.[17]

While the dataset does not explicitly match students to their classroom teachers, it does identify the person who administered each student's end-of-grade tests. There is good reason to believe that at the elementary level most of the listed proctors are, in fact, classroom teachers. In order to be more confident in our teacher-student link we take several precautionary measures. First we use a separate personnel file to eliminate proctors who are not designated as classroom teachers, or who are listed as teaching a grade that is inconsistent with the grade level of the proctored exam. We also restrict the analytic sample to include only self-contained, non-specialty classes in grades 3–5 and only include classrooms with at least 10 students for whom we can calculate an SGP and no more than 29 students (the maximum number of elementary students per classroom in North Carolina). In addition to these teacher-level restrictions, we also exclude students missing prior year test scores from the analytic sample, since pre-scores are required in the models we use to estimate teacher effectiveness.

These restrictions leave us a sample of 34,401 unique teachers and 120,267 unique teacher-year observations spanning 14 years. Descriptive statistics for the restricted and unrestricted samples are displayed in **Table 2.** We find the restricted and unrestricted samples to be very similar in terms of observable characteristics.

[Table 2 about here]

[17] This dataset has been used in many published studies (Clotfelter, et al., 2006; Goldhaber, 2007; Goldhaber and Anthony, 2007; Jackson and Bruegmann, 2009).

B. Methods of Estimating Teacher Performance

In our analysis we explore the extent to which six different teacher performance estimates compare to one another. These are generated either by SGP or VAMs as we describe below.

Student Growth Percentiles

Calculation of a student's SGP is based upon the conditional density associated with student prior scores at time *t* using prior scores as conditioning variables (Betebenner, 2008).[18] Given assessment scores for *t* occasions, $t \geq 2$, the υ-th conditional quantile for current test score, A_1 is based upon previous test scores, A_{t-1}, A_{t-2} ..., A_1 :

$$Q_{A_t}(\upsilon | A_{t-1}, \ldots, A_1) = \sum_{m=1}^{t-1} \sum_{k=1}^{7} \upsilon_{km}(A_m)\beta_{km}(\upsilon) \qquad (1)$$

The b-spline based quantile regression procedure is used to accommodate non-linearity, heteroscedasticity and skewness of the conditional density associated with the dependent variable and is denoted by 7 cubic polynomials pieced together, $k = 1, 2, \ldots, 7$, that "smooth" irregularities and previous test measures, $m = 1, \ldots, t-1$ (Betebenner, 2008).

[18] B-spline cubic basis functions, described in Wei and He (2006), are used in the parameterization of the conditional percentile function to improve goodness-of-fit, and the calculations are performed using the SGP package in R (R Development Core Team, 2008).

the program places knots at the 0.2, 0.4, 0.6, and 0.8 quantiles. Also, default arguments of the program were employed which produces rounded SGP values from 1 to 99 for each student. Considering the limitations of how SGPs are calculated by the SGP package in the statistical program R, we generate SGPs in two ways: using a maximum of three prior scores, and using one prior score.[20] However, because the teacher effectiveness estimates are very similar regardless of whether we use a single year or multiple years (r=0.96 for math; r=0.93 for reading), we choose to only report the results for SGPs based on a single prior-year score.

Student achievement, as measured by SGPs, is translated into teacher performance by taking the median SGP, which we refer to as MGP, across students for each teacher. We generate single year estimates for each teacher in the sample by taking the MGP for each teacher in each year. For teachers in the data for consecutive years, we generate 2-year teacher effectiveness estimates by pooling each teacher's students over two consecutive years and finding the MGP for that time period.[21]

Value-Added Models

A common method for estimating teacher effectiveness typically includes a fixed effect for each teacher spanning multiple time periods (which may be based on one or more years of student-teacher matched data):

---

sample observations lie in each interval. Boundaries are the points at which to anchor the B-spline basis (by default this is the range of data). For more information on Knots and Boundaries, see Racine, J. (2011). **LOSS**: Lowest Obtainable Scale Score; **HOSS**: Highest Obtainable Scale Score

[20] In particular, the conditional distribution of a student's SGP is limited to a sample of students having an equal number of prior score histories and values.

[21] We also generate teacher effectiveness estimates based on the mean of the student growth percentiles for each teacher but find the estimates to be highly correlated with the MGPs (r=0.96 in math; r=0.93 in reading). We choose to only report the results for MGPs.

$$A_{ijt} = \beta_0 A_{i(t\ prior)} + \beta_1 X_{it} + \beta_2 \kappa_{jt} + \tau_{jt} + \varepsilon_{ijt} \qquad (2)$$

where *i* indexes students, *j* indexes teachers, and *t* indexes time period. In the above general formulation, the estimated average within-teacher effect over each specified time period, *t*, is typically derived by regressing achievement at a particular point in time (A_{ijt}) on measures of prior achievement ($A_{i(tprior)}$), a vector of student background characteristics (X_{it}), a vector of classroom characteristics (κ_{jt}), and teacher fixed effects (τ_{jt}).[22] The error term (ε_{ijt}) is assumed to be uncorrelated with the predictors in the regression.

We estimate a number of variants of Equation 2. The simplest model, VAM 1, ("Lagged Score VAM") only includes same-subject student achievement in the previous year and teacher fixed effects. VAM 2 ("Student Background VAM") builds on the simple model by also including the vector of student background characteristics: gender, race/ethnicity, free/reduced price lunch status (FRL), learning disability, limited English proficiency, and parents' education level.[23] VAM 3 ("Classroom Characteristics VAM") includes student background and the following classroom-level variables: classroom percentages of FRL, disability, and minority students, percentage of students with parental education of bachelor's degree or higher, average prior-year math and reading achievement, and class size.

We also estimate a school fixed effects model (VAM 4), where we substitute the classroom-level characteristics in **Equation 2** with school fixed effects, (ζ_t):

[22] Teacher fixed effects for VAMs 1–3 are generated with the user-written Stata program `fese` (Nichols, 2008); teacher effects for VAMs 4 and 5, which require two levels of fixed effects, are generated with the user-written Stata program `felsdvreg` (Cornelissen, 2008).
[23] Note that some student background variables are not available in all years.

$$A_{ijt} = \beta_0 A_{i(t\ prior)} + \beta_1 X_{it} + \zeta_t + \tau_{jt} + \varepsilon_{ijt} \qquad (3)$$

This model specification yields estimates identified based on within school variation in student performance/teacher effectiveness.

In a final VAM specification, student background variables are replaced by individual student fixed effects, ς_i:

$$A_{ijt} = \varsigma_i + \tau_{jt} + \varepsilon_{ijt} \qquad (4)$$

The argument for estimating a student fixed effects specification is that student background variables may not adequately account for student heterogeneity whereas the inclusion of individual student effects does account for all *time invariant* observable and unobservable differences between students.[24] It is important to note that the teacher effects identification strategy of this specification (VAM 5: "Student Fixed Effects VAM") is quite different from those described above: here teacher effectiveness is identified based on within student variation in performance so students must be in the sample for multiple years. This requirement reduces the student sample from 2,398,173 to 992,669.[25]

[24] The inclusion of student fixed effects may not eliminate bias due to the student-teacher matching process as this matching may be "dynamic," i.e. based on *time-varying* factors (Rothstein, 2010).

[25] As is common in the literature (e.g. Harris et al., 2010), we instrument for the lagged same-subject achievement score using the twice-lagged same-subject achievement score to account for the fact that lagged achievement will be correlated with the error term in a model that has been first differenced. As a result, we can only include 4th and 5th grade students in the estimation sample, as 3rd grade students only have a single prior score. Since the student fixed effects specification requires that students be in the sample for at least two years, this means that the sample only includes students with test score data in 3rd,

We focus on the above five VAM specifications, along with the median student growth percentiles, in the analysis, but we also estimate a number of other variations of Equation 2**:** a specification with student *and* school-level characteristics; one that only includes same-subject and cross-subject prior achievement; one that includes student achievement from two prior years; and one that uses the same covariates as VAM 2 but excludes students with a disability in math, reading, or writing from the sample. Each of these specifications yielded teacher effectiveness estimates that were highly correlated (i.e. a correlation over 0.95 in math and 0.92 in reading) with at least one of the five VAM specifications that we describe above so, for the sake of parsimony, we do not describe these results in Section IV below.

In estimating teacher effectiveness one might incorporate one or more years of matched student-teacher data into the estimates of teacher effectiveness. VAMS 1, 2, and 5, if based on a single year, are really teacher-classroom-year effects (henceforth we refer to these below simply as teacher effects) since the teacher estimates cannot be distinguished from classroom- or school-level effects because there is no independent variation in classroom and school variables. For these VAMs, and MGPs, we estimate both 1- and 2-year teacher effects but keep the econometric specification consistent (i.e. do not add in classroom or school-level variables).[26]

It is not possible to estimate single-year effects for VAMs 3 and 4 (that independently identify school effects) because there is no variation in classroom or

---

4th, and 5th grade. In practice, we find that instrumenting makes little difference as the teacher effectiveness estimates generated from a student fixed effects model that does not instrument for lagged achievement are strongly correlated (r=0.93) with those generated from a model that does instrument for lagged achievement.

[26] While not reported, we also estimated 3-year teacher effectiveness estimates with all models, but since we find their behavior to be very similar to 2-year effects, we chose to not report them in the results.

school characteristics or school fixed effects within each year for teachers or students, but we can estimate 2-year models for these specifications.[27] For these models the student sample is based on the current and previous year (for example, the 2-year effects for teachers in 2005 are based on student achievement from 2005 and 2004; teachers who are absent from the sample in 2004 do not have a 2-year effectiveness estimate for 2005). Models for 2-year effects are estimated separately by time period and include grade and year dummy variables. We provide a summary description of each of the teacher effectiveness models in **Table 3** as a reference.

[Table 3 about here]

Finally, it is common practice to account for measurement error with an Empirical Bayes (EB) shrinkage adjustment, where each teacher effectiveness estimate is weighted toward the grand mean based on the reliability of the estimate (Aaronson et al., 2007; Boyd et al., 2008). Estimates with higher standard errors (i.e. those that are less reliable) are shrunken more toward the mean, while more reliable estimates are little affected. We find the adjusted estimates to very similar to the unadjusted estimates and choose to report results of the analysis using the unadjusted estimates.[28]

[27] In our 2-year models, school fixed effects are identified based on teachers who switch schools over time, and approximately 8 percent of the estimation sample drops due to non-mobility of teachers across schools. It is possible to estimate school fixed effects models with a single year of data (Rothstein, 2010), but these do not separately identify teacher and school effects. It is also possible to estimate single-year VAMs that include classroom characteristics using a two-stage random effects model (Johnson et al., 2012) but these models implicitly assume that teacher assignment to schools is random, which seems quite unlikely given the empirical evidence (e.g. Lankford et al., 2002).

[28] The correlations between the adjusted and unadjusted effects are greater than 0.99 in math and greater than 0.98 in reading for VAMs 1–3, and for VAM 4 (School FE) the correlations are 0.97 in math and 0.94 in reading. We found a considerable amount of estimation error with the VAM 5 estimates. In some years the estimate of the variance due to noise (weighted average of the standard errors of the teacher effects)

IV. Results

Prior to our discussion about agreement/disagreement between models when it comes to estimates of individual teacher effects, we briefly note a few findings about the estimated effect size estimates from each of the models used to estimate teacher performance. We report these effect sizes in **Table 4** for the 1-year performance estimates in math (column 1) and 1-year performance estimates in reading (column 2), and, where possible, for the 2-year performance estimates in math (column 3) and 2-year performance estimates in reading (column 4), for each of the methods of estimating teacher effectiveness. The table includes both unadjusted estimates as well as effect sizes adjusted, in the case of VAMs, for sampling error.[29]

[Table 4 about here]

Our findings suggest that a one standard deviation increase in the distribution of teacher effectiveness (for example from the median to the 84th percentile) translates to an increase of about 0.15 to 0.25 standard deviations of student achievement; magnitudes consistent with findings in other studies (e.g. Aaronson et al., 2007; Goldhaber and

---

was greater than the total variance of the effect estimates, resulting in shrunken estimates that differed considerably from the unadjusted estimates. In years with less noise, the correlations between adjusted and unadjusted VAM 5 estimates are between 0.78 and 0.96.

[29] Adjusted effect sizes are calculated using the following equation: $\sqrt{\hat{V}_{total} - \frac{1}{\sum k_j}\sum [\mathrm{SE}(\hat{\tau}_j)]^2 k_j}$ where $\hat{V}_{total}$ represents the variance of the estimated teacher effects, k_j represents the number of student observations contributing to each estimate, and $\mathrm{SE}(\hat{\tau}_j)$ represents the standard errors of the estimated teacher effects.

Hansen, forthcoming).[30] But, in addition, two patterns emerge. First, both the unadjusted and adjusted measures tend to be larger for math (except in the case of VAM 5). This is not surprising given that teacher effectiveness (and schooling resources in general) tends to explain a larger share of the variance in student outcomes in math than reading (e.g. Aaronson et al., 2007; Goldhaber and Hansen, 2010a; Koedel and Betts, 2007; Schochet and Chiang, 2010). Second, in most of the models the sampling error adjustment reduces the estimated magnitude of changes in teacher quality by about 20 percent. The exception is for VAM 5 (Student Fixed Effects VAM), where teacher effects are very imprecisely estimated so the sampling error adjustment reduces the effect size by over 80 percent.[31]

A. Correlation of Teacher Effects Across Model Specifications:

Table 5 reports both the Pearson and Spearman rank correlation coefficients (Pearson/Spearman) between the various single-year estimates of teacher effectiveness, and **Table 6** reports analogous correlations for the two-year effectiveness estimates.[32, 33]

[Tables 5 and 6 about here]

The findings in these tables are consistent with the existing literature in several respects. First, we find that the correlations are higher for estimates of teacher

[30] Research that estimates within-school teacher effects tends to find smaller effects sizes, in the neighborhood of 0.10 (Hanushek and Rivkin, 2010).
[31] We do not report effect sizes for the 2-year VAM 5 estimates because the estimated variance due to noise (the weighted average of the standard errors) is greater than the estimated total variance of the teacher effects.
[32] Recall that **Table 3** includes descriptions of the various model specifications/methods of estimating teacher effectiveness. We do not focus on the extent to which effectiveness in one subject corresponds to effectiveness in another subject. For more information on cross-subject correlations see Goldhaber et al. (2012).
[33] For similar correlation matrices using EB adjusted estimates see Table A1 and A2 in the appendix.

effectiveness in math than reading; for each of the model comparisons, the Spearman rank correlations are generally between 0.05 and 0.10 (with even greater differences in the Student FE VAM) higher for math (Panel A) than for reading (Panel B), again, consistent with findings that the signal-noise ratio tends to be higher for math than reading estimates.

Second, the addition of additional years of teacher performance data tends to increase the correlations, suggesting greater reliability of the estimates (Glazerman et al., 2010; Goldhaber and Hansen, forthcoming; McCaffrey et al., 2009; Schochet and Chaing, 2010). The increased correlation is quite modest for some models (e.g. those that just control for student background), but large in the case of the student fixed effects specification; for example the Spearman rank correlation between the Student Background VAM and Student FE VAM rises from 0.43 to 0.53 with the additional year in math and from 0.21 to 0.28 in reading. Third, the effectiveness estimates are little affected by the inclusion of student covariates beyond the base year test score (Ballou et al., 2004; Lockwood et al., 2008; Papay, 2011); specifically, we find the correlation between the Lagged Score VAM and the Student Background VAM to be over 0.90 for both math and reading.

Research on the influence of peer effects on student achievement tends to suggest that the composition of the classrooms in which students are enrolled can have an impact on their achievement. In particular, being in a school with high concentrations of student poverty (Caldas and Bankston, 1997), or low-achieving peers (Lavy et al., 2012) are found to have a negative impact on achievement. However, peer effects at the classroom-level are generally found to be small (i.e. relative to one's own background) (Burke and

Sass, 2008).[34] We too find that many of these classroom level characteristics are statistically significant, but the models in **Table 6** that include classroom level variables (VAM 3 includes a measure of poverty, prior year test scores, parental education, percent of students with a disability, and percent minority) are highly correlated (r=0.99) with the Student Background VAM specifications that does not include classroom-level controls.[35]

There is far less agreement between VAM 5 (School Fixed Effects) and the other VAM specifications; the Spearman rank correlations between this specification and models that adjust for prior year tests (with or without other student or classroom covariates) are in the neighborhood of 0.45-0.50 for reading and 0.55 for math. Thus, judging teacher effectiveness relative to other teachers in their same schools does alter the estimates of effectiveness, indicating that there is in fact some sorting of teacher effectiveness across schools.[36]

Estimates generated by the student fixed effects VAM have the lowest correlation with the other measures, with Spearman rank correlations with other single-year measures about 0.4 in math and 0.25 in reading. This finding is consistent with Harris and Sass (2006), who find a similar level of agreement (r =0.39) between Student Fixed

[34] For example, in math at the elementary level Burke and Sass (2008) find that a one point increase in mean peer achievement results in a statistically significant gain score increase of 0.04 points- equivalent to an increase 0.0015 standard deviations of achievement gains.

[35] One of the classroom level variables we include is class size and one might hypothesize that the high correlation for models with and without classroom level controls has to do with the fact that having a more challenging class (e.g. high poverty students) tends to be offset by having a lower class size. To test this, we re-estimate VAM 3 excluding class size, but find that this has little effect on the correlations. This finding is consistent with Harris and Sass (2006).

[36] Sass et al. (2010) find small average differences in teacher effectiveness across schools, with less affluent schools generally being staffed by less effective teachers, but also more variation in teacher effectiveness in higher poverty schools than in lower poverty schools.

Effects VAMs and specifications that include time-invariant student characteristics, and is likely due to the fact that fixed effects specifications are far less reliable.[37]

As we noted above, we are unaware of any research (other than Wright (2010) who compares MGPs to the EVAAS model) focusing on differences between VAM teacher effectiveness estimates and those generated from student growth percentiles. While SGP models are conceptually distinct from VAMs, it turns out that the estimates of teacher effectiveness show a high level of agreement; we find Spearman rank correlations over 0.9 for math and 0.8 for reading with VAM specifications that control for base year test scores and those that add in student or classroom level covariates.[38]

While the correlations across model specifications suggest the extent to which different models produce similar estimates of teacher effectiveness across the entire sample, they do not suggest whether these estimates may be systematically different for teachers instructing different types of students. We explore this issue in the sub-section C.

B. Magnitude of Difference between Models

The boxplots shown in **Figure 1** show the distribution of the absolute values of the differences in percentile ranks between each of the single-year models. The boxplots echo the correlations reported in the previous sub-section: we see a high degree of agreement between VAM 1, VAM 2, and MGPs, with much more disagreement between

[37] The specification that includes time-invariant student characteristics is most similar to our VAM 2 (Student Background VAM).

[38] These correlations are higher those reported in a similar comparison using school effectiveness measures (Ehlert et al., 2012). One explanation for this is that random factors (e.g. students having the flu on testing day) that influence student achievement, and hence teacher effectiveness measures, are a relatively more important when the level of aggregation is the classroom rather than the school. This is indeed what one would expect if the random shocks primarily occur at the classroom level but are not strongly correlated across classrooms in a school.

each of these measures and the Student Fixed Effects VAM (VAM 5). For instance, in math the median of the distributions of disagreement between VAM 1, VAM 2, and MGPs are between 4 and 7 percentile ranks, and the 75th percentile of the distribution is between 8 and 13 percentile ranks. The differences for the VAM 5 comparisons were much higher, with medians of 19 and 20 percentile ranks and 75th percentiles between 34 and 36 percentile ranks. The results for reading (Panel B) show slightly more disagreement between models across the board.

[Figure 1 about here]

This paper is not focused on the validity of particular models, but there is some evidence that value-added models that include lagged achievement (ideally several lagged test scores) and student covariates are less likely to be biased. Using experimental methods, Kane and Staiger (2008) find that the specification that includes controls for prior test scores, student background, and classroom composition (similar to our VAM 3) are the least likely to be biased. However, since we cannot generate single-year estimates of VAM 3, and the VAM 2 and VAM 3 estimates are highly correlated (r=0.99), we choose to focus on comparisons relative to VAM 2. In **Figure 2**, we illustrate the differences between each model and the VAM 2 estimates at different places along the effectiveness distribution. We first divide the effectiveness distribution into 20 quantiles and then plot the 10th percentile, median, and 90th percentile for each of the comparison models within each of the quantiles of VAM 2.

[Figure 2 about here]

The positive slopes for each of the lines indicate that positive values of VAM 2 are associated with positive values for each of the comparison models. Greater distance between the lines indicates greater variability in the estimates of the comparison models, and thus more disagreement between models, within each quantile of VAM 2. We see the greatest variability with the VAM 5 and VAM 4 estimates and the least variability for VAM 3 and VAM 1. The raw MGP is scaled differently from the VAMs, which makes comparisons across VAMs and MGPs difficult. For VAMs 1 and 3, the variability at the tails of the VAM 2 effectiveness distribution increases slightly, meaning that VAMs 1 and 3 disagree more at the tails than they do in the middle of the distribution. This is consistent with findings from Ballou et al. (2012); however, the pattern is not evident for VAM 4, VAM 5, or MGPs.

C. Classroom Characteristics Predicting Model Agreement/Disagreement

In order to assess how teacher performance varies according to the methodology used to estimate it, we first rank teachers on the performance distribution according to different models and then estimate the differential in percentile rankings between selected models as a function of three classroom level student characteristics: class average prior year achievement, the percentage of students receiving free or reduced price lunch, and the percentage of black students in the classroom.[39] For the sake of parsimony, we do not focus on every model combination, just differences between VAM 2 and the following

[39] We focus exclusively on prior achievement, free and reduced price lunch, and black students because there is relatively little variation in other student characteristics (e.g. Hispanic, Asian, limited English proficiency, etc.) at the classroom level in the North Carolina sample.

models: MGPs, in Panel A of **Table 7**; VAM 1 (Lagged Score), in Panel B; and VAM 5 (Student Fixed Effects), in Panel C.[40]

[Table 7 about here]

In column 1, for math, and column 5, for reading, we include linear, squared, and cubic measures of the class average prior year achievement (averaged across math and reading tests). For the comparisons with MGPs and VAM 1, for both math and reading, the relative teacher ranking according to the VAM 2 measure tends to decrease as average prior year achievement increases, but the rate in which it decreases varies across this class characteristic (i.e. the square and cubic terms are often statistically significant). Also of note is the finding that the rate of decrease is larger for reading than math. We see the opposite effect with the comparison with VAM 5; for classrooms with high prior year achievement, VAM 2 is favored over VAM 5. However, the large positive coefficient on the squared term in math indicates a U-shaped relationship, meaning that VAM 2 is generally higher than VAM 5 for teachers of classrooms with very low prior achievement.

We report comparable results for the level of free and reduced price lunch in column 2 (math) and column 6 (reading) and the percentage of black students in column 3 (math) and column 7 (reading). Here we see a similar pattern for both of the classroom characteristics. With the MGP and VAM 1 comparisons, as percent FRL and percent Black increase, teachers tend to have higher effectiveness ratings with VAM 2 rather than

[40] We use the single-year effects for each of these comparisons.

MGP or VAM 1. Again, we see the opposite effect with the VAM 5 comparison; as the classroom percentages of FRL and Black increase, VAM 5 is favored over VAM 2.

Finally, in columns 4 (math) and 8 (reading) we report the results when the various student characteristics are simultaneously included in the models. Given that there is a relatively high degree of correlation between the three classroom characteristics, it is not terribly surprising that the estimated sign on some classroom level variables are reversed and, in cases, the magnitudes of the estimated relationships change substantially. However, it turns out that the relationship across the classroom characteristics, which are shown in **Figure 3** (for the model that includes them simultaneously, column 4 for math and 8 for reading) look quite similar whether they are included in the model independently or when all three characteristics are included simultaneously. Furthermore, the generally steeper slopes for reading suggest that the effect of class composition is even greater than the effect for math.

[Figure 3 about here]

We illustrate the policy import of using different models to estimate teacher effectiveness by focusing on three different classroom types. Specifically, we define classrooms as being "advantaged," "average," or "disadvantaged" based on the aggregate student-level average prior achievement (a simple average of math and reading test scores across students in a classroom) and the percentage of the classroom enrolled in the free/reduced price lunch program. Advantaged classrooms are those in the lowest quintile of percent FRL and highest quintile of prior achievement; disadvantaged classrooms are

those in the bottom quintile of prior achievement and top quintile of percent FRL; average classrooms are in the middle quintile for both classroom characteristics.[41]

For the various model specifications, we predict teacher effectiveness for individual teachers, then average the effectiveness estimates for each of the stylized classroom types. The results of this exercise are reported in **Table 8**.

[Table 8 about here]

All models suggest that more advantaged classrooms tend to be staffed with more effective teachers, but the magnitude of the estimated teacher effectiveness difference between advantaged and disadvantaged classrooms is markedly different across models. For instance, the simple Lagged Score VAM (VAM 1) and MGP suggests extraordinarily large average percentile differentials of roughly 20 to 40 percentile points as compared to the school (VAM 4) and student (VAM 5) fixed effects specifications that suggest differentials that range from favoring disadvantaged schools (e.g. the 1-year performance estimate in reading) to favoring advantaged schools by 6 percentile points. The findings are to be expected given the way the models account for student background; i.e. those models that do not account for differences between students (other than through the base year test score comparison) will attribute student differences to teachers.

As we have stressed throughout, our comparison of models does not suggest whether one or another is more valid, but given the magnitude of the differentials presented above, the results certainly suggest that even when overall correlations between

[41] The analytic sample includes 13,164 "advantaged" classrooms, 5,541 "average" classrooms, and 13,448 "disadvantaged" classrooms according to these definitions.

measures of effectiveness are high, certain models favor certain teachers, depending on the composition of their classrooms. For example, the Lagged Score VAM and Student Background VAM are highly correlated (Pearson and Spearman correlation coefficients of 0.97); however, the average percentile rank for teachers of advantaged classes is about 7 percentile points higher for the Lagged Score VAM, while the average rank is about 8 points lower than that of the Student Background VAM for teachers in disadvantaged classes. This suggests that while models agree for most teachers in the middle of the distribution of student characteristics, the large differences in the tails of the distribution are enough to systematically favor one type of model over another for teachers with extreme classroom compositions.

D. Intertemporal Stability

In this subsection, we investigate the intemporal stability of VAMs and MGPs. The issue of intertemporal stability is an important one given that a high degree of instability is seen as a signal that VAMs are unreliable (Darling-Hammond et al., 2012) given that it is unlikely that we would see wide swings of true teacher performance from one year to the next. Moreover, higher levels of instability likely make the use of student-growth based measures of teacher effectiveness challenging from a policy standpoint (Baker, 2012; Glazerman et al., 2010).

The fact that some specific value-added models generate effectiveness estimates that are only moderately correlated from year to year has been well-documented (e.g. Goldhaber and Hansen, forthcoming; McCaffrey et al., 2009). To our knowledge there

have been no studies that investigate the intertemporal stability of teacher estimates from school fixed effect VAMs or MGPs.

In **Table 9**, we show the intertemporal correlations of performance estimates for 1 and 2-year performance estimates for math (Panel A) and reading (Panel B). The 1-year performance estimates show the correlations between adjacent years (t-1), the next column shows the correlations with one gap year (t-2), and so on. For the 2-year estimates the adjacent correlations are displayed in the t-2 column. For example, this column represents the correlation between the 2004 estimate (which is based on student achievement from 2004 and 2003) and the 2002 estimate (which is based on student achievement from 2002 and 2001). We exclude the t-1 column from the table of 2-year estimates because of the year of overlap.

[Table 9 about here]

Not surprisingly, the correlations rise across the board—for all model specifications and both subjects—when moving from a 1-year to a 2-year performance estimate, sometimes substantially. For instance, the adjacent correlation for the Lagged Score (VAM 1) reading estimates jumps from 0.41 in the single-year model to 0.54 in the two-year model. Also not surprising is the finding that adjacent performance periods are more highly correlated than are performance estimates with gap years in between them. Both of these are findings that have been previously documented for particular VAM specifications (Goldhaber and Hansen, forthcoming; McCaffrey et al., 2009).

We find the highest intertemporal correlations with the Lagged Score VAM, the Student Background VAM, and the Classroom Characteristics VAM (0.53 or greater in adjacent time periods in math, 0.35 or greater for reading), with slightly lower correlations for MGPs (0.49 or greater for adjacent time periods in math, 0.32 or greater for reading) and significantly lower correlations for the School FE and Student FE VAMs (0.32 or lower in adjacent years for math, 0.15 in reading).[42]

We would expect a higher degree of intertemporal variability for the school FE VAMs since it is likely easier to change relative position within a school than it is to change relative position across the whole workforce. For example, an above average teacher in a school of above average teachers may have within-school rankings that vary year-to-year but still be consistently above average relative to all teachers in the sample. The lower intertemporal correlations for the Student FE VAMs is also expected given that this specification is in general more noisy than estimates from VAMs that include student covariates (Sass et al., 2010).

V. Conclusions

Policymakers wishing to utilize student growth measures as an indicator of a teacher's job performance rightly worry about the properties of the estimates generated from different models, as well as the extent to which model choice might influence teacher rankings. We explore these issues in this descriptive paper examining the extent to which different methods of translating student test scores into measures of teacher

[42] McCaffrey et al. (2009) also report lower intertemporal correlations for teacher effects based on student fixed effects models relative to effects based on specifications that include student covariates.

performance produce consistent rankings of teachers, the magnitude of differences in performance estimates between methods and whether classroom characteristics predict these differences, and the stability over time of different performance estimates.

To be clear, we do not believe that the results presented here provide any definitive guidance about which model ought to be adopted; the impact on students of using a student growth measure as a factor in teacher performance ratings will depend both on how it is used and whether its use leads to behavioral changes. SGPs, for instance, are described as a means of sidestepping "many of the thorny questions of causal attribution and instead provide descriptions of student growth that have the ability to inform discussions about assessment outcomes and their relation to education quality" (Betebenner, 2008, p. 2). For the purpose of starting conversations about student achievement, SGPs might be a useful tool, but one might wish to use a different methodology for rewarding teacher performance or making high-stakes teacher selection decisions.[43]

Complicating matters is the fact that how teachers respond to the use of student growth measures may be influenced by their *perceptions* about whether the growth measures are fair and accurate, and perceptions may or may not line up with the true properties of a measure. Ultimately then it is not possible to determine whether one model appears to be preferable to another without assessing the effect of using student growth measures on the teacher workforce.

[43] Ehlert et al. (2012) make the case for using a two-stage value-added approach that assumes the correlation between growth measures of effectiveness and student background covariates is zero.

We do believe it makes sense for policymakers to be transparent with stakeholders about how teacher rankings can be affected by model choice. This is important, for our findings show that models that, broadly speaking, agree with one another (in terms of high correlations) can still generate, arguably, meaningful differences in teacher rankings that correlate with the type of students they are serving. Moreover, some models that could be seen by researchers as providing more accurate estimates of true teacher performance (i.e. school and student fixed effects specifications) generate effectiveness estimates that are far less stable over time than less saturated specifications. And, issues like intertemporal stability, or the transparency of a measure, may also influence teachers' perceptions of the measure.

References:

Aaronson, D., Barrow, L., & Sander, W. (2007). Teachers and student achievement in the Chicago Public High Schools. *Journal of Labor Economics*, *25*(1), 95–135.

Aldine Independent School District. (2012). *Supporting Quality teaching and Rewarding Effectiveness: A Guide to Proposed changes in Evaluation*.

Atteberry, A. (2011). *Defining School Value-Added: Do Schools that Appear Strong on One Measure Appear Strong on Another?* Evanston, IL.

Baker, B. D. (2012). *The Toxic Trifecta, Bad Measurement & Evolving Teacher Evaluation Policies*.

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

Ballou, D., Sanders, W. L., & Wright, P. S. (2004). Controlling for Student Background in Value-Added Assessment of Teachers. *Journal of Education and Behavioral Statistics*, *29*(1), 37–65.

Betebenner, D. (2007). Estimation of Student Growth Percentiles for the Colorado Student Assessment Program. National Center for the Improvement of Educational Assessment. Available online: http://www.cde.state.co.us/cdedocs/Research/PDF/technicalsgppaper_betebenner.pdf

Betebenner, D. (2009). Norm- and Criterion-Referenced Student Growth. *Educational Measurement: Issues and Practice*, *28*(4), 42–51.

Boyd, D. J., Grossman, P. L., Lankford, H., Loeb, S., & Wyckoff, J. (2009). Teacher Preparation and Student Achievement. *Educational Evaluation and Policy Analysis*, *31*(4), 416–440. doi:10.3102/0162373709353129

Boyd, D. J., Grossman, P., Lankford, H., Loeb, S., & Wyckoff, J. (2008). Measuring Effect Sizes: the Effect of Measurement Error. *National Conference on Value-Added Modeling*. University of Wisconsin-Madison.

Boyd, D. J., Lankford, H., Loeb, S., & Wyckoff, J. H. (2010). *Teacher Layoffs: An Empirical Illustration of Seniority vs. Measures of Effectiveness*. Washingon, DC.

Briggs, D., & Domingue, B. (2011). Due diligence and the evaluation of teachers: A review of the value-added analysis underlying the effectiveness rankings of Los Angeles Unified School District teachers by the Los Angeles Times. National Education Policy Center. Available online: http://greatlakescenter.org/docs/Policy_Briefs/Briggs_LATimes.pdf

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

Burke, M. A., & Sass, T. R. (2008). Classroom Peer Effects and Student Achievement. CALDER Working Paper 18.

Caldas, SJ Bankston, C. (1997). Effect of school population socioeconomic status on individual academic achievement. *The Journal of Educational Research*, *90*(5), 269–277.

Castellano, K. (2011). Unpacking student growth percentiles statistical properties of regression-based approaches with implications for student and school classifications. University of Iowa.

Chetty, R., Friedman, J. N., & Rockoff, J. E. (2011). The long-term impacts of teachers: Teacher value-added and student outcomes in adulthood. NBER Working Paper 17699.

Clotfelter, C. (2006). Teacher-Student Matching and the Assessment of Teacher Effectiveness. *Journal of Human Resources*, *41*(4), 778 – 820.

Clotfelter, C. T., Ladd, H. F., & Vigdor, J. L. (2006a). Teacher-Student Matching and the Assessment of Teacher Effectiveness. *Journal of Human Resources*, *41*(4), 778–820.

Clotfelter, C. T., Ladd, H. F., & Vigdor, J. L. (2006b). Teacher-student matching and the assessment of teacher effectiveness. *Journal of Human Resources*, *41*(4).

Cornelissen, T. (2008). The Stata command felsdvreg to fit a linear model with two high-dimensional fixed effects. *Stata J.*, *8*(2), 170–189.

Darling-Hammond, L., Amrein-Beardsley, A., Haertel, E., & Rothstein, J. (2012). Evaluating teacher evaluation: Popular modes of evaluating teachers are fraught with inaccuracies and inconsistencies, but the field has identified better approaches. *Phi Delta Kappan*, *93*(6), 8 – 15.

Ehlert, M., Koedel, C., Parsons, E., & Podgursky, M. (2012). Selecting Growth Measures for School and Teacher Evaluations.

Gabriel, T. (2010, September 2). A Celebratory Road Trip for Education Secretary. *New York Times*, p. A24.

Glazerman, S., & Seifullah, A. (2010). An evaluation of the Teacher Advancement Program (TAP) in Chicago: Year two impact report. Washington, D.C.: Mathematica Policy Research, Inc.

Goldhaber, D. D., Brewer, D. J., & Anderson, D. J. (1999). A Three-Way Error Components Analysis of Educational Productivity. *Education Economics*, *7*(3), 199–208.

Goldhaber, D. D., & Chaplin, D. (2012). Assessing the "Rothstein Falsification Test." Does it Really Show Teacher Value-added Models are Biased? CEDR Working Paper 2012-1.3

Goldhaber, D., & Theobald, R. (2011). Managing the Teacher Workforce in Austere Times: The Implications of Teacher Layoffs. CEDR Working Paper 2011-1.3.

Goldhaber, D. (2007). Everyone's Doing It, But What Does Teacher Testing Tell Us About Teacher Effectiveness? *Journal of Human Resources*, *42*(4), 765–794.

Goldhaber, D. (2010). When the Stakes Are High, Can We Rely on Value-Added? Exploring the Use of Value-Added Models to Inform Teacher Workforce Decisions. Center for American Progress. Washingon, DC. Available online: http://www.americanprogress.org/issues/2010/12/pdf/vam.pdf

Goldhaber, D., & Anthony, E. (2007). Can Teacher Quality Be Effectively Assessed? National Board Certification as a Signal of Effective Teaching. *Review of Economics and Statistics*, *89*(1), 134–150.

Goldhaber, D., Cowan, J., & Walch, J. (2012). Is a Good elementary teacher always good? Assessing teacher performance estimates across contexts and comparison groups. CEDR Working Paper.

Goldhaber, D., DeArmond, M., & DeBurgomaster, S. (2011). Teacher Attitudes About Compensation Reform: Implications For Reform Implementation. *Industrial & Labor Relations Review*, *64*(3).

Goldhaber, D., & Hansen, M. (forthcoming). Is It Just a Bad Class? Assessing the Stability of Measured Teacher Performance. Forthcoming in *Economica*.
Goldhaber, D., & Hansen, M. (2010a). Is It Just a Bad Class? Assessing the Stability of Measured Teacher Performance. CEDR Working Paper 2010-3.
Goldhaber, D., & Hansen, M. (2010b). Using Performance on the Job to Inform Teacher Tenure Decisions. *American Economic Review*, *100*(2), 250–255.
Goldschmidt, P., Choi, K., & Beaudoin, J. P. (2012). Growth Model Comparison Study: A Summary of Results. Council of Chief State School Officers. Available online: http://www.ccsso.org/Documents/2012/Summary_of_Growth_Model_Comparison_Study_2012.pdf
Gordon, R. J., Kane, T. J., & Staiger, D. O. (2006). Identifying Effective Teachers Using Performance on the Job. Hamilton Project White Paper. Washington, D.C.: Brookings Institution.
Haertel, E. (2009). Student Growth Data for Productivity Indicator Systems. Center for K-12 Assessment & Performance Management.
Hanushek, E. A. (1986). The Economics of Schooling - Production and Efficiency in Public-Schools. *Journal of Economic Literature*, *24*(3), 1141–1177.
Hanushek, E. A., & Rivkin, S. G. (2010). Generalizations about using value-added measures of teacher quality. *American Economic Review*, *100*(2), 267–271.
Harris, D. N., & Sass, T. R. (2011). Teacher training, teacher quality and student achievement. *Journal of Public Economics*, *95*(7-8), 798–812.
Harris, D., & Sass, T. R. (2006). Value-Added Models and the Measurement of Teacher Quality.
Harris, D., Sass, T., & Semykina, A. (2010). Value-Added Models and the Measurement of Teacher Productivity. CALDER Working Paper No. 54.
Jackson, C. K., & Bruegmann, E. (2009). Teaching Students and Teaching Each Other: The Importance of Peer Learning for Teachers. *American Economic Journal: Applied Economics*, *1*(4), 85–108.
Jacob, B. A., & Lefgren, L. (2008). Can Principals Identify Effective Teachers? Evidence on Subjective Performance Evaluation in Education. *Journal of Labor Economics*, *26*(1), 101–136.
Johnson, M., Lipscomb, S., Gill, B., Booker, K., & Bruch, J. (2012). Value-Added Models for the Pittsburgh Public Schools. Mathematica Policy Research.
Kane, T. J., & Staiger, D. O. (2008). Estimating Teacher Impacts on Student Achievement: An Experimental Evaluation. NBER.
Koedel, C., & Betts, J. R. (2007). Re-Examining the Role of Teacher Quality in the Educational Production Function. San Diego, CA: University of Missouri.
Koedel, C., & Betts, J. R. (2011). Does Student Sorting Invalidate Value-Added Models of Teacher Effectiveness? An Extended Analysis of the Rothstein Critique. *Education Finance and Policy*, *6*(1), 18–42.
Koenker, R. (2005). Quantile Regression (Econometric Society Monographs). Cambridge University Press.
LA Times. (2010). Study critical of teacher layoffs. (December 24, 2010). *Los Angeles Times*: http://www.latimes.com/news/local/teachers-investigation/la-me-teachers-study-critical-ap,0,6451518.story

Lankford, H., Loeb, S., & Wyckoff, J. (2002). Teacher Sorting and the Plight of Urban Schools: A Descriptive Analysis. *Educational Evaluation and Policy Analysis*, *24*(1), 37–62.

Lavy, V., Silva, O., & Weinhardt, F. (2012). The Good, the Bad, and the Average: Evidence on Ability Peer Effects in Schools. *Journal of Labor Economics*, *30*(2), 367–414.

Learning Point Associates. (2010). Evaluating Teacher Effectiveness: Emerging Trends Re!ected in the State Phase 1 Race to the Top Applications. Naperville, IL. Available online: http://www.wested.org/schoolturnaroundcenter/docs/lpa-evaluating-effectiveness.pdf

Lockwood McCaffrey, D., Hamilton, L., Stecher, B., Le, Vi-Nhuan, Martinez, J.F., J. R. (2007). The Sensitivity of Value-Added Teacher Effect Estimates to Different Mathematics Achievement Measures. *Journal of Education Measurement*, *44*(1), 47–67. doi:10.1111/j.1745-3984.2007.00026.x

McCaffrey, D. F., Koretz, D., Lockwood, J. R., Louis, T. A., & Hamilton, L. S. (2004). Models for Value-Added Modeling of Teacher Effects. *Journal of Educational and Behavioral Statistics*, *29*(1), 67–101.

McCaffrey, D. F., Sass, T. R., Lockwood, J. R., & Mihaly, K. (2009). The Intertemporal Variability of Teacher Effect Estimates. *Education Finance and Policy*, *4*(4), 572–606.

Nichols, A. (2008). FESE: Stata module to calculate standard errors for fixed effects. Boston College Department of Economics.

Papay, J. P. (2010). Different Tests, Different Answers: The Stability of Teacher Value-Added Estimates Across Outcome Measures. *American Educational Research Journal*, *48*(1), 163–193.

Podgursky, M. J., & Springer, M. G. (2007). Teacher Performance Pay: A Review. *Journal of Policy Analysis and Management*, *26*(4), 909–949.

Proctor, D., Walter, B., Reichardt, R., Goldhaber, D., & Walch, J. (2011). Making a Difference in Education Reform: ProComp External Evaluation Report 2006-2010. *Prepared for the Denver Public Schools*.

R Development Core Team. (2010). R: A language and environment for statistical computing. Vienna, Austria: R Foundation for Statistical Computing.

Racine, J. S. (2011). A Primer on Regression Splines. CRAN.R-Project. Available online: http://cran.r-project.org/web/packages/crs/vignettes/spline_primer.pdf

Rivkin, S. G., Hanushek, E. A., & Kain, J. F. (2005). Teachers, Schools, and Academic Achievement. *Econometrica*, *73*(2), 417–458.

Rockoff, J. E., & Speroni, C. (2011). Subjective and objective evaluations of teacher effectiveness: Evidence from New York City. *Labour Economics*, *18*(5), 687–696.

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

Rothstein, Jesse. (2009). Student Sorting and Bias in Value-Added Estimation: Selection on Observables and Unobservables. *Education Finance and Policy*, *4*(4), 537–571.

Sass, T., Hannaway, J., Xu, Z., Figlio, D., & Feng, L. (2010). Value Added of Teachers in High-Poverty Schools and Lower-Poverty Schools. Washingon, DC: The Urban Institute.

Schochet, P. Z., & Chiang, H. S. (2010). Error Rates in Measuring Teacher and School Performance Based on Student Test Score Gains. NCEE 2010-4004. *National Center for Education Evaluation and Regional Assistance*.

Seattle Public Schools (SPS). (2010). Collective Bargaining Agreement. Seattle, WA: Seattle Public Schools & Principal's Association of Seattle Schools.

Toch, T., & Rothman, R. (2008). *Rush to Judgment: Teacher Evaluation in Public Education*. Washington, D.C.: Education Sector.

Todd, P. E., & Wolpin, K. I. (2003). On the Specification and Estimation of the Production Function for Cognitive Achievement. *Economic Journal*, *113*(485), F3–F33.

Tucker, P. (1997). Lake Wobegon: Where All the Teachers Are Competent (Or, Have We Come to Terms with the Problem of Incompetent Teachers?). *Journal of Personnel Evaluation in Education*, *11*(2), 103–26.

Wei, Y., & He, X. (2006). Conditional growth charts. *The Annals of Statistics*, *34*(5), 2069–2097.

Weisburg, D., Sexton, S., Mulhern, J., & Keeling, D. (2009). The Widget Effect: Our National Failure to Acknowledge and Act of Differences in Teacher Effectiveness. The New Teacher Project.

Wright, P. S. (2010). An Investigation of Two Nonparametric Regression Models for Value-Added Assessment in Education. SAS Institute Inc: http://www.sas.com/resources/whitepaper/wp_16975.pdf

TABLES AND FIGURES:

Table 1. Plans for Race to the Top Winners

RttT 1st and 2nd Round Winners	Quantitative Teacher Evaluation Method	Percent of Teacher Evaluation based on Quantitative Measure (For teachers in tested subjects only)
Delaware	Not determined, but "likely that the standard will include value-added student growth analysis as part of the methodology."	"In Delaware, student growth is not one factor among many; instead satisfactory student growth is the minimum requirement for any educator to be rated effective."
Tennessee	Value-Added Model	35% on value-added (out of a total of 50% based on student achievement measures)
District of Columbia	Value-Added Model	50%
Florida	Value-Added Model	40% by end of grant
Georgia	Value-Added Model	50%
Hawaii	Value-Added Model	50%
Maryland	Growth Model	50% based on student growth (though not explicitly defined by student growth model)
Massachusetts	Growth Model	"...will be a cornerstone of evaluation protocols to be implemented statewide over the next four years."
New York	Value-Added & Growth Models	40% overall, divided between state and locally selected growth measures
North Carolina	Value-Added &/or Growth Model (local areas decide on a pre-approved growth measure in stage I, 2010-2012. State will adopt uniform system based on local experiences in stage II, 2012-2014.)	Does not use percent system. Teachers must be proficient in each evaluation category, including student growth, or subject to an intervention and potential dismissal.
Ohio	Value-Added Model	"...include student growth as a significant factor..."
Rhode Island	Value-Added and Growth Models	Value-added measure will count toward 51% of teacher evaluation by 2013–14

Table 2. Descriptive Statistics for Restricted and Unrestricted Samples

	Unrestricted Sample **Mean**	Restricted Sample **Mean**
Female	0.93	0.92
	(0.26)	(0.27)
White	0.84	0.85
	(0.36)	(0.36)
Black	0.14	0.13
	(0.35)	(0.34)
Hispanic	0.00	0.00
	(0.06)	(0.06)
Other Non-White	0.01	0.01
	(0.11)	(0.10)
Master's or Higher	0.28	0.27
	(0.45)	(0.45)
First Year Teacher	0.07	0.07
	(0.26)	(0.26)
Experience	12.46	12.42
	(9.78)	(9.84)
Self-Contained Classroom	0.90	1.00
	(0.30)	(0.00)
N (teacher-years)	169,041	120,267

Note: standard deviations in parentheses

Table 3. Descriptions of Teacher Effectiveness Measures

Model Name	**Control Variables**	**Single-Year/Two-Year Effects**	**Teacher Effectiveness Modeling Comparisons**
Lagged Score (VAM 1)	Prior year same-subject score	Single- and two-year effects	Variation in student achievement across all teachers
Student Background (VAM 2)	Prior year math and reading score, FRL status, disability status, parental education level, ELL status, race/ethnicity, gender	Single and two-year effects	Variation in student achievement across all teachers
Classroom Characteristics (VAM 3)	Same as VAM 2, but also includes classroom-level variables: class size, average prior year math and reading scores, %FRL, %parents with bachelors or higher, %disability, %minority	Two-year effects only	Variation in student achievement across all teachers
School Fixed Effects (VAM 4)	Same as VAM 2, but also includes school fixed effects	Two-year effects only	Variation in student achievement across teachers in each school
Student Fixed Effects (VAM 5)	Prior year same-subject score (instrumented using twice lagged score) and student fixed effects	Single and two-year effects	Within-student variation in achievement across time
MGP	Prior year same-subject score	Single and two year effects	Teacher-level median of the student-level SGPs

Table 4. Teacher Effect Sizes for Different Models

		Math Unadjusted/ Adjusted Effect Size	**Reading Unadjusted/ Adjusted Effect Size**
VAM 1	1 year	0.26/0.23	0.21/0.17
VAM 1	2 year	0.23/0.21	0.17/0.15
VAM 2	1 year	0.25/0.22	0.19/0.15
VAM 2	2 year	0.22/0.20	0.16/0.13
VAM 3	2 year	0.22/0.20	0.16/0.13
VAM 4	2 year	0.35/0.30	0.28/0.20
VAM 5	1 year	0.41/0.07	0.46/0.08
VAM 5	2 year	0.29/N/A	0.29/N/A

Notes: Adjusted effect sizes are calculated by subtracting out the estimated error variance (weighted average of the standard errors of teacher effects) from the total variance and taking the square root of the adjusted variance.

Table 5. Pairwise Pearson's and Spearman Rank Correlation Matrix of Single-Year Teacher Effectiveness Estimates [a]

Panel A. Math

	VAM 1 (Lagged Score)	VAM 2 (Student Background)	VAM 5 (Student FE)	MGP
VAM 1 (Lagged Score)	-			
VAM 2 (Student Background)	0.97/0.97	-		
VAM 5 (Student FE)	0.38/0.41	0.40/0.43	-	
MGP	0.93/0.93	0.91/0.91	0.36/0.39	-

Panel B. Reading

	VAM 1 (Lagged Score)	VAM 2 (Student Background)	VAM 5 (Student FE)	MGP
VAM 1 (Lagged Score)	-			
VAM 2 (Student Background)	0.92/0.91	-		
VAM 5 (Student FE)	0.23/0.25	0.25/0.27	-	
MGP	0.88/0.87	0.82/0.81	0.21/0.23	-

Note: Pearson's r/Spearman rank correlation

Table 6. Pairwise Pearson's and Spearman Rank Correlation Matrix of Two-Year Teacher Effectiveness Estimates

Panel A. Math

	VAM 1 (Lagged Score)	VAM 2 (Student Background)	VAM 3 (Classroom Characteristics)	VAM 4 (School FE)	VAM 5 (Student FE)	MGP
VAM 1 (Lagged Score)	-					
VAM 2 (Student Background)	0.97/.096	-				
VAM 3 (Classroom Characteristics)	0.96/0.95	0.99/0.99	-			
VAM 4 (School FE)	0.51/0.54	0.52/0.55	0.51/0.54	-		
VAM 5 (Student FE)	0.48/0.49	0.51/0.53	0.51/0.52	0.26/0.29	-	
MGP	0.94/0.94	0.92/0.92	0.91/0.91	0.49/0.52	0.47/0.48	-

Panel B. Reading

	VAM 1 (Lagged Score)	VAM 2 (Student Background)	VAM 3 (Classroom Characteristics)	VAM 4 (School FE)	VAM 5 (Student FE)	MGP
VAM 1 (Lagged Score)	-					
VAM 2 (Student Background)	0.91/0.90	-				
VAM 3 (Classroom Characteristics)	0.91/0.91	0.99/0.99	-			
VAM 4 (School FE)	0.41/0.46	0.46/0.52	0.45/0.51	-		
VAM 5 (Student FE)	0.28/0.30	0.32/0.33	0.31/0.32	0.14/0.17	-	
MGP	0.9/0.89	0.83/0.83	0.84/0.83	0.38/0.43	0.28/0.29	-

Note: Pearson's /Spearman rank correlation

Table 7. Predicting the Difference between VAM 2 and Other Measures Using Classroom Characteristics

Panel A. Difference between VAM 2 and MGP (VAM 2 Ranking - MGP Ranking)

	MATH				READING			
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Average Prior Achievement	-3.779***			3.365***	-10.89***			-2.584***
	(0.121)			(0.145)	(0.172)			(0.208)
Average Prior Achievement^2	2.850***			1.619***	-1.368***			-2.256***
	(0.136)			(0.144)	(0.194)			(0.207)
Average Prior Achievement^3	-0.990***			-2.078***	1.045***			-0.409*
	(0.149)			(0.148)	(0.213)			(0.213)
% FRL		0.122***		0.133***		0.279***		0.164***
		(0.0128)		(0.0131)		(0.0184)		(0.0189)
% FRL^2		-0.000562*		-0.000648**		-0.000768*		-0.000358
		(0.000304)		(0.000309)		(0.000438)		(0.000444)
% FRL^3		6.99e-06***		2.42e-06		1.75e-07		-4.22e-06
		(2.07e-06)		(2.10e-06)		(2.97e-06)		(3.01e-06)
% Black			0.111***	0.100***			0.103***	0.0907***
			(0.00845)	(0.00843)			(0.0122)	(0.0121)
% Black^2			0.000273	3.22e-05			0.00302***	0.00121***
			(0.000238)	(0.000240)			(0.000344)	(0.000345)
% Black^3			-1.43e-06	-5.89e-07			-2.45e-05***	-9.96e-06***
			(1.75e-06)	(1.77e-06)			(2.53e-06)	(2.54e-06)
Constant	0.166***	-4.672***	-2.973***	-7.473***	1.295***	-9.912***	-4.468***	-8.047***
	(0.0423)	(0.158)	(0.0724)	(0.185)	(0.0602)	(0.227)	(0.105)	(0.266)
Observations	120,267	120,267	120,267	120,267	120,267	120,267	120,267	120,267
R-squared	0.023	0.054	0.067	0.083	0.059	0.069	0.074	0.098

Notes: Standard errors in parentheses; *** p<0.01, ** p<0.05, * p<0.1

Panel B. Difference between VAM 2 and VAM 1 (VAM 2 Ranking - VAM 1 Ranking)

	MATH				READING			
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Average Prior Achievement	-10.33***			-2.346***	-19.41***			-9.834***
	(0.0630)			(0.0666)	(0.0988)			(0.113)
Average Prior Achievement^2	1.211***			-0.340***	-1.317***			-2.432***
	(0.0712)			(0.0661)	(0.112)			(0.112)
Average Prior Achievement^3	1.561***			0.323***	3.671***			1.913***
	(0.0781)			(0.0679)	(0.122)			(0.115)
% FRL		0.169***		0.114***		0.295***		0.119***
		(0.00640)		(0.00602)		(0.0109)		(0.0102)
% FRL^2		0.000176		-5.78e-05		0.00121***		0.00123***
		(0.000152)		(0.000142)		(0.000259)		(0.000240)
% FRL^3		1.29e-06		-6.50e-07		-1.25e-05***		-1.36e-05***
		(1.03e-06)		(9.63e-07)		(1.76e-06)		(1.63e-06)
% Black			0.0876***	0.0856***			0.113***	0.117***
			(0.00421)	(0.00387)			(0.00738)	(0.00654)
% Black^2			0.00239***	0.000608***			0.00471***	0.000673***
			(0.000119)	(0.000110)			(0.000208)	(0.000186)
% Black^3			-1.62e-05***	-4.16e-06***			-3.64e-05***	-6.94e-06***
			(8.73e-07)	(8.13e-07)			(1.53e-06)	(1.37e-06)
Constant	0.164***	-8.226***	-4.542***	-7.687***	0.919***	-14.24***	-6.789***	-9.144***
	(0.0221)	(0.0790)	(0.0361)	(0.0849)	(0.0347)	(0.134)	(0.0632)	(0.144)
Observations	120,267	120,267	120,267	120,267	120,267	120,267	120,267	120,267
R-squared	0.274	0.356	0.369	0.474	0.346	0.319	0.291	0.450

Notes: Standard errors in parentheses; *** p<0.01, ** p<0.05, * p<0.1

Panel C. Difference between VAM 2 and VAM 5 (VAM 2 Ranking - VAM 5 Ranking)

	MATH				READING			
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Average Prior Achievement	3.366***			-5.303***	8.576***			1.203
	(0.599)			(0.749)	(0.674)			(0.844)
Average Prior Achievement^2	6.824***			5.593***	1.991**			2.490**
	(0.799)			(0.883)	(0.899)			(0.996)
Average Prior Achievement^3	0.657			2.606***	1.240			2.194**
	(0.812)			(0.829)	(0.913)			(0.934)
% FRL		-0.326***		-0.213***		-0.322***		-0.159**
		(0.0623)		(0.0665)		(0.0702)		(0.0750)
% FRL^2		0.000428		-0.00200		0.00172		-0.00126
		(0.00150)		(0.00157)		(0.00169)		(0.00177)
% FRL^3		1.57e-05		2.42e-05**		-6.68e-06		9.25e-06
		(1.04e-05)		(1.09e-05)		(1.17e-05)		(1.23e-05)
% Black			-0.00367	0.0163			0.0323	0.0315
			(0.0424)	(0.0425)			(0.0478)	(0.0479)
% Black^2			-0.00321***	-0.000777			-0.00296**	0.000663
			(0.00124)	(0.00126)			(0.00140)	(0.00142)
% Black^3			3.08e-05***	9.08e-06			1.65e-05	-8.13e-06
			(9.40e-06)	(9.56e-06)			(1.06e-05)	(1.08e-05)
Constant	-1.536***	10.24***	1.724***	9.035***	-0.839***	10.72***	2.062***	6.741***
	(0.214)	(0.750)	(0.344)	(0.927)	(0.240)	(0.845)	(0.387)	(1.045)
Observations	31,150	31,150	31,150	31,150	31,150	31,150	31,150	31,150
R-squared	0.008	0.019	0.004	0.022	0.013	0.020	0.005	0.022

Notes: Standard errors in parentheses; *** p<0.01, ** p<0.05, * p<0.1

Table 8. Average Percentile Ranks for Typical Classrooms

Panel A. Math			
1 Year Percentile Ranks	Advantaged	Average	Disadvantaged
MGP	60.2	49.9	42.1
Lagged Score VAM	64.5	50.6	39.3
Student Background VAM	57.7	50.2	47.7
Student FE VAM	51.6	47.8	48.8
2 Year Percentile Ranks			
MGP	60.7	49.3	41.1
Lagged Score VAM	65.1	50.3	38.2
Student Background VAM	57.8	49.9	47.7
Classroom Characteristics VAM	60.1	49.7	46.6
School FE VAM	51.9	51.2	48.7
Student FE VAM	50.8	52.3	48.4
Panel B. Reading			
1 Year Percentile Ranks	Advantaged	Average	Disadvantaged
MGP	65.0	49.9	36.3
Lagged Score VAM	70.1	51.1	31.9
Student Background VAM	57.1	50.7	44.8
Student FE VAM	49.2	49.7	51.8
2 Year Percentile Ranks			
MGP	66.6	49.6	33.8
Lagged Score VAM	71.8	50.6	29.0
Student Background VAM	58.2	50.6	43.6
Classroom Characteristics VAM	60.3	50.2	42.8
School FE VAM	51.0	51.4	49.4
Student FE VAM	49.2	51.1	50.3

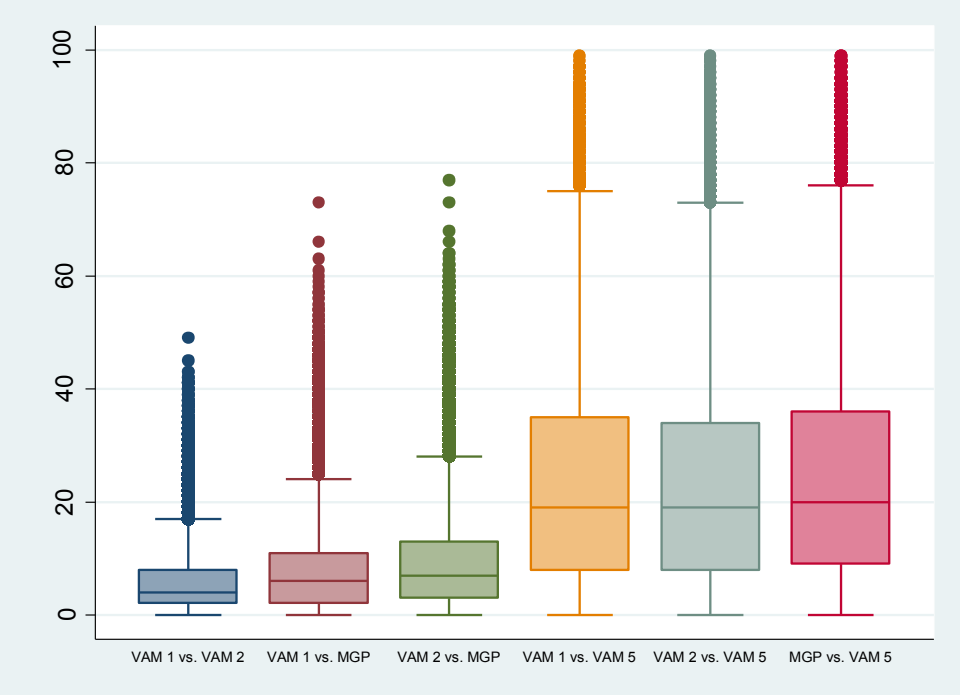
Notes: Advantaged is defined as Q5 for average prior achievement and Q1 for % Free lunch (13,164 teacher-years; 11% of the sample); average is defined as Q3 for average prior achievement and Q3 for % Free lunch (5,541 teacher-years; 5% of the sample); disadvantaged is defined as Q1 for average prior achievement and Q5 for % Free lunch (13,448 teacher-years; 11% of the sample)

Table 9. Intertemporal Stability of Effectiveness Estimates (Correlations Over Time)

Panel A. Math							
Single-Year Estimates	t-1	t-2	t-3	t-4	t-5	t-6	t-7
VAM 1, 1 year	0.54	0.48	0.45	0.41	0.39	0.36	0.35
VAM 2, 1 year	0.53	0.47	0.43	0.40	0.38	0.35	0.33
VAM 5, 1 year	0.22	0.12	0.11	0.11	0.09	0.10	0.01
MGP, 1 year	0.49	0.43	0.39	0.36	0.34	0.32	0.30
Two-Year Estimates							
VAM 1, 2 year	-	0.64	0.58	0.54	0.51	0.49	0.45
VAM 2, 2 year	-	0.63	0.57	0.53	0.50	0.47	0.43
VAM 3, 2 year	-	0.64	0.58	0.54	0.51	0.48	0.44
VAM 4, 2 year	-	0.24	0.20	0.18	0.17	0.18	0.17
VAM 5, 2 year	-	0.32	0.27	0.24	0.22	0.25	0.21
MGP, 2 year	-	0.59	0.53	0.49	0.46	0.43	0.40
Panel B. Reading							
Single-Year Estimates	t-1	t-2	t-3	t-4	t-5	t-6	t-7
VAM 1, 1 year	0.41	0.37	0.34	0.31	0.29	0.27	0.25
VAM 2, 1 year	0.35	0.31	0.28	0.25	0.24	0.22	0.21
VAM 5, 1 year	0.13	0.02	0.06	0.05	0.07	0.07	0.09
MGP, 1 year	0.32	0.28	0.25	0.24	0.22	0.21	0.19
Two-Year Estimates							
VAM 1, 2 year	-	0.54	0.49	0.46	0.42	0.39	0.36
VAM 2, 2 year	-	0.47	0.42	0.39	0.36	0.32	0.30
VAM 3, 2 year	-	0.49	0.44	0.40	0.37	0.33	0.30
VAM 4, 2 year	-	0.14	0.11	0.10	0.08	0.09	0.08
VAM 5, 2 year	-	0.15	0.14	0.17	0.12	0.13	0.19
MGP, 2 year	-	0.43	0.39	0.36	0.34	0.31	0.28

Figure 1. Absolute Values of the Differences between the Percentile Ranks According to Each Single-Year Measure

Panel A. Math



Panel B. Reading

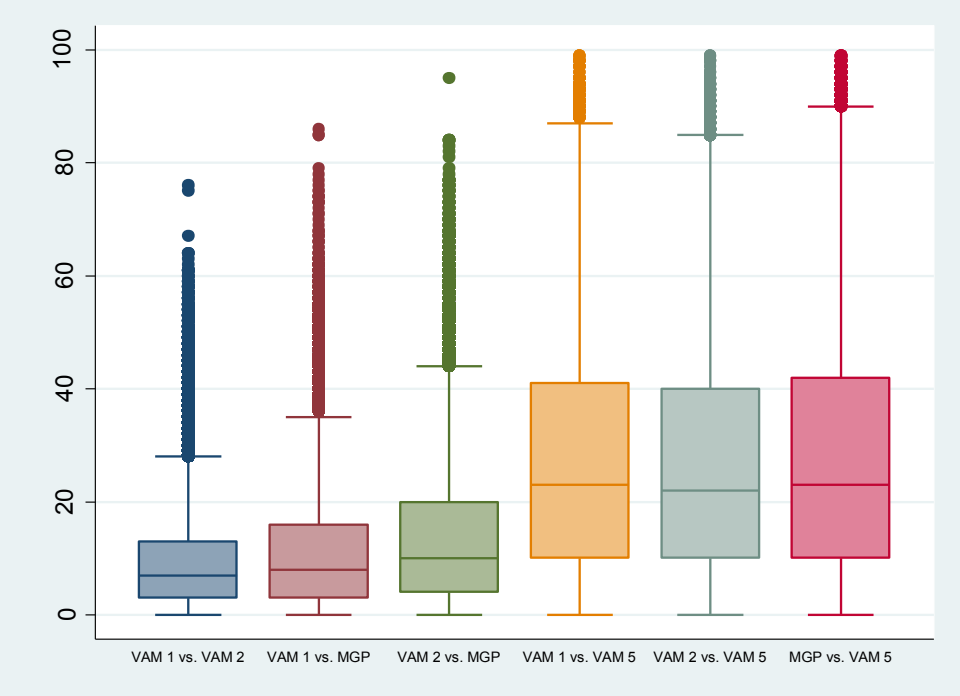
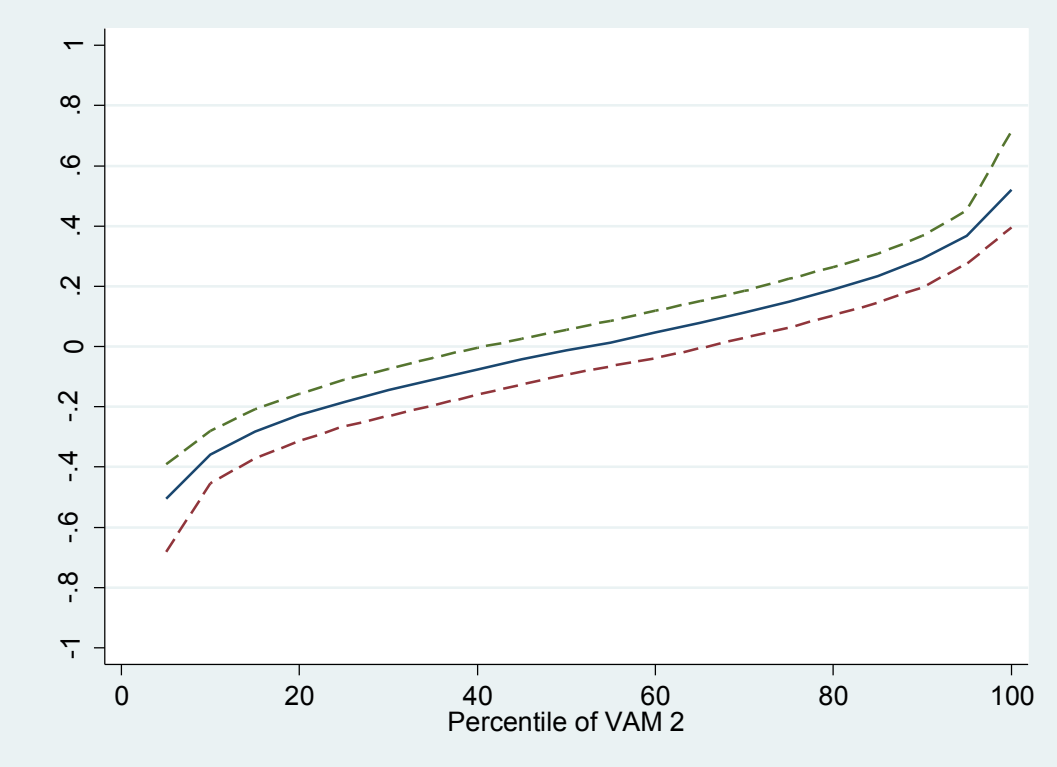
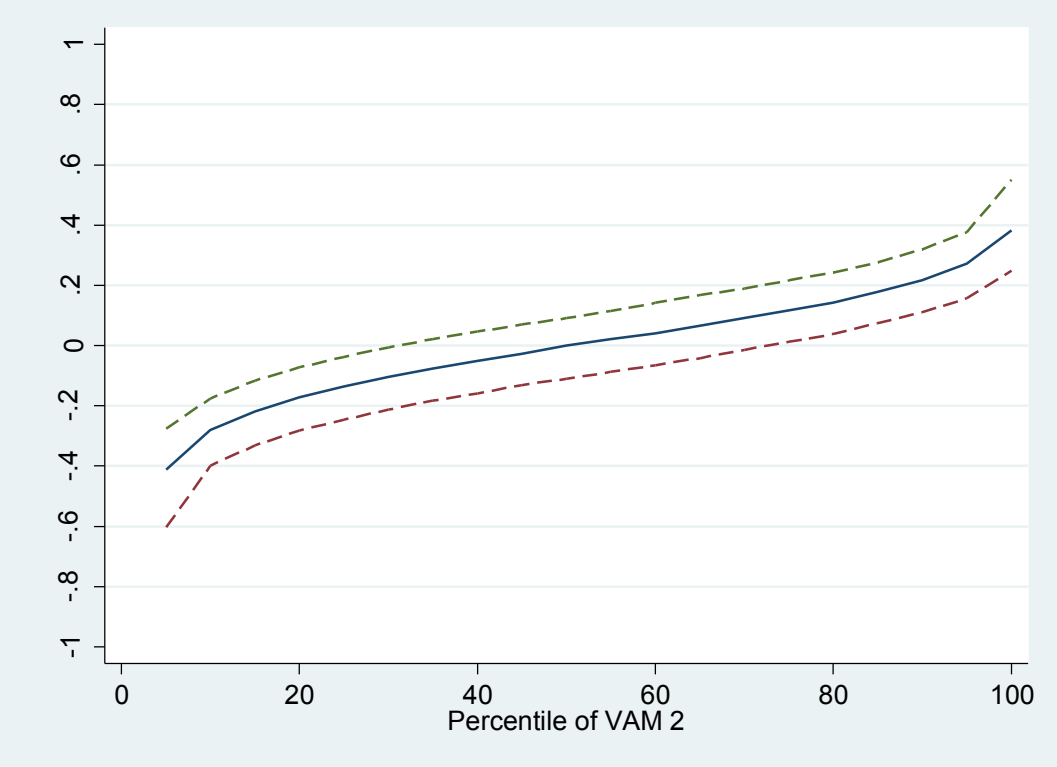


Figure 2. Effectiveness Estimates Across the VAM 2 Effectiveness Distribution (10th Percentile, Median, 90th Percentile)

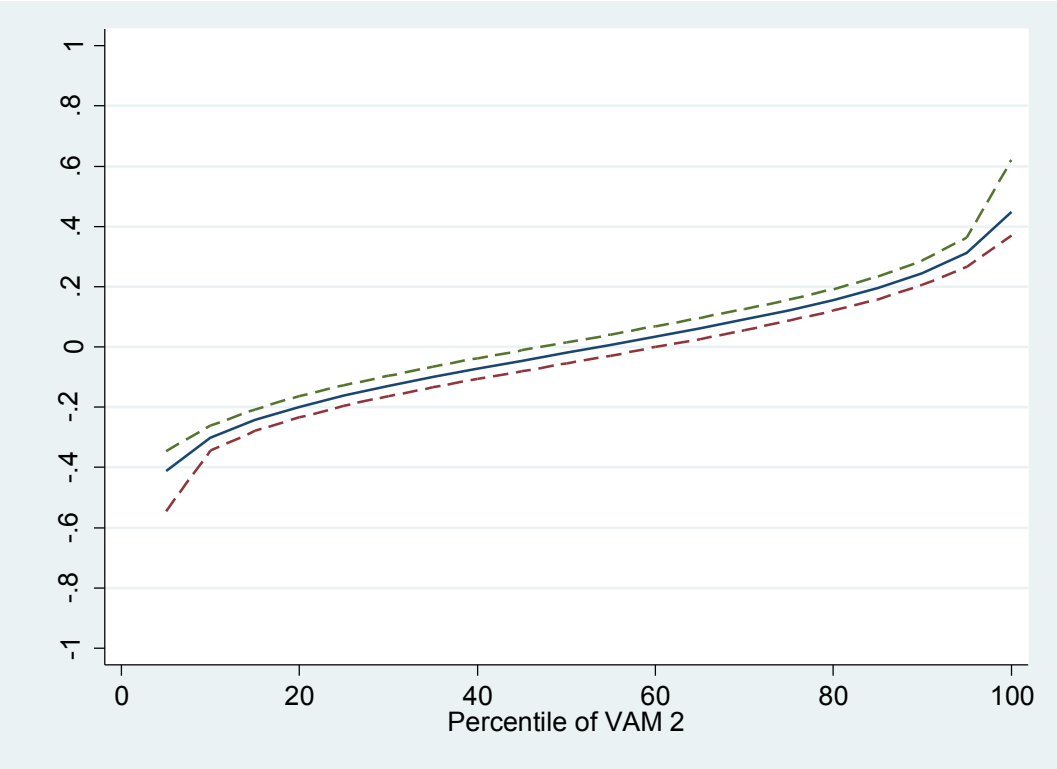
VAM 1, Math:



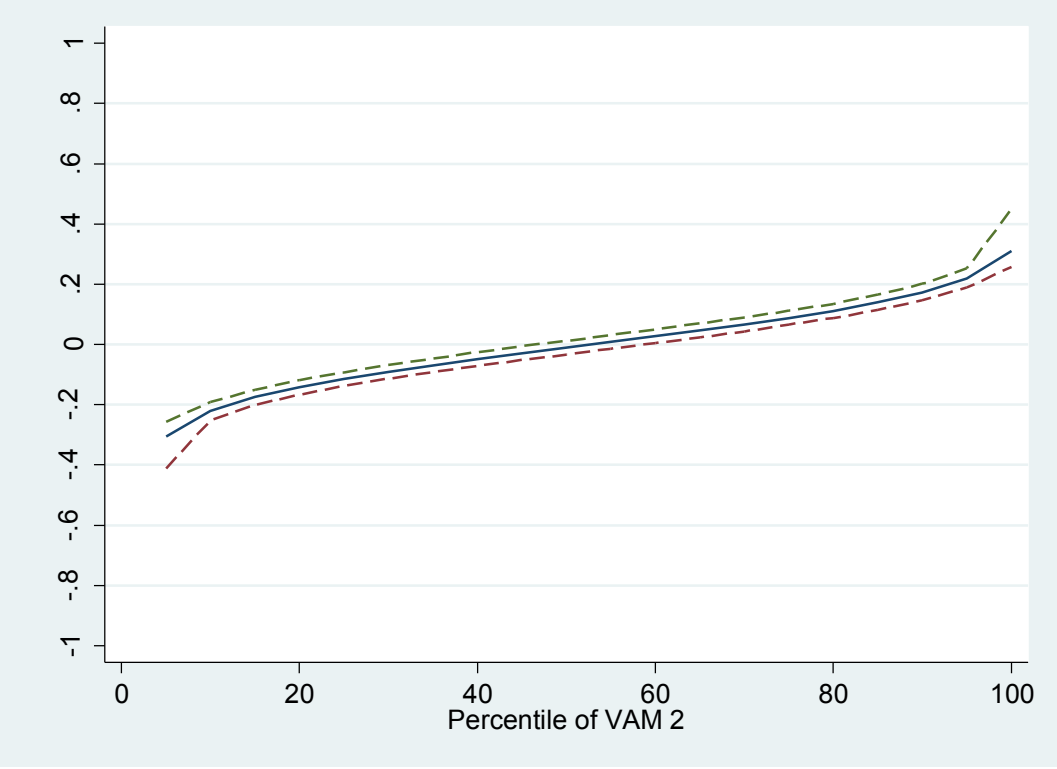
VAM 1, Reading:



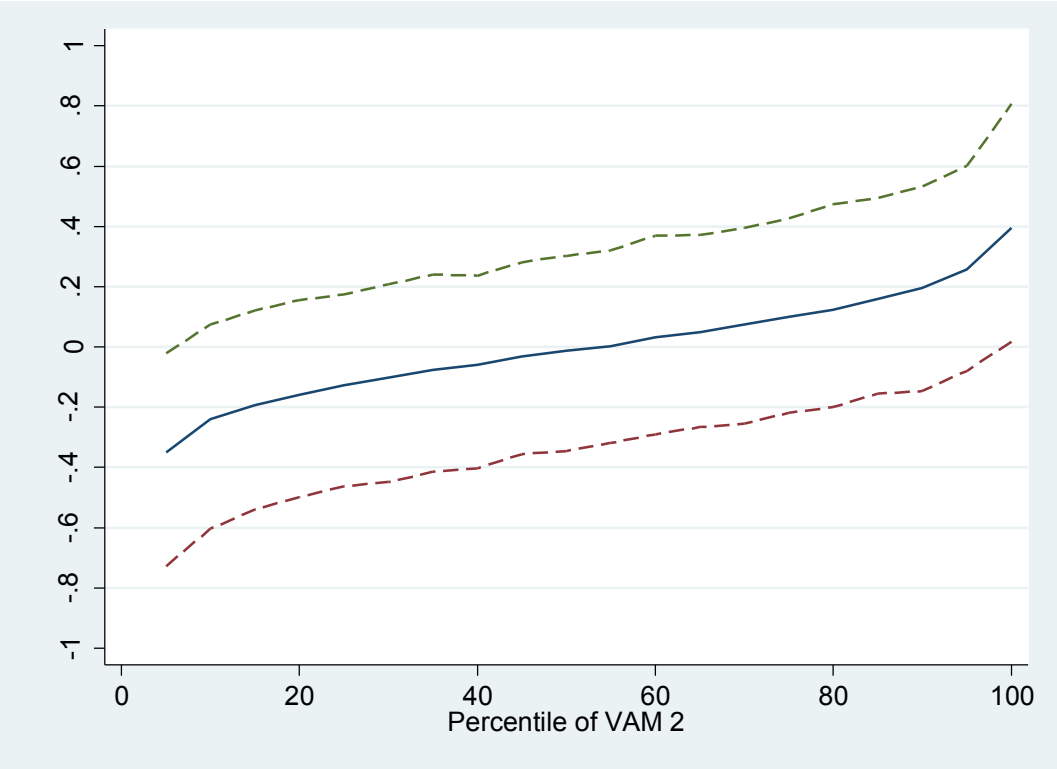
VAM 3, Math:



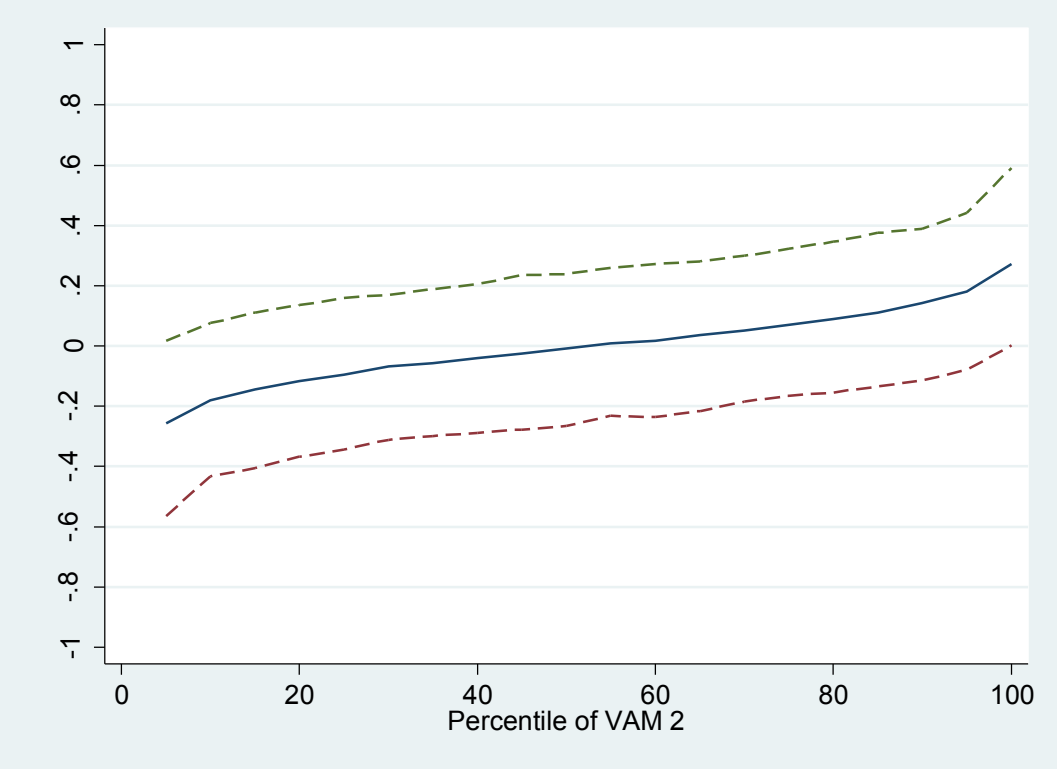
VAM 3, Reading:



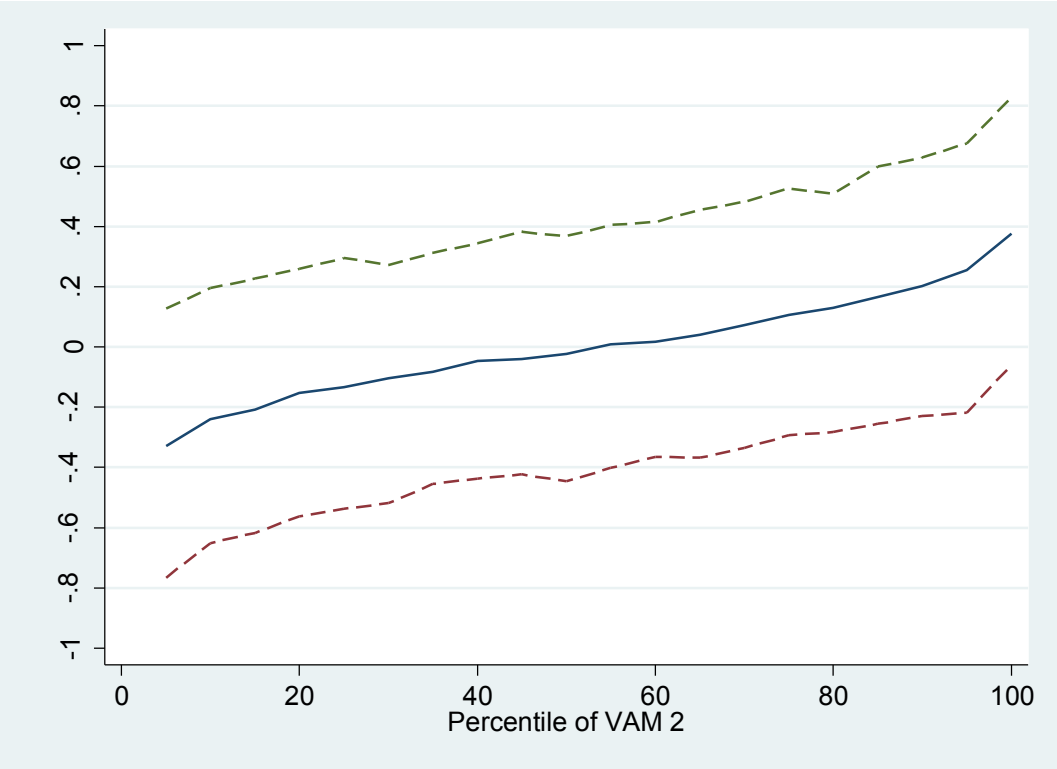
VAM 4, Math:



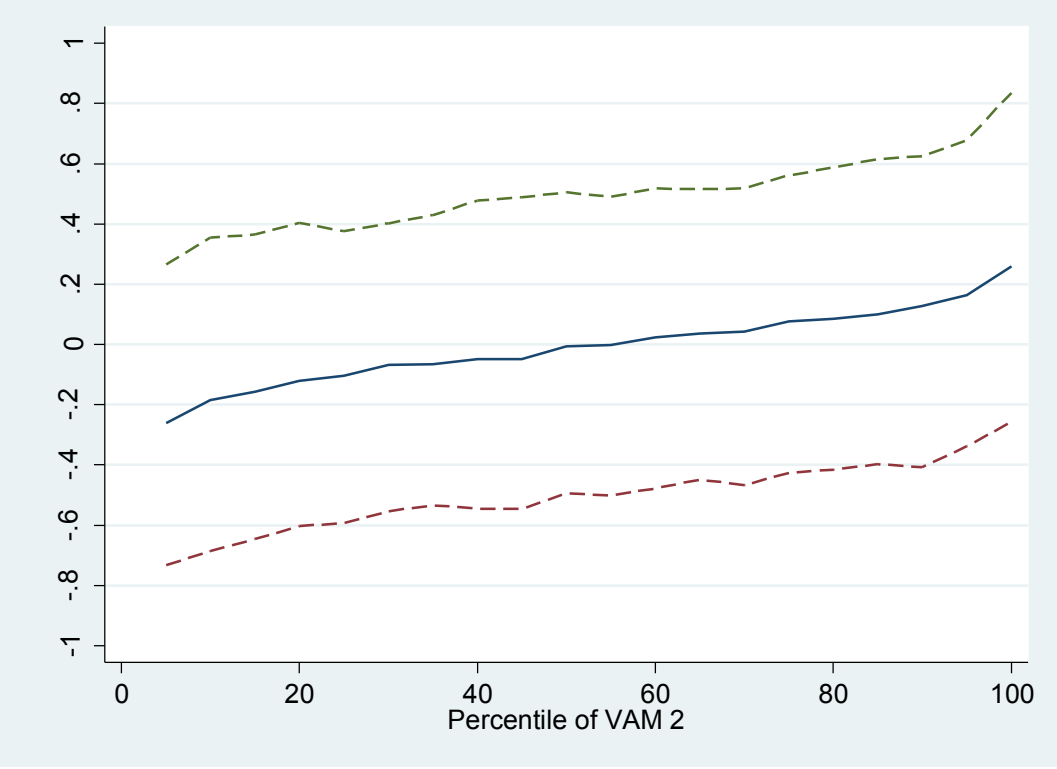
VAM 4, Reading:



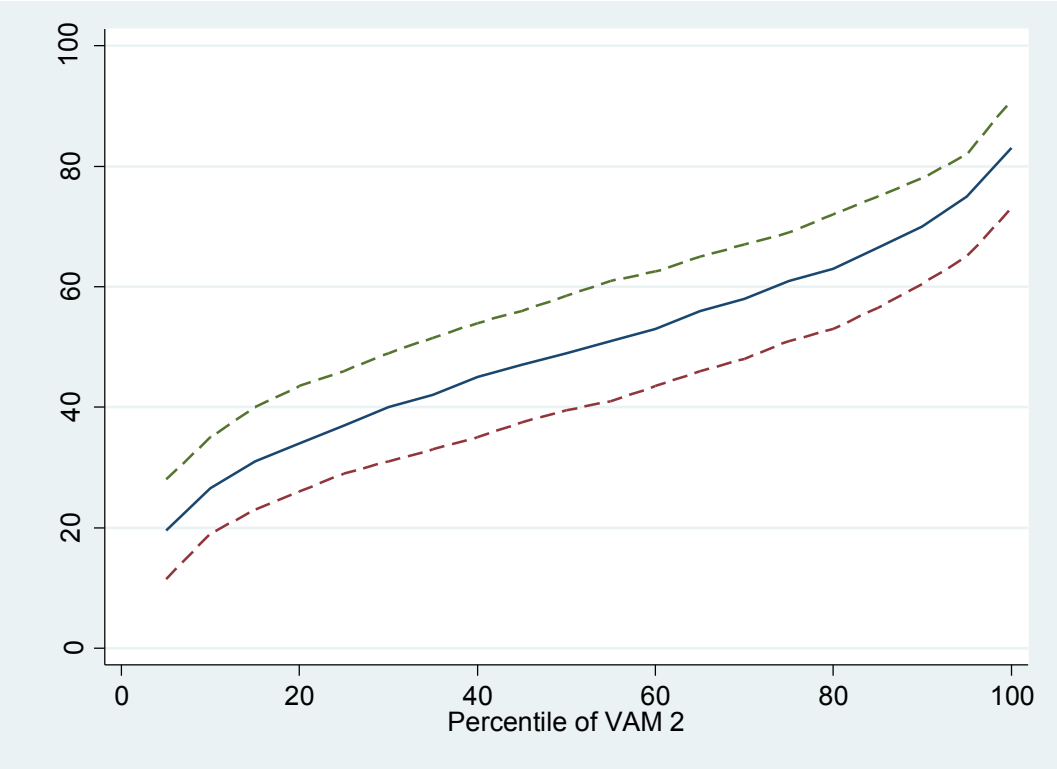
VAM 5, Math:



VAM 5, Reading:



MGP, Math:



MGP, Reading:

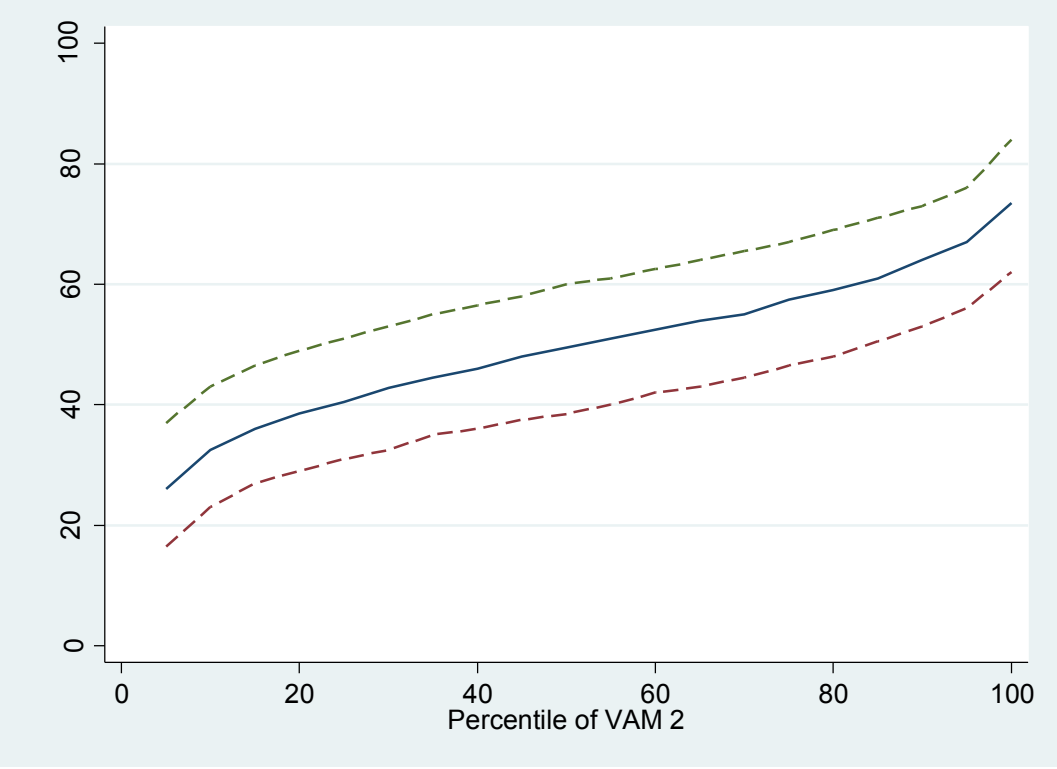
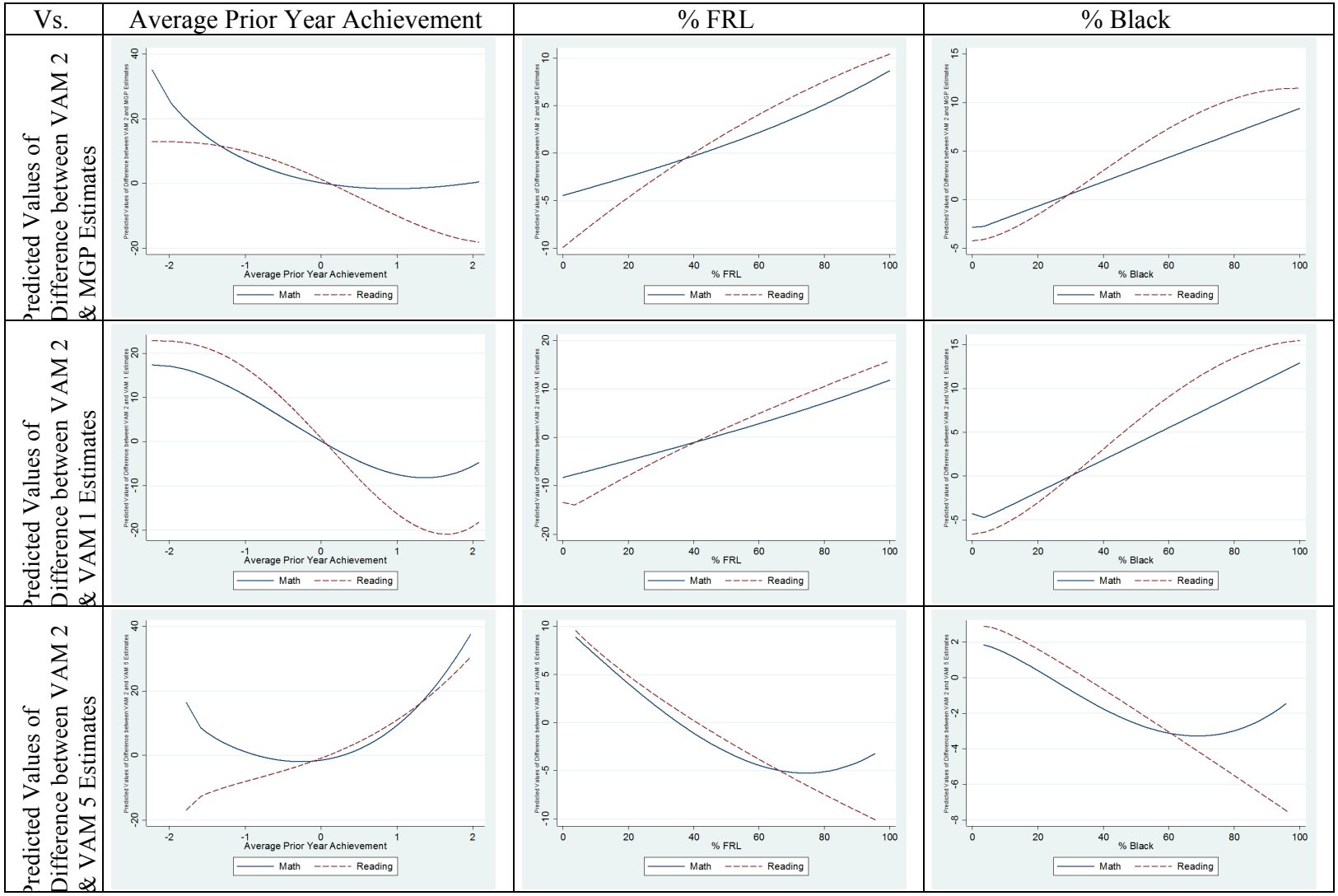


Figure 3. Predicted Differences between Models vs. Classroom Characteristics



Appendix

Table A1. Pairwise Pearson's and Spearman Rank Correlation Matrix of Single-Year Teacher Effectiveness Estimates, with EB Adjusted VAMs

Panel A. Math

	VAM 1 (Lagged Score)	VAM 2 (Student Background)	MGP
VAM 1 (Lagged Score)	-		
VAM 2 (Student Background)	0.97/0.97	-	
MGP	0.93/0.93	0.91/0.91	-

Panel B. Reading

	VAM 1 (Lagged Score)	VAM 2 (Student Background)	MGP
VAM 1 (Lagged Score)	-		
VAM 2 (Student Background)	0.91/0.90	-	
MGP	0.87/0.87	0.81/0.81	-

Notes: Pearson's r/Spearman rank correlation. VAM 5 is excluded from the matrix because the high degree of estimation error in some years yields EB adjusted estimates that are very different from the unadjusted estimates. MGPs are not adjusted for estimation error.

Table A2. Pairwise Pearson's and Spearman Rank Correlation Matrix of Two-Year Teacher Effectiveness Estimates, with EB Adjusted VAMs

Panel A. Math

	VAM 1 (Lagged Score)	VAM 2 (Student Background)	VAM 3 (Classroom Characteristics)	VAM 4 (School FE)	MGP
VAM 1 (Lagged Score)	-				
VAM 2 (Student Background)	0.97/.096	-			
VAM 3 (Classroom Characteristics)	0.96/0.95	0.99/0.99	-		
VAM 4 (School FE)	0.49/0.51	0.50/0.52	0.49/0.51	-	
MGP	0.94/0.94	0.93/0.92	0.91/0.91	0.47/0.49	-

Panel B. Reading

	VAM 1 (Lagged Score)	VAM 2 (Student Background)	VAM 3 (Classroom Characteristics)	VAM 4 (School FE)	MGP
VAM 1 (Lagged Score)	-				
VAM 2 (Student Background)	0.90/0.90	-			
VAM 3 (Classroom Characteristics)	0.91/0.90	0.99/0.99	-		
VAM 4 (School FE)	0.37/0.42	0.41/0.48	0.41/0.47	-	
MGP	0.90/0.89	0.83/0.82	0.83/0.82	0.34/0.40	-

Notes: Pearson's r/Spearman rank correlation. VAM 5 is excluded from the matrix because the high degree of estimation error in some years yields EB adjusted estimates that are very different from the unadjusted estimates. MGPs are not adjusted for estimation error.