CEDR Working Paper 2012-7

Is a Good Elementary Teacher Always Good? Assessing Teacher Performance Estimates Across Subjects

Dan Goldhaber, James Cowan, & Joe Walch

CEDR Working Paper 2012-7
University of Washington Bothell

The Center for Education Data & Research
University of Washington Bothell
3876 Bridge Way N. Ste 201 Seattle, WA 98103
206-547-5585
www.cedr.us

The research presented here is based primarily on confidential data from the North Carolina Education Research Center (NCERDC) at Duke University, directed by Clara Muschkin and supported by the Spencer Foundation. We wish to acknowledge the North Carolina Department of Public Instruction for its role in collecting this information and making it available and the Institute of Educational Studies at the Department of Education for providing financial support for this project. The views expressed in this paper do not necessarily reflect those of the University of Washington.

Suggested Citation
Goldhaber, D., Cowan, J., & Walch, J. (2012). Is a Good Elementary Teacher Always Good? Assessing Teacher Performance Estimates Across Subjects. CEDR Working Paper 2012-7. University of Washington, Seattle, WA.

© 2012 by Dan Goldhaber, James Cowan, and Joe Walch. All rights reserved.
Short sections of text, not to exceed two paragraphs, may be quoted without explicit permission, provided that full credit, including © notice, is given to the source.

1. The Use of Value-Added Teacher Effect Estimates

Policymakers are now using student growth-based measures of teacher effectiveness for a number of high-stakes personnel decisions. This policy direction is supported by research showing that teacher effectiveness varies widely and the variation has educationally meaningful consequences for student test achievement (Aaronson et al., 2007; Nye et al., 2004; Rivkin et al., 2005).[1] Some research cautions about the use of value added, raising issues as to the validity (Rothstein, 2009) and stability (McCaffrey et al., 2009) of effectiveness measures, as well as the possibility that teacher value-added effects "fade out" over time (Jacob et al., 2010; Konstantopoulos, 2007).[2] But, recent research (Chetty et al., 2011) provides a measure of external validity to value-added estimates, showing that value-added estimates of the impact of individual elementary and middle school teachers are statistically significant predictors of such later life student outcomes as college attendance and labor market earnings.

Regardless of the academic debate about value-added, it seems clear that policymakers are likely to accelerate the use of student growth based measures to inform high-stakes personnel decisions such as tenure and compensation. Indeed, current policy initiatives such as Race to the Top and the Teacher Incentive Fund have created financial incentives for states and districts to incorporate these measures into their teacher evaluation systems.

One of the key assumptions underlying the policy use of value-added at the elementary level, where teachers teach multiple subjects, is that teachers who are effective in one subject

[1] The literature typically finds teacher effect size estimates in the neighborhood of 0.10 and 0.25 standard deviations. The estimates are typically in the neighborhood of 0.10-0.15 for within-school estimates and are 0.15-0.25 for estimates that include between-school differences in teacher effectiveness. See, for instance, Goldhaber and Hansen (forthcoming) and Hanushek and Rivkin (2010) for a more thorough discussion of the teacher effect size literature.

[2] There is an active debate over how to interpret findings on validity and stability and whether value-added measures ought to be used for personnel decisions. See, for instance, Darling-Hammond et al. (2012), Glazerman et al. (2010), Goldhaber and Chaplin (2012), Hill (2009), and Harris (2009).

area also tend to be effective in other areas.[3] Surprisingly, this assumption, which we explore in this paper, has received little empirical attention.

The correlation of teacher effects across subjects has significant policy implications. For instance, clearly it is important to know whether policies that reward or sanction teachers based on value-added are likely to be rewarding or sanctioning teachers who are effective or ineffective across the subjects they are responsible for teaching. Also, should it turn out that teachers are often differentially effective, it might suggest elementary schools should think about some type of departmentalization, allowing teachers to specialize in a subject as is most common at the middle- and high-school levels (and does happen informally in some elementary schools).[4]

The gap in the literature on cross-subject teacher effectiveness is surprising as there is a burgeoning literature that explores the intertemporal stability of estimates as well as the stability of value-added estimates across model specification, and high and low stakes tests in the same subject area. This literature generally shows teacher effect estimates are highly correlated across model specification (Ballou et al., 2004; Goldhaber et al., 2012; Papay, 2011; Kane & Staiger, 2008). For instance, Goldhaber et al. (2012) finds correlations of 0.4-0.95 across models that include different combinations of student covariates, student fixed effects, and school fixed effects. The adjacent year correlations are far lower, however, as Goldhaber and Hansen (forthcoming) and McCaffrey et al. (2009) report adjacent year correlations in the neighborhood of 0.3 in reading and 0.5 for math. Finally, Lockwood et al. (2007) and Papay (2011) find that estimates of teacher effectiveness are sensitive to changes in the testing instrument, with correlations in value-added estimates across exams of about 0.1-0.6.[5] Taken as a whole, the

[3] The terms "teacher value-added", "teacher effectiveness", and "teacher performance" are used interchangeably here.
[4] See Public Impact (2012) for a proposal and discussion.
[5] Lockwood et al. (2007) estimate Pearson correlations while Papay (2011) estimates Spearman rank correlations.

literature suggests that teacher effectiveness exhibits educationally meaningful variation across test content and classrooms.

Few studies, however, look directly at the question of whether value-added measures are correlated across subjects. In a working paper, Koedel and Betts (2007) use several years of elementary level data to assess the cross-subject correlations of teacher effectiveness. They find a lower-bound cross-subject correlation of about 0.35 and an upper-bound (adjusted for sampling error) correlation of 0.64. Cross-subject correlations between mathematics and reading value-added estimated for use in New York City demonstrated correlations (unadjusted for sampling error) of between 0.4 and 0.55, depending on grade (Value-Added Research Center, 2010). The only published study on this issue by Lefgren and Sims (2012) implicitly shows a positive correlation. Specifically, it shows that the ability of past value-added measures to predict teachers' future value-added increases when composite math and reading measures of value-added are utilized.

In this paper we add to the sparse literature exploring teacher effectiveness across subjects. In particular, we use a 7-year panel of statewide data from elementary schools in North Carolina to estimate teacher value-added in math and reading using a variety of model specifications. We estimate correlations in value-added measures within years across subjects of about 0.6 and correlations across subjects in consecutive years of 0.3-0.4. Correcting for the sampling error in these estimates, we find correlations in underlying teaching effectiveness of 0.8-0.9 within years and 0.5-0.6 in consecutive years. We further find that these results are mostly robust to changes in observable classroom characteristics and are reflected in average changes in student achievement when teachers move across schools. These findings for cross-subject, consecutive-

year correlations are similar in size to those observed in consecutive years within the same subject.

2. Data

We use administrative records collected by the North Carolina Department of Public Instruction (NCDPI) and managed by Duke University's North Carolina Education Research Data Center (NCERDC). These data include information on student performance on standardized tests in math and reading (in grades 3 through 8) that are administered as part of the North Carolina accountability system. We standardize student test scores are standardized to have a mean of zero and a standard deviation of one within grades and years. The student data also include individual information about students, such as gender, race and ethnicity, disabilities, and FRL status. In order to use a stable set of exams and covariates across years, we use data for teachers and students from school years 1999 through 2005.

Students in the North Carolina data are linked to the proctor of their end-of-course exam. In order to ensure that we are matching students with their actual classroom teachers, we restrict our sample to self-contained classrooms in grades 3-5 with fewer than 29 students (the maximum for elementary classrooms in North Carolina). To ensure the reliability of our value-added estimates, we further restrict the sample to classrooms with at least 10 students.[6] Furthermore, because we estimate correlations between successive years of teacher effects, we drop observations from students in a class taught by a teacher they have had previously. The resulting

[6] The North Carolina data do not include explicit ways to match students to their classroom teachers. They do, however, identify the proctor of each student's end-of-grade tests, and in elementary school the exam proctors are generally the teachers for the class. We utilize the listed proctor as our proxy for a student's classroom teacher, but take several precautionary measures to reduce the possibility of inaccurate matches: 1) restricting our sample to those matches where the listed proctors have separate personnel file information and classroom assignments that are consistent with their teaching the specified grade and class for which they proctored the exam; 2) restricting the data sample to self-contained, non-specialty (such as special education or honors) classes.

sample (our analysis sample) includes 1,369,790 student-teacher observations; of these 734,110 are unique students and 21,633 are unique teachers.

In Panel A of **Table 1**, we report summary statistics for all students in a classroom with a valid student-teacher match. Comparing columns (3) and (4) of Panel A, it is apparent that students in matched classrooms are similar, though are somewhat more likely to be white and free-lunch eligible; they also have test scores of about 0.05 standard deviations higher than the student population as a whole. T-tests between the excluded and included samples show most of these differences to be statistically significant. This is not surprising given that low-achieving students are somewhat more likely to be mobile (Xu et al., 2009; Wright, 2001) and therefore less likely to have both a pre- and post-test score.

In Panel B of the table, we report descriptive statistics for teachers in 2005 (the last year in our sample), which is approximately representative of cross-sectional means over other years in the sample. As shown, teachers are primarily white and female. In terms of credentials, a minority holds master's degrees or higher or certifications from an approved North Carolina education program; a far higher proportion of the sample are fully licensed (that is, those teachers not holding a temporary or provisional license).

3. Analytic Approach

3.1. Estimating Value-added Measures of Teacher Effectiveness

There is a growing body of literature that examines the implications of using value-added models (VAMs) to identify causal impacts of schooling inputs, and indeed the contribution that individual teachers make toward student learning gains (Ballou, 2005; Ballou et al., 2004, 2012; McCaffrey et al., 2004; Rothstein, 2010; Todd and Wolpin; 2003).

Researchers typically utilize some variant of the following VAM:

$$A_{ijkst} = \alpha A_{is(t-1)} + X_{it}\gamma + C_{jt}\beta + \tau_j + \varsigma_k + \phi_t + \epsilon_{ijkst} \quad (1)$$

where i represents students, j represents teachers, k represents schools, s represents subject area (math or reading), and t represents the school year. Student achievement, A_{ijkst}, is regressed against: prior student achievement, $A_{ijks(t-1)}$; a vector of student and family background characteristics (for example, age, race and ethnicity, disability and free or reduced-price lunch (FRL) status, or parental education level), X_{it}; a vector of classroom characteristics (such as class size or average student characteristics, including achievement), C_{jt}; and teacher, τ_j, school, ς_k, and year, ϕ_t, fixed effects. The error term is associated with a particular student in a particular year ε_{ijkst}.[7]

If equation (1) is specified correctly, then $\hat{\tau}_j$ provides an estimate of the *time invariant* value-added contribution of teacher j to student learning. In policy applications, however, teacher effectiveness is often estimated as a one-year value-added measure. To analyze the stability of these measures across subjects, we estimate models separately by grade and year and estimate a teacher-by-classroom-by-year fixed effect, as in

$$A_{ijst} = \alpha A_{is(t-1)} + X_{it}\gamma + \tau_{jt} + \epsilon_{ijst} \quad (2)$$

In (2), the teacher-year effect is now confounded with classroom explanatory variables, the school effect, and the year effect. Importantly, this includes any annual variation in student achievement. Because the effects of excluded classroom level factors and annual random shocks are likely correlated across exams, these factors will likely induce spurious positive correlation between math and reading value-added estimated in the same year.

[7] Note that we focus on self-contained classrooms so that subject area does not vary by teacher, class, or school. However, the annual class grouping of students implies shared (and potentially unobservable) environmental factors that will influence the performance of the entire class, contributing to positive intra-class correlation among students in the same classroom that should be accounted for by clustering students at the classroom level.

Efforts to identify causal effects of teachers (or any schooling inputs) are further complicated by both the complexity of the schooling process, which can lead to model mis-specification, as well as teacher assignment and parental preferences, which can lead to nonrandom matches between teachers and students.[8] If sorting practices result in a common bias in math and reading value-added estimates, estimated correlations will overstate the true relationship between math and reading effectiveness.[9] In addition, to the extent that these practices are consistent within teachers over time, correlations of math and reading value-added estimated over several years, or correlations between one-year estimates from consecutive years, will likely also be biased upward.

Rothstein (2010) develops falsification tests of the assumptions underlying commonly used value-added methods using non-experimental data. Specifically, using longitudinal panel data from North Carolina, he shows that 5th grade teachers have statistically significant effects on 4th grade achievement gains. In a companion paper (Rothstein, 2009), he models how plausible student assignment mechanisms could introduce bias into value-added measurements. He finds that if principals sort students according to set rules based on their entire history of achievement, value-added estimates using one year of test score history could be biased. However, he also finds that regressions using multiple years of student test scores have minimal

[8] While we do not explicitly focus on these issues in this paper, other research has addressed them, and we are mindful of the implications. Todd & Wolpin (2003) discuss how the specification of the learning process corresponds to commonly used value-added estimates. Koedel and Betts (2008) show that value-added teacher effectiveness estimates may be sensitive to ceilings in the testing instrument. Our data show little evidence of a test ceiling, so we do not feel it should pose a problem since the impacts are fairly small in tests with only moderately skewed distributions, such as those we use here. For instance, the skewness of the distributions on test scores in our sample ranges between -0.397 and -0.233 in reading, and -0.302 and 0.301 in math (skewness = 0 for symmetric distributions) whereas Koedel and Betts (2008) find minimum competency tests have skewness measures ranging from -2.08 to -1.60, and these have the most consequential impacts on teacher effectiveness estimates and rankings.

[9] Sorting may happen both formally through ability tracking of students, or more informally, for instance, if "good" teachers are rewarded with choice class assignments or parents exert influence to get their children into choice classrooms.

bias under many of the assignment rules he considers.[10] This model is simply Equation (2) with additional regressors, as below:

$$A_{i,j,t,s,g=5} = \alpha \mathbf{A}_{\mathbf{i(history)}} + X_{i,t,g=5}\gamma + \tau_{j,t,g=5} + \varepsilon_{i,j,t,s,g=5},$$
$$\mathbf{A}_{\mathbf{i(history)}} = \left[A_{i,R,g=4} \middle| A_{i,M,g=4} \middle| A_{i,R,g=3} \middle| A_{i,M,g=3} \middle| A_{i,R,g=2} \middle| A_{i,M,g=2} \right]. \quad (3)$$

In the analyses below, we include the sample of 5th grade teachers and estimate value-added models of this form.

Other research suggests bias in value added may be negligible. Kane and Staiger (2008) exploit an experiment where teachers were randomized to classrooms. They fail to find evidence of bias in several common value-added models estimated on pre-experimental data. Chetty et al. (2011) perform several tests of the assumptions underlying value-added models. They exploit a richer dataset that includes socioeconomic information reported on parents' tax forms and find that these variables do not predict teacher assignments conditional on the variables commonly included in value-added regressions. They further find that movements of teachers across schools produce achievement gains in the incoming schools consistent with out-of-sample value-added estimates. Furthermore, Goldhaber and Chaplin (2012) show that Rothstein's test will falsify estimates under plausible assignment mechanisms that are consistent with unbiased value-added models.

In sum, the research on VAM specification suggests that performance estimates may be sensitive to the choice of variables included in student achievement regressions. We attempt to control for this sensitivity by estimating correlations on several standard models. In addition, we further test the robustness of the estimated correlations by estimating value-added models that appear robust to bias caused by non-random assignment (Rothstein, 2009; Rothstein, 2010).

[10] Rothstein's (2009) conclusion on the minimal bias in these estimates stems from the significance of this vector in explaining variation in student achievement. This vector captures almost all of the variation in student achievement, leaving very little room for non-random sorting or the like to bias student estimates.

3.2. Estimating Cross-Subject Correlations

We use Pearson's correlation coefficient $\left(\rho_{mr} = \sigma_{mr}/\sigma_m\sigma_r\right)$ to measure the correlation of estimated value-added across subjects. Direct pairwise correlations within years of observations on teachers provide an estimate of the stability of performance across subjects; however, teacher effectiveness is measured with error on each exam. Thus, the directly calculated correlation coefficient reflects instability in both performance and measurement. As noted above, the instability may reflect purely random variation due to the testing instrument or more systematic variation due to student sorting or classroom level random factors. We attempt to correct for both forms of sampling error by estimating correlations that correct for the estimation error in the value-added estimates and correlations across subjects in different years.

We estimates the correlation of true classroom effects across subjects, including classroom-level factors, by decomposing estimated performance into the true classroom effect and a random error:

$$\hat{\tau}_{jst} = \tau_{jst} + \varepsilon_{jst} . \qquad (4)$$

Under the assumption that errors in classrooms effects are uncorrelated, the correlation coefficient based on these two measures takes the following form:[11]

$$corr(\hat{\tau}_{j,m,t}, \hat{\tau}_{j,r,t}) = \hat{\rho}_{mr} = \frac{\text{cov}(\tau_{j,m,t}, \tau_{j,r,t})}{[\text{var}(\tau_{j,m,t}) + \text{var}(\varepsilon_{j,m,t})]^{1/2}[\text{var}(\tau_{j,r,t}) + \text{var}(\varepsilon_{j,r,t})]^{1/2}} . \qquad (5)$$

Hence, under the restrictive assumption that classroom errors are uncorrelated, the covariance of the teacher effects is isolated in the numerator. The denominator represents the noisy estimates of performance in both subjects. We estimate the standard errors of the true effectiveness τ_{jst} by

[11] This assumption is clearly problematic if there are classroom shocks common to both exams. We return to this point below. However, these correlation coefficients can be interpreted as the correlation between classroom-level measures of effectiveness that include any sources of common classroom-level variation while correcting for sampling error in the value-added estimates.

subtracting the sampling variance of the fixed effects estimators from their sample variance to obtain the corrected correlation $\tilde{\rho}_{mr}$:

$$corr(\tau_{j,m,t}, \tau_{j,r,t}) = \tilde{\rho}_{mr} = \frac{\text{cov}(\tau_{j,m,t}, \tau_{j,r,t})}{[\text{var}(\tau_{j,m,t})]^{1/2}[\text{var}(\tau_{j,r,t})]^{1/2}}. \quad (6)$$

Following Aaronson et al. (2007), we estimate the variance of ϵ_{jst} with the mean of the standard errors across all fixed effects.[12] The variance in the estimated teacher effects that remains after subtracting the mean standard error comprises the denominator in Equation (6).

While the cross-subject correlations of teacher effectiveness measures estimated above give an indication of the relationship of mathematics and reading instruction within a given classroom, they will overstate the correlation in true teacher effectiveness across subjects if there are classroom shocks that are common across exams. To investigate this possibility, we estimate correlations between mathematics (reading) value-added in one year and reading (mathematics) value-added in the next year. While these consecutive-year comparisons will be biased upward in the presence of systematic student sorting of the kind considered by Rothstein (2009; 2010), we show below that they are insensitive to several different models, including those Rothstein (2009) shows to be robust to plausible assignment mechanisms.

4. Results

We begin our discussion of the empirical results by reporting the standard deviation of the estimated teacher value-added. The first column in **Table 2** displays the standard deviation of the estimated teacher fixed effects, while the adjusted effect sizes in the second column have been corrected for estimation error by subtracting the mean variance of the estimators and recalculating the standard deviations. The reported effect sizes are comparable to those estimated

[12] We use heteroskedasticity-robust standard errors of the fixed effects, which are estimated with the Stata user-written command `fese`, written by Austin Nichols (2008).

in other settings (Chetty et al., 2011; Kane & Staiger, 2012) Comparison of the magnitudes of these effect estimates across subjects shows a considerably higher variance in the distribution of teacher quality in math relative to reading, consistent with the existing literature (e.g. Kane & Staiger, 2012; Lefgren and Sims, 2012).

In order to visually display the covariance of math and reading value-added, we estimate a bivariate kernel density in **Figure 1**.[13] For ease of illustration, we trim the (empirical Bayes adjusted) value-added sample of the top and bottom 1% of observations and standardize each measure to have mean zero and standard deviation 1. Shading on the figure represents the value of the density function, with darker regions having higher probability density. For instance, reading horizontally along the gridlines, the shading suggests the contour of the conditional distribution of math value-added given a constant value of reading value-added. The figure suggests substantial variability in math and reading value-added, but a clear positive correlation.

We use the joint kernel density estimate to construct conditional densities for select values of math and reading value-added in **Figure 2**. Each cell in Figure 2 represents a conditional distribution of reading (math) value-added evaluated at one of the quartiles of math (reading) value-added. Hence, Figure 2(a) depicts the conditional distribution of reading value-added evaluated at the 25th percentile of the math value-added distribution. Intuitively, it estimates the distribution of reading value-added for teachers assessed to be a 25th percentile math value-added teacher. Accordingly, most of the probability density is concentrated below the sample mean of the reading value-added distribution. This pattern is reflected in the other rows of Figure 2, as the distribution shifts to the right with higher levels of the conditioning value-added.

[13] We use a bivariate normal kernel. The kernel density estimates use the empirical Bayes estimates of value-added, which we also use in the regressions below. The correlations reported in the next section use the regression fixed effects coefficients and are corrected for sampling error using the estimated standard errors.

4.1. Correlation of Value-Added Across Math and Reading

Table 3 reports the cross-subject correlation of teacher effects. We find that the correlations across subjects are largely insensitive to the choice of model. For the full sample of data, the within-year, cross-subject correlations are about 0.6. For context, Papay (2011) finds correlations of value-added measures assessed from different tests *in the same year and subject* in the range of 0.3-0.55.

Comparisons of value-added measures within classrooms will reflect any teaching ability common to both subjects as well as the classroom effects. When we look across years to remove the contribution of classroom-level shocks, the correlations are approximately 0.35-0.4. However, even without any bias in the value-added estimation of the sort considered above, consecutive-year comparisons may also be an imperfect measure of the correlation in underlying effectiveness if there are real year-to-year variations in teacher effectiveness that are unrelated to experience or classroom dynamics (Goldhaber and Hansen, forthcoming). To put these correlations in perspective, therefore, they are of approximately the same magnitude as those reported for consecutive-year correlations in teacher effectiveness within the same subject.

As discussed above, these correlations represent the correlation between two noisy estimates of teacher effectiveness and thus not the correlation between teachers' "true" mathematics and reading effectiveness. After correcting for sampling error (see equation 4 above), the correlations across subjects are about 0.9 within the same year and 0.5-0.6 for consecutive years.

The correlations for the 5th grade sample are reported in Panel B of Table 3. The reported correlations for the 5th grade sample are smaller than those for the entire sample, but again remain stable across specifications, including the specification with the entire student test

history. The unadjusted correlations are approximately 0.5 within the same year and 0.3 for consecutive years. With the sampling error corrections, these correlations increase to approximately 0.8 within years and 0.4-0.5 for consecutive years. Together, these results suggest that assignment mechanisms are not responsible for the cross-subject correlations observed in panel A.

Figure 3 plots the within-year correlations in math and reading value-added across quantiles of the lagged test scores in math and reading. The unadjusted correlations are lower for classrooms with higher average baseline test scores, a finding that appears to reflect greater variability in the estimated coefficients, i.e. the adjusted correlations are stable across the distribution of baseline achievement. These results are suggestive that assignment rules are not driving the correlations reported in **Table 3**.

Correlations between value-added in consecutive years could overstate the correlation between persistent teaching effectiveness in math and reading if a significant portion of the observed variation in teaching skill is transitory (Goldhaber and Hansen, forthcoming) or reflects student-teacher sorting (Koedel and Betts, 2011; Rothstein, 2010). **Table 4** shows correlations of year *t* effectiveness in one subject with value-added in the other subject for years *t+1*, *t+2*, and *t+3*. Moving across the rows, the correlations do indeed decline, suggesting some of the correlations between successive years represents general teaching effectiveness that is transient. However, the differences in correlations between years *t+1* and *t+3* are modest and are similar to those estimated for the same subject across similar periods of time (Goldhaber and Hansen, forthcoming).

4.2. Robustness Tests

As we noted above, there is no consensus on the appropriate specification of VAMs, and research has raised concerns that teacher effectiveness estimates from value-added models are biased (Rothstein, 2010). Furthermore, if bias stemming from non-random assignments is constant within teachers, it may influence both within-year and across-year comparisons of teacher value-added estimates. Of particular concern is the possibility that students are assigned to teachers based on their achievement history and the teacher's known skill in teaching reading or mathematics. While we condition for lagged achievement in both subjects, principals may use a longer record of achievement when determining teacher assignments.

To test whether student assignments bias the cross-subject correlations in teacher effectiveness, we use the value-added measures estimated in (2) to predict student achievement in out-of-sample regressions. Because value-added estimates measure teacher effectiveness with error, we use empirical Bayes value-added estimates for all the regressions. Using empirical Bayes measures eliminates the attenuation bias due to measurement error (Jacob & Lefgren, 2008). The interpretation of the regression coefficients is now somewhat different than the correlations reported in the previous section. Consider the baseline regression

$$A_{ijst} = \alpha A_{is(t-1)} + X_{it}\gamma + \hat{\tau}_{j(t-1)}\,\rho + \epsilon_{ijst} \qquad (7)$$

which is identical to (2), except that we have replaced teacher indicators with teacher value-added estimated in the prior period. By the Frisch-Waugh-Lovell theorem, this regression is equivalent to a regression of student achievement on the residuals obtained from a regression of the value-added measure on the included covariates. Therefore, ρ measures the ratio of the covariance of residual value-added and student achievement to the variance of residual value-added. Hence, we test whether the results in the previous section are driven by correlated biases in value-added estimates by removing variation in estimated teacher effectiveness that is

correlated with contemporaneous predictors of student achievement. If student sorting explains the correlation between reading and math effectiveness, then removing these factors from the teacher value-added should attenuate the relationship between the cross-subject value-added and student achievement. If, on the other hand, the coefficients on estimated value-added remain constant, it suggests that the component of teaching effectiveness that is common across subjects is not related to observable student characteristics that were not included in the original value-added estimation model.

The benefit of this approach is that we can test different sets of assumptions to identify the relationship between math and reading value-added by adding different variables to Equation (7). Because we use reading value-added estimated in another set of years to predict student achievement in math, we assume that random classroom effects are uncorrelated with value-added estimated in prior years, conditional on different sets of covariates. However, note that this approach is conservative in the sense that correlation of value-added with student characteristics does not necessarily imply the value-added or the cross-subject correlations are biased.

Table 5 displays coefficients from regressions of the form discussed above. Because we test whether sorting on previous test score gains influences our results, we restrict our analysis to 4^{th} and 5^{th} grade teachers for whose students we have twice-lagged test scores. Panel A contains results for math achievement. The coefficient on reading value-added in the math achievement regression is 0.685 in column (5). As we add twice-lagged achievement, the coefficient falls to 0.672. Adding classroom characteristics and school-by-year fixed effects in columns (7) and (8), which are not included in the estimation of the value-added, reduces the coefficient to 0.620. Results are similar for the prediction of reading achievement with math value-added. We estimate a coefficient of 0.277 in column (1) of Panel B, which falls to 0.274 with the addition of

twice-lagged achievement, classroom characteristics, and school-by-year fixed effects. Further lags of student achievement, observable classroom characteristics, and school-by-year factors appear to explain virtually none of the relationship between math and reading effectiveness.

Recall from equation (7) that the coefficients reflect both the covariance between the cross-subject value-added and student achievement and the variance of value-added. Given the difference in the variance of value-added by subject, we should expect the coefficients to be of different magnitudes. When we account for the fact that math and reading value-added have different variances, the results that include twice-lagged achievement, classroom characteristics, and school-by-year fixed effects (Panel A, column 8 and Panel B, column 4) suggest a correlation of math and reading value-added of 0.35-0.50.[14]

Student sorting on observable characteristics explains very little of the covariance in math and reading effectiveness. This is despite the fact that there appears to be sorting on twice-lagged achievement. In **Table 6**, we regress students' twice-lagged achievement on their current teacher's value-added estimated in the previous year and all the covariates included in equation (2). The coefficients are generally positive and statistically significant, consistent with Rothstein (2010). We find that lagged gains in math are associated with current math and reading value-added and that lagged gains in reading are associated with current reading value-added. The coefficient on math value-added in the regression on twice-lagged reading achievement is nearly zero and not significant. The relationship between lagged achievement gains and current teacher effectiveness is shown graphically in **Figure 4**, which demonstrates that the coefficients are small in magnitude compared to the contemporary effects of teachers and, as the results in Table

[14] We multiply the coefficient in Table 5 by the square root of the residual value-added in the same subject and divide by the residual value-added in other subject. Both of these numbers are given in the bottom row of Table 5.

5 indicate, are not large enough to explain much of the covariance in math and reading value-added.

While we find that sorting on observable characteristics of students has modest impacts on the cross-subject correlations, it remains possible that sorting on unobserved characteristics of students generates a more substantial bias. We test for sorting on unobservables by following Chetty et al. (2011) and examine changes in cross-subject value-added that result from the movement of teachers across school-grade cells. For each school-grade-year cell, we generate a student-weighted difference in value-added of year *t* from year *t-1*. Both years' value-added are estimated in year *t-2*, so differences in value-added arise from changes in teacher staffing from one year to the next. We then use the differenced value-added in a regression with differenced mean achievement as the dependent variable. Because we average within school-grade-year cells, such an approach does not rely on non-random sorting within cells. Rather, the average change in student achievement should reflect the average change in value-added.

Table 7 shows the results of this test. The coefficient of reading value-added in the math achievement regressions is 0.583, compared to 0.685 in Table 5. The coefficient on math value-added in the reading achievement regressions is 0.355, compared to 0.277 in Table 5. These coefficients suggest correlations of math and reading value-added of approximately 0.35-0.6, which are broadly consistent with the results in **Tables 3** and **5**. Columns (3) and (4) of **Table 7** further show that changes in value-added in other grades in the same school have no predictive power for student achievement.

Finally, we use a method similar to one used in Harris and Sass (2011) and re-estimate the value-added on a sub-sample of data for which there appears to be random sorting of students to teachers. Specifically, we drop school-grade cells for which there are statistically significant

differences in any of the following variables: lagged math or reading score, free lunch eligibility, parental education, female, African American, and limited English proficiency. That is, we drop cases where it looks like the make-up of teachers' classrooms is not representative of the school as a whole in a given grade. **Table 8** presents analogous results to Table 5 for this sample. Again, the results are largely similar to those in **Table 5**, with estimated coefficients on reading value-added of 0.630 and on math value-added of 0.282 with twice-lagged achievement, classroom covariates, and school fixed effects.

On the whole, the robustness tests presented in **Tables 5-8** suggest that sorting on observable and unobservable student characteristics has very little effect on the correlation between math and reading effectiveness. We find that adding further lags of student achievement, classroom characteristics, and school-by-year fixed effects do not affect the relationship between estimated value-added and student achievement in the other subject. And quasi-experimental tests of bias in value-added of the kind suggested by Chetty et al. (2011) further confirm that the correlation of true math and reading value-added in proximate years is in the range of 0.35-0.5.

5. Summary and Conclusions

We began this paper by noting the policy interest in using VAM to estimate teacher performance and then using the resulting performance estimates to make high-stakes personnel decisions, such as determining job retention or pay. Policies relying on the use of value-added measures of teacher effectiveness depend on the accuracy, precision, and stability of these estimates. To date, there has been relatively little research on how teacher effectiveness varies across subject and whether value-added assessments are stable across subjects.

We find that correlations across subjects display approximately the same stability as those estimates across different years within the same subject. The results suggest that while there are differences in measured effectiveness across subjects, value-added estimates from mathematics and reading exams reflect a similar underlying measure of teaching effectiveness.

These patterns are documented in **Table 9**, which shows the tabulations of teachers by quintile of both math and reading value-added. Most of the weight in this table is along the diagonal, which is consistent with **Figure 1**. Across all cells, 75% of teachers are in the same or an adjacent quintile of teacher effectiveness across subjects. Teachers considered effective according to their value-added in one subject also tend to be effective in the other.

Thus far we have focused on the correlation between math and reading value-added. However, principals and administrators may be more concerned with the degree to which teacher effectiveness differs by subject. In middle and high schools, it is standard for teachers to specialize in one subject. This is less common in the elementary grades and does not occur in the self-contained classrooms we consider here. To give a sense of the magnitude of cross-subject variation within teachers, we conduct a simulation that assigns teachers to particular subjects based on their value-added. For each school-grade-year cell, we assign each teacher to teach two sections based on the difference between standardized math and reading value-added. For instance, in a school with three third grade classrooms, the teacher with the highest difference between math and reading value-added teaches two sections of math, the teacher with the lowest difference teaches two sections of reading, and the third teaches a section of each. To ensure the results do not reflect unusual year-to-year variation in teacher effectiveness, we average a teacher's value-added across all years and weight by the number of students.

While informative, this exercise is unlikely to estimate the true gains from specializing teachers at the elementary school level. Our assessments of teacher effectiveness are estimated from a system in which teachers must devote their lesson planning time to both subjects and cannot focus on specializing their instruction in one particular subject. Hence, simulations using value-added estimated from contained classrooms likely understate the gains from specialization. On the other hand, switching classrooms may prove disruptive or students may receive more individual attention when they have a single teacher. Nonetheless, this exercise provides policy-relevant context for the magnitude of our results.

We plot the kernel density estimators of the math and reading value-added in **Figure 5.** In both plots in **Figure 5**, allowing teachers to teach subject in which they are relatively better results in higher mean value-added. For mathematics, the difference in mean value-added is 0.19 standard deviations of teacher effectiveness, which corresponds to about 0.04 standard deviations of student achievement. For reading the mean difference is 0.18, which corresponds to 0.03 standard deviations of student achievement. These differences represent about 5-6% of the average annual gains in math achievement in grades 3-5 and 7-8% of the annual reading gains (Hill et al., 2008). Thus, while overall our findings tend to confirm that teachers who are effective in math also tend to be effective in reading, they also show that there may be meaningful gains associated with specialization at the elementary level.

References

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

Ballou, D. (2005). Value-added assessment: Lessons from Tennessee. In *Value-added Models in Education: Theory and Application*, ed. R. Lissitz. Maple Grove, MN: JAM Press.

Ballou, D., Mokher, C.G., and Cavaluzzo, L. (2012). Using value-added assessments for personnel decisions: How omitted variables and model specification influence teachers' outcomes. Unpublished manuscript.

Ballou, D., Sanders, W., and Wright, P. (2004). Controlling for student background in value-added assessment of teachers. *Journal of Educational and Behavioral Statistics*, 29(1), 37-66.

Boyd, D., Lankford, H., Loeb, S., and Wyckoff, J (2010). Teacher layoffs: An empirical illustration of seniority v. measures of effectiveness. CALDER working paper, July 20, 2010.

Chetty, R. and Friedman, J.N. and Rockoff, J. (2011). The long-term impacts of teachers: Teacher value-added and student outcomes in adulthood. National Bureau of Economic Research Working Paper #17699.

Clotfelter, C.T., Ladd, H., and Vigdor, J. (2006). Teacher-student matching and the assessment of teacher effectiveness. *The Journal of Human Resources,* 41(4), 778-820.

Darling-Hammond, L., Amrein-Beardsley, A., Haertel, E. and Rothstein, J. (2012). Evaluating teacher evaluation. *Phi Delta Kappan*, March, 2012.

Glazerman, S., Loeb, S., Goldhaber, D., Staiger, D., Raudenbush, S. and Whitehurst, G. (2010). Evaluating teachers: The important role of value-added. Brookings Institution.

Goldhaber, D. and 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. and Chaplin, D. (2012). Assessing the "Rothstein falsification test." Does it really show teacher value-added models are biased? CEDR Working Paper 2012-1.3. University of Washington, Seattle, WA.

Goldhaber, D., Gabele, B., and Walch, J. (2012). Does the model matter? Exploring the relationship between different student achievement-based teacher assessments. CEDR Working Paper 2012-6. University of Washington, Seattle, WA.

Goldhaber, D. and Hansen, M. (2010) Using performance on the job to inform teacher tenure decisions. *American Economic Review* 100(2), 250-255.

Goldhaber, D., and Hansen, M. (Forthcoming) Is it just a bad class? Assessing the long-term stability of estimated teacher performance. *Economica*.

Goldhaber, D., and Theobald, R. (2011) Managing the teacher workforce in austere times: The implications of teacher layoffs. CEDR Working Paper 2011-1.2. University of Washington, Seattle, WA.

Hanushek, E. (2009). "Teacher deselection." In *Creating a New Teaching Profession*, ed. Dan Goldhaber and Jane Hannaway. Washington, DC: Urban Institute Press.

Hanushek, E. and Rivkin, S. (2010). Generalizations about using value-added measures of teacher quality. *American Economic Review* 100, no. 2 (2010): 267-71.

Harris, D.N. (2009). Teacher value-added: Don't end the search before it starts. *Journal of Policy Analysis and Management*, 28(4), 693-699.

Harris, D.N. and Sass, T.R. (2011). Teacher training, teacher quality, and student achievement. *Journal of Public Economics*, 95, 798-812.

Hill, C.J., Bloom, H.S., Black, A.R., and Lipsey, M.W. (2008). Performance trajectories and performance gaps as achievement effect-size benchmarks for educational interventions. *Journal of Research on Educational Effectiveness, 1*(4), 289-328.

Hill, H. (2009). Evaluating value-added models: A validity argument approach. *Journal of Policy Analysis and Management*, 28(4), 700-709.

Jackson, C. and 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., and Lefgren, L. (2008). Can principals identify effective teachers? Evidence on subjective performance evaluation in education. *Journal of Labor Economics, 26*(1), 101-135.

Jacob, B. A., Lefgren, L., and Sims, D. P. (2010). The persistence of teacher-induced learning. *Journal of Human Resources*, *45*(4), 915-943.

Kane, T. and Staiger, D. O. (2008) Estimating teacher impacts on student achievement: An experimental evaluation. National Bureau of Economic Research Working Paper #14607.

Kane, T. and Staiger, D.O. (2012). *Gathering feedback for teaching.* Seattle, Washington: Bill and Melinda Gates Foundation.

Koedel, C. and Betts, J.R. (2007). Re-examining the role of teacher quality in the educational production function. Unpublished manuscript.

Koedel, C. and Betts, J.R. (2008). Value-added to what? How a ceiling in the testing instrument influences value-added estimation. National Center on Performance Incentives Working Paper 2008-21.

Koedel, C. and 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.

Konstantopoulos, S. (2007). How long do teacher effects persist? IZA Discussion Paper No. 2893. Bonn, Germany: Institute for the Study of Labor (IZA).

Lefgren, L. and Sims, D. (2012). Using subject test scores efficiently to predict teacher value-added. *Educational Evaluation and Policy Analysis*, 34(1), 109-121.

Lockwood, J.R., McCaffrey, D.F., Hamilton, L.S., Stecher, B.M., Le, V., and Martinez, F. (2007). The sensitivity of value-added teacher effect estimates to different mathematics achievement measures, *Journal of Educational Measurement*, 44(1), 47-67.

Nichols, A. (2008). fese: Stata module calculating standard errors for fixed effects. Available from http://ideas.repec.org/c/boc/bocode/s456914.html":http://ideas.repec.org/c/boc/bocode/s456914.html

Papay, J.P. (2011). Different tests, different answers. *American Educational Research Journal*, 48(1),163-193.

McCaffrey, D.F., Lockwood, J.R., Koretz, D., Louis, T., and 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. and Mihaly, K. (2009). The intertemporal variability of teacher effect estimates. *Education Finance and Policy*, 4(4), 572–606.

Nye, B., Konstantopoulos, S., and Hedges, L.V. (2004). How large are teacher effects? *Educational Evaluation and Policy Analysis* 26, 237-257.

Public Impact. (2012). *Redesigning schools to reach every student with excellent teachers: Financial planning for elementary subject specialization.* Chapel Hill, North Carolina.

Rivkin, S.G., Hanushek, E.A., and Kain, J.F. (2005). Teachers, schools, and student achievement. *Econometrica*, 73(2), 417-458.

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

Rothstein, J. (2010). Teacher quality in educational production: Tracking, decay, and student achievement. *Quarterly Journal of Economics*,125(1), 175–214.

Schochet, P. and Chiang, H. (2010). Error rates in measuring teacher and school performance based on student test score gains. *(NCEE 2010—4004).* Washington, D.C: National Center for Education Evaluation and Regional Assistance, Institute of Education Sciences, US Department of Education.

Staiger, D.O., and Rockoff, J.E. (2010). Searching for effective teachers with imperfect information. *Journal of Economic Perspectives*, 24(3), 97–118.

Todd, P.E., and Wolpin, K.I. (2003). On the specification and estimation of the production function for cognitive achievement. *The Economic Journal*, 113, F3-F33.

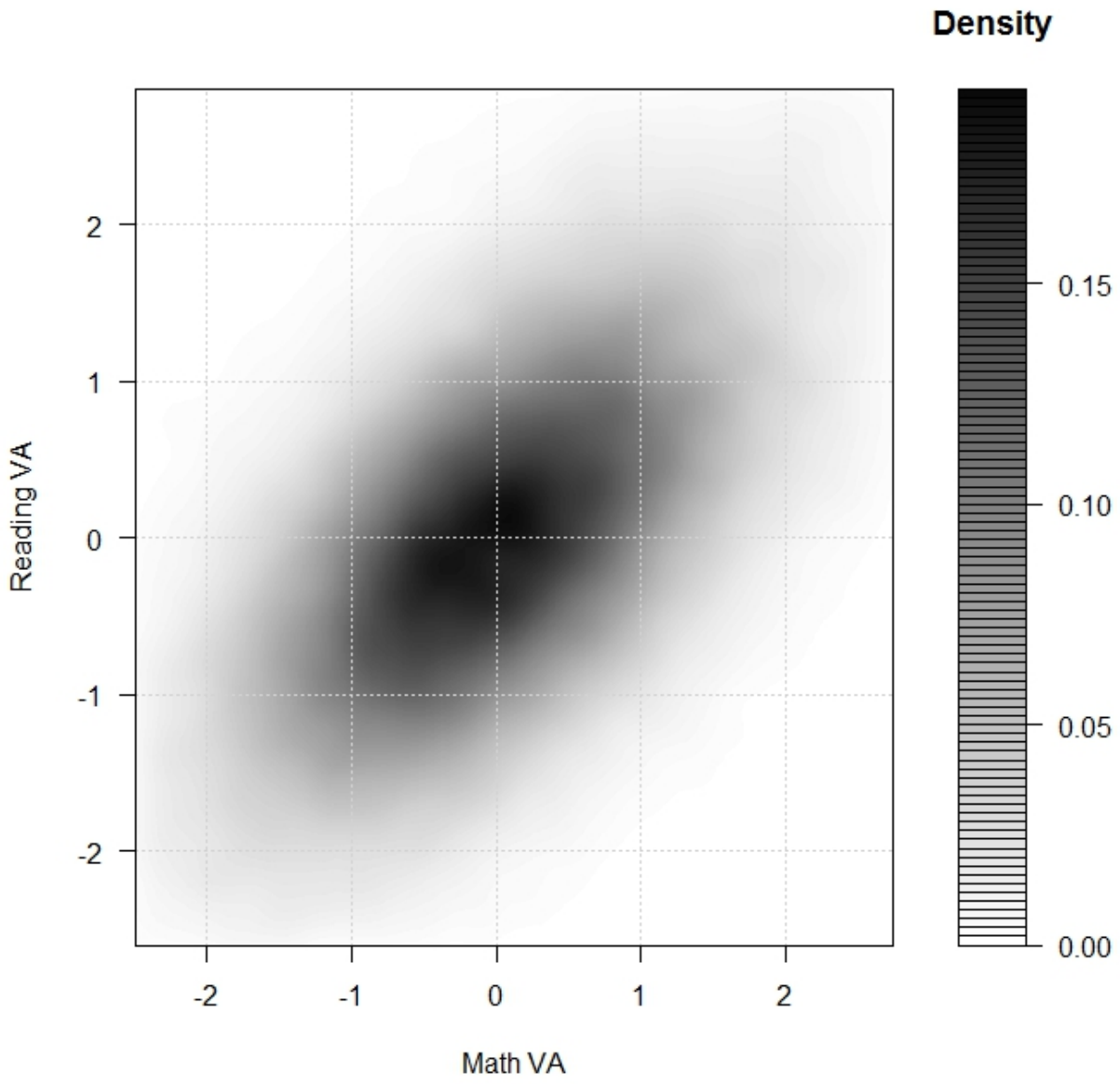
Value-Added Research Center (2010). *NYC teacher data initiative: technical report on the NYC value-added model.* University of Wisconsin-Madison.

Weisberg, D., Sexton, S., Mulhern, J. and Keeling, D. (2009). *The widget effect: Our national failure to acknowledge and act on differences in teacher effectiveness*. New York, NY: The New Teacher Project.

Wright, D. (2001). Student mobility: a negligible and confounded influence on student achievement. *Journal of Educational Research*, 92(6), 347-353.

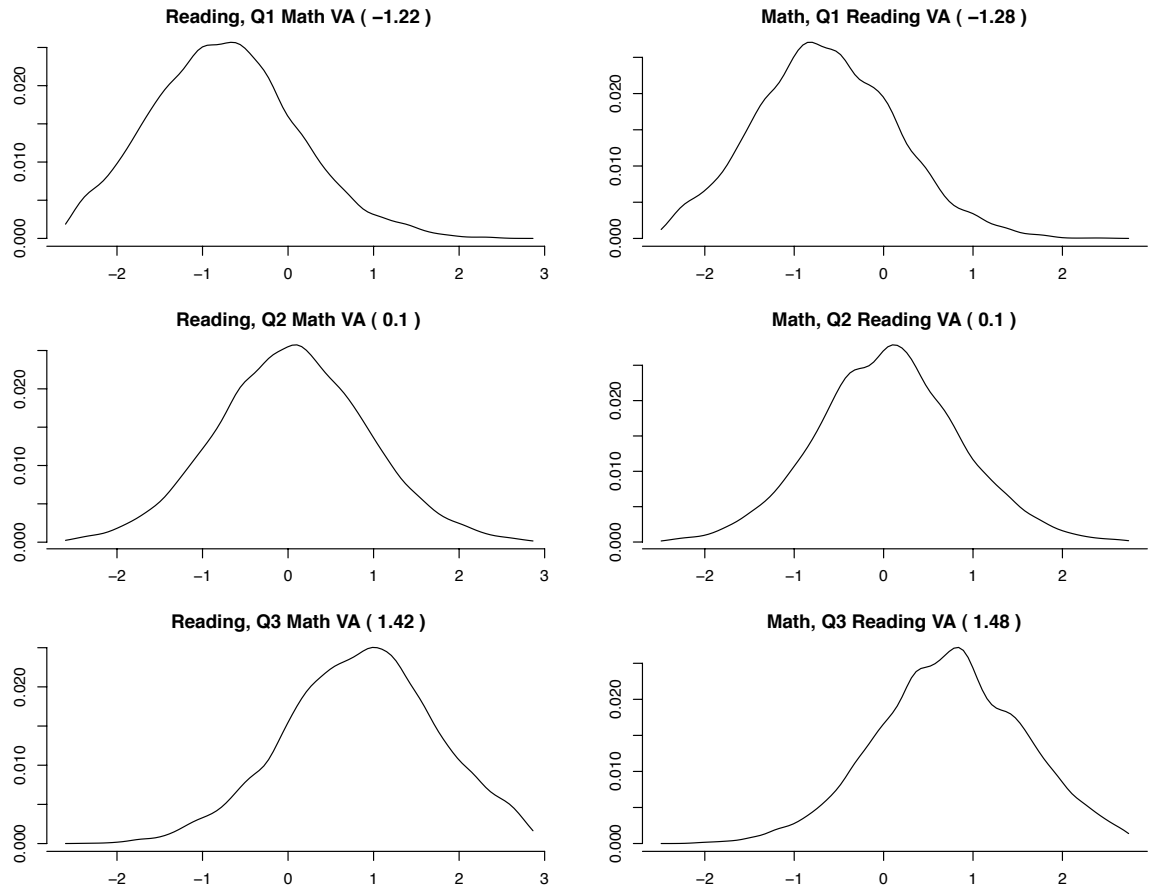
Xu, Z., Hannaway, J., and D'Souza, S. (2009). Student transience in North Carolina: The effect of school mobility on student outcomes using longitudinal data. CALDER Working Paper #22.

Figure 1: Kernel Density Estimate of the Joint Distribution of Math and Reading VA



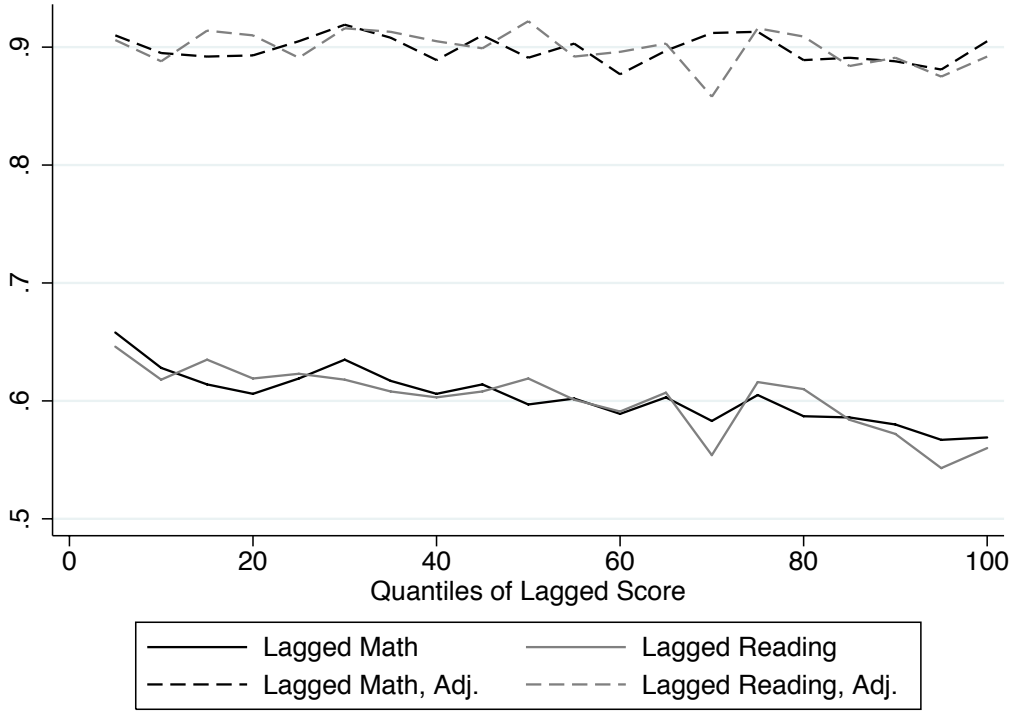
Notes: Figure depicts the joint distribution of math and reading empirical Bayes value-added. For this figure, math and reading value-added have been trimmed of the top and bottom 1% of observations and standardized to have a mean of 0 and standard deviation of 1. The joint distribution has been estimated with a bivariate normal kernel. The distribution has been estimated at 10,000 points on the support of math and reading value-added. Shading represents the value of the estimated distribution function.

Figure 2: Kernel Density Estimates of Conditional Distributions of Math and Reading VA



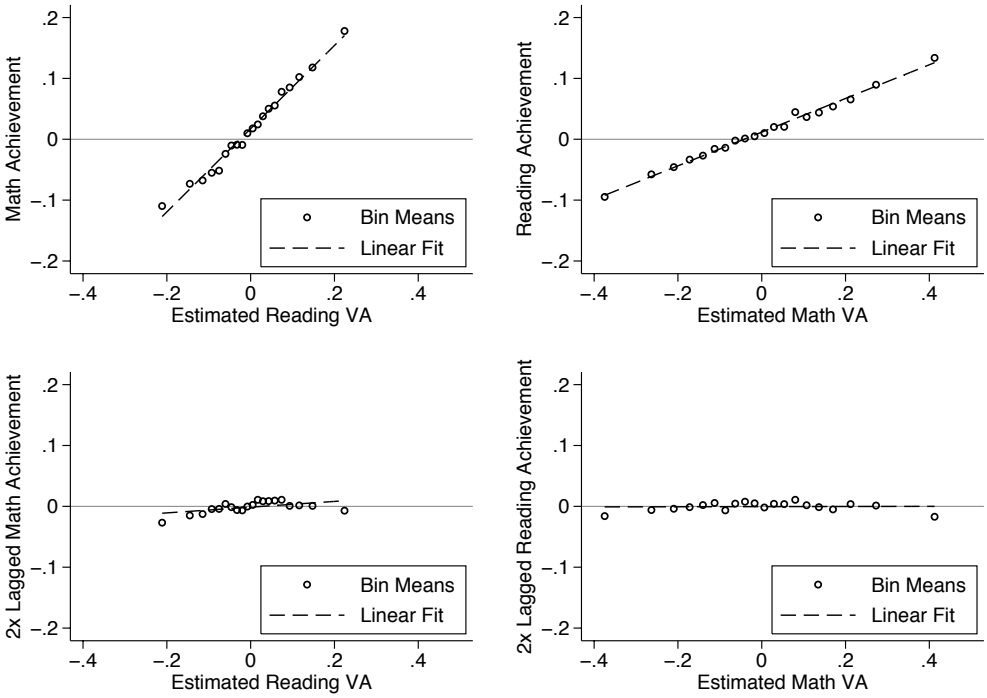
Note: Figures plot kernel density estimates of the conditional distribution of value-added derived from the joint distribution plotted in Figure 1. Value-added measures are empirical Bayes estimates censored at the top and bottom 1% standardized to have mean 0 and standard deviation 1. Left column plots conditional density of reading value-added at given quartile of the math value-added distribution. Right column plots conditional density of math value-added at given quartile of reading value-added distribution. Standardized empirical Bayes value-added at given quartile of conditioning distribution in parentheses.

Figure 3. Correlation between Math and Reading Value-Added by Quantile of Baseline Student Achievement



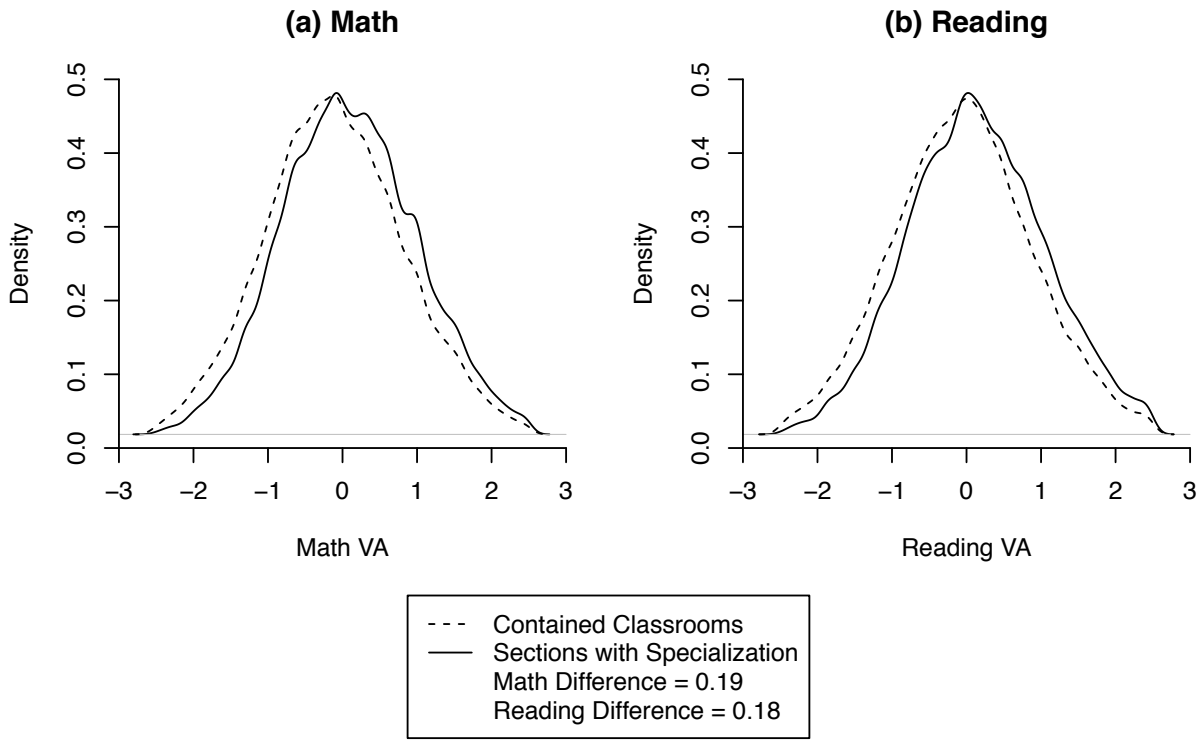
Notes: Figure depicts unadjusted and adjusted correlations between math and reading value-added calculated within each 5% interval of the math and reading baseline achievement distributions. Value-added estimated with lags in both math and reading.

Figure 4: Student Achievement and Cross-Subject Value-Added Estimated in Prior Year



Figures show residuals from math and reading achievement regressed on the control variables in the value-added model plotted against residuals of estimated value-added from the previous year regressed on the value-added controls. The residuals have been collapsed into 20 equally spaced bins based on quantiles of residual estimated value-added. The plotted lines are constructed from coefficients of regressions of residual achievement on residual value-added using the student-level data. These regressions are identical to those reported in Panel A, column (5) and Panel B, column (1) of Table 5 and columns (2) and (3) of Table 6.

Figure 5: Distribution of Teacher Value-Added with Contained Classrooms and Specialized Sections



Notes: Figures display kernel density estimates of estimated value-added. VA measures are averages over all available years. Contained classrooms are classrooms as they exist in the data. Sections assign teachers to teach one subject based on estimated VA as described in the text. Teacher VA has been standardized to have mean 0 and standard deviation 1.

Table 1. Descriptive Means and Standard Deviations

		Included	Excluded	Difference
Panel A: Students				
Female	.49	.496	.486	.00956***
	(.5)	(.5)	(.5)	(.000564)
Parental education: BA+	.148	.163	.136	.0268***
	(.355)	(.37)	(.343)	(.000404)
Black	.289	.289	.29	-.00014
	(.454)	(.453)	(.454)	(.000511)
Hispanic	.06	.0448	.0712	-.0264***
	(.237)	(.207)	(.257)	(.000259)
Other non-white	.0559	.0513	.0593	-.00799***
	(.23)	(.221)	(.236)	(.000256)
Free-lunch eligible	.395	.43	.369	.0606***
	(.489)	(.495)	(.483)	(.000552)
Learning disability: math	.0246	.0212	.0271	-.00589***
	(.155)	(.144)	(.162)	(.000171)
Learning disability: reading	.0502	.0461	.0533	-.00712***
	(.218)	(.21)	(.225)	(.000244)
Learning disability: writing	.0465	.0442	.0483	-.00406***
	(.211)	(.206)	(.214)	(.000236)
Limited English proficiency	.029	.0231	.0334	-.0103***
	(.168)	(.15)	(.18)	(.000184)
Std. math score	0	.0579	-.105	.163***
	(1)	(.977)	(1.03)	(.00145)
Std. reading score	0	.0455	-.0838	.129***
	(1)	(.977)	(1.04)	(.00146)
N	2113342	1369790	743552	
Panel B: Teachers				
Female	.924	.926	.913	.0127*
	(.266)	(.262)	(.281)	(.00735)
Black	.148	.145	.158	-.0131
	(.355)	(.352)	(.365)	(.0096)
Hispanic	.00319	.00301	.00391	-.000902
	(.0564)	(.0548)	(.0624)	(.00161)
Other non-white	.0131	.0107	.0223	-.0116***
	(.114)	(.103)	(.148)	(.0037)
Approved NC training program	.352	.355	.338	.0177
	(.477)	(.479)	(.473)	(.0127)
Experience	14.7	14.6	15.3	-.692***
	(9.36)	(9.37)	(9.27)	(.249)
Fully licensed	.0391	.0262	.0893	-.0631***
	(.3)	(.237)	(.467)	(.0114)
Master's degree	.294	.284	.332	-.0486***
	(.455)	(.451)	(.471)	(.0125)
N	8604	6847	1757	

Notes: Standard deviations (last column: standard error of t-test) in parentheses. In Panel A, observations are student-years. In Panel B, observations are teachers for 2005 (last year of sample).

Table 2: Standard Deviation of Teacher Value-added Estimates

	Unadjusted		Adjusted	
	Math	Reading	Math	Reading
Panel A: All Teachers				
Lagged score	.259	.215	.231	.170
Lagged score, both subjects	.252	.199	.225	.155
Lagged scores & student covariates	.244	.191	.217	.147
N	67967	67967	67967	67967
Panel B: 5th Grade Teachers				
Lagged score	.237	.187	.213	.143
Lagged score, both subjects	.231	.177	.207	.134
Lagged scores & student covariates	.226	.172	.202	.128
Full test history	.226	.173	.198	.120
N	21045	21045	21045	21045

Notes: Adjusted effect sizes have been calculated by subtracting the student-weighted means of the teacher effects standard errors from the sample variance of the teacher effects.

Table 3: Cross-Subject Correlations

	Corr. w/in Years		Corr. across Years			
			Math(t), Reading(t+1)		Math(t+1), Reading(t)	
	Unadj.	Adj.	Unadj.	Adj.	Unadj.	Adj.
Panel A: All Teachers						
Lagged score	.612	.871	.408	.578	.399	.559
Lagged score, both subjects	.629	.908	.374	.539	.378	.538
Lagged scores & student covariates	.609	.897	.35	.517	.353	.516
N	67768	67768	43149	43149	43149	43149
Panel B: 5th Grade Teachers						
Lagged score	.536	.784	.347	.507	.332	.482
Lagged score, both subjects	.547	.810	.302	.449	.308	.455
Lagged scores & student covariates	.528	.799	.28	.427	.285	.433
Full test history	.48	.793	.275	.462	.283	.46
N	20983	20983	13442	13442	13442	13442

Notes: Adjusted correlations have been calculated by subtracting the student-weighted means of the teacher effects standard errors from the sample variance of the teacher effects.

Table 4: Cross-Subject Correlations in Other Years

	t+1		t+2		t+3	
	Unadj.	Adj.	Unadj.	Adj.	Unadj.	Adj.
Panel A: All Teachers						
Math (t)	.344	.512	.315	.472	.28	.426
Reading (t)	.345	.505	.322	.469	.299	.438
N	20480	20480	20480	20480	20480	20480
Panel B: 5th Grade Teachers						
Math (t), Lagged scores & covariates	.281	.421	.261	.394	.229	.357
Reading (t), Lagged scores & covariates	.283	.416	.29	.428	.263	.39
Math (t), Full test history	.296	.48	.242	.397	.217	.371
Reading (t), Full test history	.292	.46	.283	.444	.267	.429
N	6271	6271	6271	6271	6271	6271

Table 5: Out-of-sample Predictions of Cross-Subject Value-Added

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Panel A. Math Achievement								
Math VA	0.671***	0.667***	0.664***	0.632***				
	(0.00757)	(0.00738)	(0.00738)	(0.00721)				
Reading VA					0.685***	0.672***	0.662***	0.620***
					(0.0155)	(0.0150)	(0.0150)	(0.0149)
Lagged math score	0.666***	0.516***	0.516***	0.516***	0.672***	0.522***	0.522***	0.510***
	(0.00144)	(0.00159)	(0.00159)	(0.00151)	(0.00148)	(0.00167)	(0.00166)	(0.00155)
Lagged reading score	0.169***	0.0985***	0.0983***	0.0966***	0.166***	0.0962***	0.0960***	0.0985***
	(0.00134)	(0.00142)	(0.00142)	(0.00135)	(0.00138)	(0.00147)	(0.00147)	(0.00137)
2x lagged math score		0.239***	0.239***	0.241***		0.239***	0.239***	0.245***
		(0.00142)	(0.00142)	(0.00134)		(0.00148)	(0.00148)	(0.00137)
2x lagged reading score		0.0371***	0.0370***	0.0365***		0.0364***	0.0362***	0.0367***
		(0.00124)	(0.00124)	(0.00118)		(0.00128)	(0.00128)	(0.00120)
N	527108	527108	527108	527108	527108	527108	527108	527108
Panel B. Reading Achievement								
Math VA	0.277***	0.276***	0.271***	0.274***				
	(0.00650)	(0.00623)	(0.00621)	(0.00622)				
Reading VA					0.546***	0.535***	0.526***	0.476***
					(0.0116)	(0.0109)	(0.0109)	(0.0111)
Lagged math score	0.238***	0.148***	0.147***	0.152***	0.241***	0.150***	0.150***	0.150***
	(0.00137)	(0.00157)	(0.00157)	(0.00152)	(0.00135)	(0.00157)	(0.00156)	(0.00152)
Lagged reading score	0.566***	0.430***	0.430***	0.425***	0.563***	0.428***	0.428***	0.426***
	(0.00150)	(0.00155)	(0.00155)	(0.00153)	(0.00149)	(0.00155)	(0.00155)	(0.00153)
2x lagged math score		0.0704***	0.0700***	0.0707***		0.0702***	0.0699***	0.0715***
		(0.00148)	(0.00148)	(0.00143)		(0.00147)	(0.00147)	(0.00142)
2x lagged reading score		0.221***	0.221***	0.221***		0.221***	0.221***	0.221***
		(0.00135)	(0.00135)	(0.00135)		(0.00135)	(0.00135)	(0.00134)
N	527108	527108	527108	527108	527108	527108	527108	527108
Student chars	Y	Y	Y	Y	Y	Y	Y	Y
Class chars	N	N	Y	Y	N	N	Y	Y
School-by-year FE	N	N	N	Y	N	N	N	Y
Residual VA variance	0.0351	0.0351	0.0348	0.0221	0.0106	0.0106	0.0105	0.00697

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$. Standard errors clustered by school-cohort in parantheses. EB Math or Reading VA estimated in prior year is independent variable. All regressions include all variables used in value-added estimation model. Residual VA variance gives sample variance of residuals from regression of value-added on all other covariates included in regression.

Table 6: Sorting on Twice-Lagged Student Achievement

	(1) Math	(2) Math	(3) Reading	(4) Reading
Math VA	0.0166**		0.00108	
	(0.00729)		(0.00637)	
Reading VA		0.0471***		0.0343***
		(0.0131)		(0.0121)
Lagged math score	0.595***	0.595***	0.220***	0.220***
	(0.00154)	(0.00154)	(0.00154)	(0.00154)
Lagged reading score	0.210***	0.210***	0.544***	0.543***
	(0.00149)	(0.00149)	(0.00163)	(0.00163)
N	527108	527108	527108	527108

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$. Standard errors clustered by school-cohort in parantheses. EB Math or Reading VA estimated in prior year is independent variable. All regressions include all variables used in value-added estimation model. Test indicated in column title.

Table 7: Teacher Mobility and Changes in Student Achievement

	(1)	(2)	(3)	(4)
Panel A. Differenced Math Achievement				
Diff. avg. math VA	0.584***			
	(0.0454)			
Diff. avg. reading VA		0.583***		
		(0.0755)		
Diff. avg. math VA, other grades			0.0310	
			(0.0601)	
Diff. avg. reading VA, other grades				0.0818
				(0.0938)
N	11827	11827	11505	11505
Panel B. Differenced Reading Achievement				
Diff. avg. math VA	0.355***			
	(0.0385)			
Diff. avg. reading VA		0.518***		
		(0.0654)		
Diff. avg. math VA, other grades			0.0128	
			(0.0538)	
Diff. avg. reading VA, other grades				0.0889
				(0.0835)
N	11827	11827	11505	11505
Residual VA variance	0.00526	0.00194	0.00265	0.000984

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$. Standard errors clustered by school-cohort in parantheses. Dependent variable is first differenced mean student achievement by grade and school. Independent variable is first differencedestimated value-added by grade and school (rows 1-2) or for other grades in the same school (rows 3-4). Both VA measures in difference are estimated in year t-2. All regressions also include year effects. Residual VA variance gives the sample variance of the residuals of a regression of math value-added (columns 1,3) or reading value-added (columns 2,4) on year effects.

Table 8: Out-of-sample Predictions of Cross-Subject Value-Added (Sample of Classrooms with Apparently Random Assignment)

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Panel A. Math Achievement								
Math VA	0.678***	0.675***	0.673***	0.630***				
	(0.0106)	(0.0104)	(0.0104)	(0.0112)				
Reading VA					0.691***	0.676***	0.671***	0.622***
					(0.0219)	(0.0216)	(0.0216)	(0.0222)
Lagged math score	0.668***	0.518***	0.518***	0.517***	0.674***	0.524***	0.524***	0.513***
	(0.00202)	(0.00227)	(0.00227)	(0.00215)	(0.00209)	(0.00239)	(0.00238)	(0.00220)
Lagged reading score	0.169***	0.0994***	0.0992***	0.0976***	0.166***	0.0968***	0.0966***	0.0988***
	(0.00185)	(0.00199)	(0.00198)	(0.00192)	(0.00190)	(0.00205)	(0.00204)	(0.00194)
2x lagged math score		0.237***	0.238***	0.241***		0.237***	0.237***	0.243***
		(0.00209)	(0.00208)	(0.00199)		(0.00218)	(0.00217)	(0.00202)
2x lagged reading score		0.0363***	0.0363***	0.0365***		0.0353***	0.0353***	0.0366***
		(0.00175)	(0.00175)	(0.00168)		(0.00181)	(0.00181)	(0.00170)
N	260163	260163	260163	260163	260163	260163	260163	260163
Panel B. Reading Achievement								
Math VA	0.287***	0.288***	0.284***	0.282***				
	(0.00921)	(0.00883)	(0.00883)	(0.00996)				
Reading VA					0.565***	0.555***	0.549***	0.490***
					(0.0171)	(0.0162)	(0.0163)	(0.0174)
Lagged math score	0.241***	0.150***	0.150***	0.155***	0.243***	0.153***	0.153***	0.153***
	(0.00192)	(0.00221)	(0.00221)	(0.00215)	(0.00190)	(0.00219)	(0.00219)	(0.00214)
Lagged reading score	0.566***	0.428***	0.428***	0.423***	0.563***	0.426***	0.426***	0.424***
	(0.00209)	(0.00225)	(0.00224)	(0.00222)	(0.00208)	(0.00224)	(0.00223)	(0.00222)
2x lagged math score		0.0697***	0.0695***	0.0702***		0.0693***	0.0691***	0.0705***
		(0.00209)	(0.00209)	(0.00204)		(0.00209)	(0.00208)	(0.00203)
2x lagged reading score		0.224***	0.224***	0.224***		0.224***	0.224***	0.224***
		(0.00199)	(0.00199)	(0.00198)		(0.00199)	(0.00199)	(0.00198)
N	260163	260163	260163	260163	260163	260163	260163	260163
Student chars	Y	Y	Y	Y	Y	Y	Y	Y
Class chars	N	N	Y	Y	N	N	Y	Y
School-by-year FE	N	N	N	Y	N	N	N	Y
Residual VA variance	0.0353	0.0353	0.0350	0.0178	0.0107	0.0107	0.0106	0.00586

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$. Standard errors clustered by school-cohort in parantheses. EB Math or Reading VA estimated in prior year is independent variable. All regressions include all variables used in value-added estimation model. Residual VA variance gives sample variance of residuals from regression of value-added on all other covariates included in regression.

Table 9: Frequency by Quintiles of Yearly Math and Reading Value-Added

	Reading VA					
Math VA	1	2	3	4	5	Total
1	9.7	5.1	2.9	1.6	0.6	19.9
2	5.1	5.4	4.6	3.2	1.5	19.8
3	3.0	4.6	5.0	4.5	2.8	19.9
4	1.5	3.2	4.6	5.4	5.0	19.7
5	0.5	1.5	2.8	5.2	9.9	19.9
Total	19.8	19.8	19.9	19.9	19.8	

Notes: Each cell contains percentage of teacher-year observations with given quintile of math and reading VA.