Working Paper 64

MATHEMATICA
Policy Research

Working PAPER

BY ELIAS WALSH, DALLAS DOTTER, AND ALBERT Y. LIU

Can More Teachers Be Covered? The Accuracy, Credibility, and Precision of Value-Added Estimates with Proxy Pre-Tests

August 2018

ABSTRACT

Value-added models used for evaluating teachers typically rely on controls for previous-grade student achievement to isolate teachers' contributions to students' current achievement. As a consequence, these models are most commonly used for subjects in which students are tested in consecutive grades. However, states and districts have little information about how value-added models perform in grades when tests in the same subject are not available from the previous year. In those grades, *proxy pre-tests*—prior-grade test scores from other subjects—are often used as controls. Using proxy pre-tests could allow states and districts to increase the number of teachers for whom value-added models can be used (for example, by including science teachers, rather than only teachers of math and English language arts). In this paper, we use statewide data from Oklahoma to investigate whether value-added models that rely on proxy pre-tests can credibly, accurately (with limited bias), and precisely measure teachers' contributions to student achievement. We find that not incorporating same-subject pre-tests affects value-added estimates much more than does excluding other student background characteristics. This difference appears to be a result of more bias in the proxy pre-test estimates rather than less precision. Despite evidence of bias, we discuss how these estimates may still reflect important information about a teacher's effectiveness. We also note that empirical Bayes shrinkage, an approach typically used to address precision, might also be used to address bias so that value-added estimates that rely on proxy pre-tests can be given more appropriate weights in teachers' evaluations.

This working paper would not have been possible without the support of the Oklahoma State Department of Education (OSDE). We are especially grateful to Kerri White, Susan Pinson, and Jason Perez at OSDE, and Jackie Skapik at UPD. We are also grateful to Duncan Chaplin for his helpful comments. Clare Wolfendale, Matthew Jacobus, and Lisa McCusker provided excellent programming support.

I. INTRODUCTION

Many state and district-wide systems for evaluating teacher effectiveness have incorporated value-added models that use statistical methods to isolate a teacher's contribution to student achievement. However, these models can only be used for teachers of students in grades and subjects in which standardized tests are given. Furthermore, value-added models must include previous test scores to account for how students were performing when they entered a teacher's classroom. Therefore, students must also have been tested in the previous school year, ideally in the same subject as the test used to measure achievement. As a consequence of these requirements, it is typical for fewer than a quarter of teachers in a district to have evaluations that include value-added estimates. For example, in the 2012–2013 school year, only 14 percent of teachers in the District of Columbia Public Schools had them (Walsh and Dotter 2014). Often, this small percentage includes only teachers of math and English language arts (ELA) in grades 4 to 8 in which testing occurs annually, although some states also estimate value added for some high school teachers using end-of-course exams.

Some states or districts may be interested in increasing the number of teachers who receive value-added estimates, but one approach to doing so—expanding testing to additional grades or subjects—can be costly and controversial. When expanding testing is not an option, additional teachers can be given value-added estimates by relaxing some presumed data requirements for value-added models. Rather than requiring a previous test score in the same subject as the test used to measure achievement (which we refer to as the *pre-test*), previous test scores in other subjects (or proxy pre-tests) might be used to account for students' skills before entering a teacher's classroom. For example, Oklahoma tests students in science in grades 5 and 8, but not in grades 4 or 7. Instead of accounting for grade 7 science performance when measuring the contributions to science achievement made by a grade 8 science teacher, a value-added model might rely on math and ELA achievement in grade 7. However, if these grade 7 proxy pre-test scores are not sufficiently related to science skills at the end of grade 8, then the value-added model could produce biased and imprecise estimates of teachers' contributions. Additionally, accounting for scores on the grade 5 science test from three years earlier would be unlikely to resolve this concern because tests in earlier grades (1) have weaker relationships with contemporaneous achievement, (2) would not account for the contributions of a student's teachers in grades 6 and 7, and (3) would exclude students who do not have test scores for both grades 5 and 8 due to mobility during the three school years.

This paper investigates whether value-added models that rely on previous test scores in other subjects can credibly, accurately (that is, with limited bias), and precisely measure teachers' contributions to student achievement. Using data from the state of Oklahoma, we explore three research questions about value-added estimates that do not account for a test score in the same subject:

1. **Credibility**: Compared to a value-added model with a pre-test in the same subject as the outcome, are the relationships between prior achievement and outcomes in a value-added model with only proxy pre-tests strong enough to credibly account for students' prior achievement?

2. **Accuracy**: How is the accuracy of value-added estimates that include pre-tests from multiple subjects affected by omitting the same-subject pre-test? In answering this question, we assume that the value-added model with the pre-test and proxy pre-tests is unbiased.
3. **Precision**: How does the precision of value-added estimates from models that use only proxy pre-tests compare to those that also include a same-subject pre-test?

We find that proxy pre-tests less credibly predict achievement compared a value-added model that also includes the same-subject pre-test. We also find substantial consequences from excluding the same-subject pre-test from a value-added model. Omitting the pre-test produces larger differences than those typically found for other changes to value-added models, such as excluding background characteristics, and may introduce bias that accounts for 11 to 12 percent of the dispersion in the resulting estimates. Finally, value-added estimates that rely on the proxy pre-tests are only slightly less precise than those that also include the same-subject pre-test.

II. BACKGROUND

Among the states and districts that have used growth or value-added models to generate results used in teachers' evaluations, many have focused on teachers of math and reading because their students are tested in these subjects in consecutive grades. Additionally, some districts have provided these results to teachers, even when test scores are not available in consecutive grades in a subject. For example, Charleston County School District, Los Angeles Unified School District, Pittsburgh Public Schools, and Tulsa Public Schools calculated value-added estimates for science teachers in some grades based on models that included prior scores only from math and reading (Resch and Deutsch 2015; Value-Added Research Center 2013a; Rotz et al. 2014; Value-Added Research Center 2013b).

Value-added estimates based on models that include only proxy pre-tests may be less accurate than value-added models with a same-subject pre-test, but we are not aware of previous analyses that seek to understand the quality of these value-added estimates. However, several studies have examined the consequences of excluding some prior test scores from a value-added model. For example, Goldhaber et al. (2014) exclude prior scores in the "opposite" subject (for example, excluding prior reading scores from the model for math teachers) and find correlations above 0.92 with models that include those scores. Johnson et al. (2015) include test scores in the same subject from both of the previous two years and find correlations above 0.96 with models that include only test scores from the previous year. Using a quasi-experimental approach to measure bias in value-added estimates, Chetty et al. (2014) found that excluding prior scores in the opposite subject led to 5 percentage points more bias compared to estimates from a model using the prior scores they identified as including 5 percent bias.[1] It is reasonable to hypothesize that the consequences of excluding prior scores in the same subject would be larger.

[1] We selected these value-added models from several that Chetty et al. (2014) reported on to isolate the impact of excluding the test scores. The benchmark model for this comparison included prior scores in two subjects, but did not include the other student characteristics that Chetty et al. (2014) included in their primary value-added model. The 10 percent bias result for the model with only the same-subject pre-test scores was statistically significant, but

However, whether to use biased value-added estimates in evaluations depends on the quality of results from the alternative approach. When a growth model or value-added estimate is unavailable for a teacher, a state or district may use a separate evaluation component to evaluate teachers without value-added estimates, or increase the weight on one or more components that contribute to all teachers' evaluation results to replace the absent value-added estimates.[2] The amount of bias that might be acceptable will depend on the properties of these other evaluation measures.

Thus, an important contribution of our paper is to examine the consequences of excluding prior test scores from the same subject in a value-added model. This contribution is particularly important because states and districts have used estimates from value-added models that do not include these prior scores in teachers' evaluations. Our results can be used to inform decisions about whether to use value-added estimates with this limitation or replace them with other evaluation measures.

III. METHODS

A. Data

We use administrative data from the state of Oklahoma, provided by the Oklahoma State Department of Education (OSDE), to estimate value added for teachers of math, reading, and science classes during the 2013–2014 school year. The data cover all districts in the state and include state standardized test scores, student background characteristics, links between teachers and students, and students' school enrollment.

We analyzed students' 2014 test scores from the Oklahoma Core Curriculum Tests (OCCTs) in math, reading, and science. We refer to these test scores as *post-test scores*. We also used 2013 scores in the same three subjects, which we refer to as *pre-test scores*. To be included in our analysis, students' test score records had to meet certain conditions based on when they were tested. Students enrolled in grades 4 through 8 during the 2013–2014 school year were eligible to be included if they had an OCCT math or reading post-test score and a pre-test score in the same subject. Approximately 13 percent of students with 2014 test scores in math or reading did not have a pre-test score in the same subject. For the science value-added model, we included students enrolled in grades 5 and 8 during the 2013–2014 school year if they had both an OCCT math and reading pre-test score. We excluded a small number of students who had conflicting test score records for the same subject and school year or who had scores that were not in the

neither the 5 percent bias in the benchmark model nor the 2 percent bias that Chetty et al. (2014) found in their primary model were statistically significant.

[2] As examples of the first approach, the states of Louisiana and New York have used an alternative measure of teachers' contributions to student growth based on whether the teacher meets goals he or she sets for the year (often called student learning objectives). Examples of the second approach include the state of New Jersey and the District of Columbia Public Schools, which assigned a classroom observation measure a larger weight in evaluations for teachers without value-added estimates to account for the weight given to value-added estimates for those teachers who have them.

valid range for the subject. We standardized the test scores from each subject, grade, and year by subtracting the mean score and dividing by the student-level standard deviation.

For our three research questions, we focused on separate analyses for each post-test grade and subject. Table 1 summarizes the available pre- and post-test data by grade and subject, and indicates which data we used to examine each of our research questions. To examine our first research question about credibility, we used post-test scores from grade 6 in math and reading. Grade 6 is the only grade for which we could estimate a value-added model that included pre-test scores in math, reading, and science. When excluding a same-subject pre-test from a grade 6 value-added model, two proxy pre-tests remain in the model as would typically be the case in a value-added model for a subject such as science or social studies for which only math and reading pre-tests are available in many states. We also used post-test scores in math, reading, and science from grades 5 and 8, allowing us to estimate value-added models in all three subjects for these grades, but with only math and reading scores as pre-tests. To examine our second and third research questions about accuracy and precision, we again used post-test scores in math, reading, and science from grade 6 so that we have pre-tests in all three subjects. For the third research question, we also report on the precision of the science value-added estimates for teachers of students in grades 5 and 8.

Table 1. Available state assessment scores in Oklahoma for calculating value added during the 2013–2014 school year

	Are prior test scores available?			
Post-test subject and grade	Math	Reading	Science	Used in our analysis?
Panel A: Math				
Grade 4	Yes	Yes	No	No
Grade 5	Yes	Yes	No	Research question 1
Grade 6	Yes	Yes	Yes	Research questions 1, 2, and 3
Grade 7	Yes	Yes	No	No
Grade 8	Yes	Yes	No	Research question 1
Panel B: Reading				
Grade 4	Yes	Yes	No	No
Grade 5	Yes	Yes	No	Research question 1
Grade 6	Yes	Yes	Yes	Research questions 1, 2, and 3
Grade 7	Yes	Yes	No	No
Grade 8	Yes	Yes	No	Research question 1
Panel C: Science				
Grade 5	Yes	Yes	No	Research questions 1 and 3
Grade 8	Yes	Yes	No	Research questions 1 and 3

We made additional sample exclusions to avoid including highly imprecise value-added estimates in our analysis. We excluded 1 percent of tested students from the analysis because they were not linked to a teacher eligible to be in the value-added model. To be eligible, a teacher had to teach at least 10 students in the grade and subject corresponding to the value-added model. Because the results would be imprecise, we do not estimate a value-added measure for teachers with so few students contributing to the estimate.

Ultimately, we use value-added estimates in grade 5 for 1,598 math teachers, 1,922 reading teachers, and 1,445 science teachers. For grade 6, we use estimates for 1,110 math teachers and 1,469 reading teachers. For grade 8, we use estimates for 937 math, 1,229 reading, and 763 science teachers. Depending on the grade and subject, these value-added models included test scores from between 30,285 and 42,127 students. Descriptive statistics for the students who meet requirements to be included in the value-added models are reported in Walsh et al. (2015).

In addition to prior test scores, we accounted for a set of student background characteristics in the value-added models. These include whether the student was eligible for free or reduced-price lunch, gender, race and ethnicity, whether the student had an individual education plan with or without accommodations, whether the student had limited English proficiency with or without accommodations, whether the student transferred schools during the 2013–2014 school year, and the fraction of the 2012–2013 school year the student attended school.

We imputed data for students who were included in the analysis file but had missing values for one or more student characteristics. Because some districts did not provide the state with student-level attendance records from the 2012–2013 school year, we used the typical attendance rate from the student's school and grade in place of the student's individual attendance rate for 45 percent of students in the analysis.[3] For other characteristics, including pre-test scores in subjects different from the post-test, our imputation approach used non-missing values of the characteristic, as well as non-missing values of other student characteristics to predict the value of the missing characteristic. Fewer than 3 percent of students in the value-added analysis files had a characteristic besides attendance imputed. Most imputed values were for missing pre-test scores in content areas different from the OCCT post-test.

Finally, we measured the proportion of instructional time a teacher spent with a student during the school year using data from the state's pilot roster verification program or information on the time periods students were enrolled in each of their schools. In the 2012–2013 school year, OSDE implemented a pilot roster verification program in selected schools, which was expanded to additional schools in the 2013–2014 school year. Roster verification is a process by which records of teachers' monthly shares of instruction for each student and course are submitted and either verified or corrected by teachers and school administrators. In recording their share of instructional time with a student, teachers rounded to the nearest quarter. Thus, 0, 25, 50, 75, and 100 percent were the possible responses. After these shares were reported, they were verified or corrected by those teachers and school administrators. The roster verification process differed slightly in Tulsa Public Schools, where teachers rounded to the nearest 10 percent.

Roster verification was not implemented statewide and was not always fully implemented in the pilot schools, so we supplemented these data with administrative data from the state

[3] Value-added estimates from models that include or exclude attendance are correlated above 0.999.

containing school withdrawal and admission dates for students. For teachers without verified roster data, we used these data to measure the proportion of instructional days each student in the teacher's classroom was officially enrolled in the school (see Walsh et al. 2015 for details on how we used the roster verification and school enrollment data to construct the measure of dosage used in the value-added model).

B. Overview of the value-added model

We estimated value-added results for teachers in math, reading, and science using the model estimated for the state of Oklahoma (described in Walsh et al. 2015), which shares features of value-added models that have been used for teacher evaluation in other states.[4] We estimated regression models separately for each grade and subject combination (we drop a subject subscript for ease of notation). The post-test score depends on pre-test scores, student background characteristics, the student's teacher, and unmeasured factors. For a given teacher *t* and student *i* in grade *g*, the regression equation is as follows:

$$(1)\quad Y_{tig} = \lambda_{Mg} M_{i(g-1)} + \lambda_{Rg} R_{i(g-1)} + \lambda_{Sg} S_{i(g-1)} + \boldsymbol{\beta}'_g \mathbf{X}_i + \boldsymbol{\delta}'_g \mathbf{T}_{tig} + \varepsilon_{tig}\ .$$

In this equation, Y_{tig} is the post-test score in math, reading, or science from grade *g*, $M_{i(g-1)}$ is the math pre-test for student *i* from the previous grade *g-1*, and $R_{i(g-1)}$ is the reading pre-test from the previous grade. For students in the math and reading value-added models for grade 6, we also include $S_{i(g-1)}$, the science pre-test taken in the previous grade.

The pre-test scores in both equations capture prior inputs into student achievement; we estimated the associated coefficients—λ_{Mg}, λ_{Rg}, and λ_{Sg}—using a procedure that corrects for measurement error in these pre-test scores. Specifically, we implemented an errors-in-variables correction (Buonaccorsi 2010) that used information about the reliability of the state tests, available from the test publisher, to net out the known amount of measurement error (CTB/McGraw-Hill 2013a, 2013b). The vector $\mathbf{X}_i$ denotes the control variables for student background characteristics. Equation (1) does not include students' peers because the data linking students to teachers did not identify specific classrooms in the case of teachers with multiple classrooms during a school year.

The vector $\mathbf{T}_{tig}$ consists of binary variables for each teacher. A teacher who taught multiple grades had variables in each grade regression model. For example, a teacher who taught math in grades 4 and 5 had one variable in $\mathbf{T}_{1ti4}$ for the grade 4 regression and one in $\mathbf{T}_{1ti5}$ for the grade 5 regression. Each teacher-student observation corresponding to grade *g* has one nonzero element

[4] For example, value-added models used in Baltimore (American Institutes for Research 2014); Charleston (Resch and Deutsch 2015); District of Columbia Public Schools (Isenberg and Walsh 2014); Florida (American Institutes for Research 2013); Hillsborough County, Florida (Value-Added Research Center 2014); Los Angeles (Value-Added Research Center 2013a); Louisiana (Louisiana's VAA Model 2013); New York City (Value-Added Research Center 2010); Pittsburgh (Rotz et al. 2014); and Tulsa (Value-Added Research Center 2013b) all incorporate empirical Bayes shrinkage. Additionally, all of these models but Louisiana use an errors-in-variables approach to account for measurement error in prior scores, and all but Baltimore, Florida, and Louisiana use fixed effects for teachers.

in $\mathbf{T}_{tig}$. The coefficient vector $\boldsymbol{\delta}$ contains the initial estimates of teacher effectiveness for each teacher.

To account for team teaching, we used the full roster method (Hock and Isenberg 2016, Isenberg and Walsh 2015). In this approach, each student contributed one observation to the model for each teacher to whom he or she was linked. Thus, the unit of observation in the analysis file is a teacher-student combination. We estimated the coefficients by using weighted least squares, weighting each record based on the dosage associated with the teacher-student combination. For example, a student linked to a teacher for 100 percent of instructional time (whether or not the student also received instruction in the subject from another teacher) would contribute twice as much to the estimated coefficient as another student linked to the same teacher for 50 percent of instructional time. We used a cluster-robust sandwich variance estimator (Liang and Zeger 1986; Arellano 1987) to produce consistent standard errors in the presence of heteroskedasticity and correlation in the regression error term.[5]

As is typically done for value-added results used for evaluations, we applied empirical Bayes shrinkage to the estimates (Herrmann et al. 2016) using the empirical Bayes procedure outlined in Morris (1983). Doing so reduces the risk that teachers, particularly those with relatively few students in their grade, will receive a very high or very low effectiveness measure by chance.

As a final step, we rescale the value-added estimates based on their standard deviations within each subject and grade. States and districts frequently scale or standardize value-added estimates using a similar approach to allow them to be combined with other measures in an evaluation system. As a result, differences across grades or subjects in the dispersion of teachers' effects measured in units of student-level achievement often have no consequence for teachers' evaluation scores.

For some analyses described in the next section, we use the unadjusted value-added estimates that have not been subject to empirical Bayes shrinkage or standardized.

C. Empirical approach

We explore our three research questions by examining the results of value-added models that include or exclude certain student test scores from the previous year, and by comparing those results to those from value-added models that include the previously excluded test scores.

[5] Because of computational limitations with the errors-in-variables procedure, we used two regression steps to obtain these standard errors, following Isenberg and Walsh (2014). The first regression was the errors-in-variables regression from Equation 1. We then used the measurement-error corrected values of the pre-test coefficients to calculate adjusted post-test scores by subtracting the predicted effects of the pre-test scores from the post-test scores. In the second step, we obtained the clustered standard errors from a regression of the adjusted post-test scores on the same student background characteristics, grade indicators, and teacher indicators used in the first step.

1. Assessing credibility of value-added models that rely on proxy pre-test scores by examining relationships between post-test and pre-test scores from multiple subjects

The credibility, or face validity, of an evaluation measure may influence how easily it can be adopted into an educator evaluation system. A measure that is not credible may not be embraced by teachers and principals, potentially undermining the goals of the evaluation system.

Value-added models rely on the relationships between test scores and background characteristics, including previous test scores in one or more subjects, to isolate teachers' contributions to achievement. If proxy pre-test scores do a poor job of predicting subsequent student achievement compared to how well same-subject pre-test scores predict subsequent achievement, then value-added estimates that do not account for same-subject pre-test scores may be biased; rather than measuring the contributions of teachers, they may reflect differences in achievement determined by characteristics of students outside of teachers' control. Thus, a value-added model with weaker relationships between pre-test scores and subsequent achievement is less credible.

To assess the credibility of value-added models that rely on proxy pre-tests to account for prior student achievement, we conduct two analyses that examine relationships between different sets of test scores in value-added models. First, we examined how the estimated relationships between post-tests and the included pre-tests change when omitting the same-subject pre-test from math and reading value-added models. Second, we examined how well math and reading pre-test scores predict subsequent science achievement compared to how well they predict subsequent math and reading achievement. For both analyses, we examined (1) the coefficients on the same-subject and proxy pre-tests from the value-added models, and (2) the amount of variation in the outcome explained by the test scores and student characteristics in the value-added model.

To conduct the first analysis, we used value-added models for math and ELA teachers of students in grade 6. Grade 6 math and ELA students in Oklahoma are tested in math, ELA, and science at the end of grade 5, so it is possible to estimate two versions of the grade 6 value-added model in math or reading: (1) a version that accounts for grade 5 pre-test scores in all three subjects, and (2) a version that omits the grade 5 same-subject pre-test score from the model (that is, math pre-test score from the math value-added model and reading pre-test score from the reading value-added model).

For the second analysis that examines the relationships in science value-added models, we examine value-added models for students in grades 5 and 8, when students are tested in all three subjects. In the previous grade, students in grades 5 and 8 are tested in math and reading but are not tested in science. Thus, for these grades, we can examine how relationships with the same pre-tests compare in value-added models for science, math, and reading achievement. As in Oklahoma, few states and districts administer science tests to students in every grade, so this analysis provides insight on whether the feasible science value-added model with no same-subject pre-test can be credible.

As noted above, we assess the relationships in two ways. First, we examine coefficients on the pre-test scores in the value-added models to measure whether the relationships with proxy pre-tests are weaker than those obtained from models that include same-subject pre-tests. If so, value-added estimates from models that rely on proxy pre-tests may be less credible because they may capture factors outside of teachers' control. Across value-added models, we compare the magnitude of the individual coefficients on each pre-test, as well as the sum of coefficients across all of the included pre-tests. Summing the coefficients in each value-added model and comparing these sums provides a measure of how students' predicted scores differ from each other under each model. Specifically, each coefficient sum is how much larger a student's predicted score is compared to another student who performed 1 standard deviation lower on each included pre-test assessment. A smaller sum means that the predicted scores reflect a smaller difference between a student who was high-achieving at baseline compared to another who was low-achieving, which would suggest that the model may be less credible because it may not sufficiently account for baseline achievement.[6]

Our second approach addresses a concern that involves examining only the coefficients on pre-test scores. The coefficients on the pre-tests do not measure the contribution of other background characteristics in the value-added model. For example, even if math and reading scores do a poor job of predicting subsequent science scores, other background characteristics may address the deficiency. To account for these other factors, our second approach examines the combined explanatory power of all covariates in the value-added models. The *R*-squared of a regression from each model is used to describe the amount of variation in the test score that can be collectively explained by the included covariates, with and without the same-subject scores.

2. Assessing accuracy of value-added models that rely on proxy pre-test scores by examining consequences of omitting same-subject pre-test scores

Omitting previous test scores in the same subject as the outcome measure may lead to value-added estimates that provide less accurate measures of teacher effectiveness. We examine this possibility by measuring how much value-added estimates change when excluding these same-subject pre-test scores. We use the same math and ELA value-added models in grade 6 from the previous research question, comparing results from models that account for grade 5 scores in all three subjects and those that omit the grade 5 same-subject pre-test score from the model. If grade 5 science scores provide a reasonable proxy for math or reading skills at the end of grade 6, then the results of these two value-added models (within each subject) will be similar.

We compare the results of the two value-added models in five ways. First, we present the correlation of teachers' value-added estimates based on the two models. An estimated correlation near 1.0 would suggest that excluding students' previous same-subject scores did not lead to differences in teachers' value-added estimates.

Second, because correlations can be difficult to interpret, we also characterize the magnitude of the differences for individual teachers as absolute differences. We report the average absolute difference and the largest absolute difference. We examine the largest difference because value-

[6] Examining the coefficient sums provides a signal of credibility but not a precise measure of how much the predicted scores change. Even if the coefficient sums from two value-added models were identical, predicted scores from the two models could differ because (1) relationships between the other background characteristics in the model could differ, and (2) the differences in previous test scores for actual pairs of students are not identical for each prior test subject.

added estimates are used to evaluate individual teachers; it is important to consider the extent of changes that some individual teachers might experience. Because we scale value-added estimates to have a standard deviation of 1 within each combination of grade and subject, a difference of 1 unit represents a standard deviation of teacher effectiveness, or the difference between a median teacher and one ranked at the 84th percentile. Another way to conceptualize the importance of these differences is by considering the 5-point scale from 1.0 to 5.0 adopted by Oklahoma (Walsh et al. 2015), which is similar to evaluation scales used elsewhere, including in the District of Columbia Public Schools (Isenberg and Walsh 2014). Teachers who receive the average value-added estimate are assigned a 3.0 while teachers who receive an estimate that is 1 standard deviation above or below the average receive a 2.0 or 4.0, and the approximately 5 percent of teachers who receive a value-added estimate that is 2 standard deviations above or below the average receive a 1.0 or 5.0. Under this framework, a difference of 1 standard deviation would move a teacher from one evaluation category to another. In practice, value-added estimates would be combined with other measures to obtain the final evaluation results, so a change of only 1 standard deviation in a teacher's value-added estimate would have a smaller influence on the teacher's combined result.

Third, we also explore how many teachers might be misclassified as among the most or least effective teachers based on the value-added model that excludes the previous math scores. We define the most effective teachers as the top 20 percent of teachers based on value-added estimates and the least effective as the bottom 20 percent. Among teachers classified as being among the least effective in the state based on the value-added model that includes baseline math scores, we present the percentage that remain there after excluding the math scores. Large percentages would suggest that excluding the math scores misclassifies few teachers (relative to the value-added model that includes the math scores). We calculate the analogous percentage for the most effective teachers. Interpreting these percentages as misclassification rates depends on assuming that the value-added estimates from the model that includes all three pre-tests are unbiased, an assumption we discuss below.

Fourth, in addition to the above ways we described the magnitude of the changes in value-added estimates, we also examined how excluding the same-subject pre-test score affects value-added estimates for teachers of low-achieving students. Excluding a prior test score is likely to lead teachers of low-achieving students to be ranked lower by the value-added estimates because the lower test scores of their students is outside of the teachers' control, but not accounted for in the value-added model. To measure how teachers of low-achieving students are affected, we identified the 20 percent of teachers who have the lowest-achieving students based on the same-subject previous assessment, and then measured the proportion of these teachers who were identified as among the least or most effective teachers in the state based on the value-added model. If excluding the same-subject test score from the value-added model leads teachers of low-achieving students to be ranked lower, the percentage of these teachers identified among the least effective will increase.

Finally, we also estimate the amount of bias in the value-added estimates that rely on proxy pre-tests. To compute this bias, we again rely on the assumption that the value-added estimates from the model that includes all three pre-tests is unbiased so that differences can be interpreted as the result of bias. Specifically, we measure bias by decomposing the total variance in the value-added estimates $\hat{\delta}$ into three components: (1) variance due to differences in the true effectiveness of teachers δ; (2) variance due to estimation error μ, assumed to be independent of teachers' true effects; and (3) variance due to bias θ, assumed to be independent of both δ and μ. Under these assumptions, the total variance can be written as follows:

$$(2)\quad V\left[\hat{\delta}\right]=V\left[\delta+\mu+\theta\right]=V\left[\delta\right]+V\left[\mu\right]+V\left[\theta\right].$$

By assuming that estimates from the value-added model with all three pre-tests $\hat{\delta}_U$ are unbiased, the amount of variance due to bias in estimates from the model that relies on proxy pre-tests $\hat{\delta}_P$ is given by the following equation:

$$(3)\quad V\left[\theta_P\right]=V\left[\hat{\delta}_P\right]-V\left[\mu_P\right]-V\left[\delta\right]=\left(V\left[\hat{\delta}_P\right]-V\left[\mu_P\right]\right)-\left(V\left[\hat{\delta}_U\right]-V\left[\mu_U\right]\right).$$

The quantities on the right-hand side of Equation 3 can be estimated. We estimate the variance of the estimation error using the following equation:

$$(4)\quad \hat{V}\left[\mu\right]=\sum\hat{\sigma}_{\mu}^{2}\big/N\,,$$

where $\hat{\sigma}_\mu$ is the estimated standard error of a teacher's value-added estimate and N is the number of teachers with value-added estimates. We estimate $\hat{V}\left[\hat{\delta}\right]$ as the observed variance of the value-added estimates across teachers. We then estimate the fraction of dispersion in the value-added estimates using the model with proxy pre-tests that is due to bias using $\hat{V}\left[\theta_P\right]\big/\hat{V}\left[\hat{\delta}_P\right]$.[7] We measure all variances in student test score units, using the unadjusted value-added estimates obtained prior to applying empirical Bayes shrinkage or standardizing. This approach allows us to directly compare variances across the value-added models for the same subject. If we were to use the standardized estimates instead, then $\hat{V}[\hat{\delta}_P] = \hat{V}[\hat{\delta}_U] = 1$ by construction, preventing us from measuring the full extent to which excluding the pre-test score affects dispersion.

This approach measures the bias in the proxy pre-test model when (1) the model that includes all pre-tests is unbiased, and (2) the bias is not correlated with teachers' true effects. The first of these assumptions is unlikely to hold, but the bias may not be large. For example, using a quasi-experimental approach to measure bias in value-added estimates, Chetty et al. (2014) found that bias accounted for about 5 percent of estimates from a value-added model similar to the one we estimate. If this value-added model is biased, then our estimate of bias in the proxy pre-test model will likely be too small, although the estimated bias could be too large if the

[7] We also report an alternative measure of the amount of bias in the value-added estimates that may be more comparable to results from the quasi-experimental method for measuring bias developed by Chetty et al. (2014): the proportion of variance in the value-added estimates due to bias after removing variance due to estimation error, $\hat{V}\left[\theta_P\right]\big/\hat{V}\left[\delta_P+\theta_P\right]$.

differences offset another source of bias that we do not measure. If the second assumption fails, then whether our estimate of bias is too large or too small will depend on whether the bias tends to help or hurt estimates for relatively more effective teachers. We cannot test these assumptions with our data because we do not observe teachers' true effects. However, rather than providing conclusive evidence, our goal is to produce a measure of bias that serves as a starting point for understanding how the accuracy of value-added estimates is affected by relying on proxy pre-tests.

3. Assessing precision of value-added estimates from models that rely on proxy pre-test scores

Our final research question is focused on the precision of value-added estimates that do not account for same-subject test scores. The precision of an estimate indicates how certain we can be that the estimate reflects the teachers' actual effectiveness (assuming for the moment that the estimates are not systematically in error). Uncertainty can arise from several sources, including having fewer students contribute to the teachers' estimate, good or bad luck affecting the test scores of some individual students (such as illness), events outside of the teachers' control that affected the performance of students in a classroom (such as a barking dog on the day of the test), and the chance that a teacher was assigned to teach a specific set of students who have certain unmeasured characteristics associated with achievement. Excluding previous test scores in a subject increases this last source of uncertainty (and perhaps some of the others as well) because fewer characteristics are accounted for in the value-added model that might affect student achievement.

To assess precision, we compare the reliability and average 95 percent confidence intervals (or margins of error) for value-added estimates that do and do not account for previous same-subject test scores. A smaller reliability or larger confidence interval means that the estimates are less precise. We estimate reliability using $\hat{V}\left[\mu\right]/\hat{V}\left[\hat{\delta}\right]$, where $\hat{V}\left[\mu\right]$ is calculated using Equation 4. As we do when estimating bias, we again use pre-shrinkage value-added estimates to estimate these variances (standardizing the estimates has no effect on the measured reliability). So that they better correspond to the reliability estimates, we measure the margins of error using standardized value-added estimates that were not subject to empirical Bayes shrinkage. Although shrinking the estimates does affect the margins of error, this choice has little consequence for our results because it has little influence on how much the margins of error change when excluding the pre-test.

IV. RESULTS

A. Value-added estimates that do not account for a same-subject pre-test may be less credible

We conducted two analyses to assess the credibility of value-added estimates that rely on proxy pre-tests. From the results of both analyses, we conclude that while that the proxy pre-tests do compensate for the omission of a same-subject pre-test to some extent, the proxy pre-test models appear less credible than those that include the pre-test.

In our first analysis, we estimated value added for grade 6 math and ELA teachers. We then examine how the estimated relationships between grade 6 test scores and grade 5 test scores that are used to generate predicted scores change when omitting the pre-test. Although pre-tests are almost always available and included in value-added models for these subjects, these comparisons provide direct information about the credibility of proxy pre-test models compared to those that include the same-subject pre-test.

In both math and ELA, the estimated relationships between grade 6 test scores and the proxy pre-test scores become larger when omitting the same-subject grade 5 pre-test (Table 2), partially offsetting the lost information about prior achievement. For example, in the math value-added models, the coefficient on grade 5 reading scores increases from 0.107 to 0.274, and the coefficient on the grade 5 science scores increases from 0.149 to 0.593. The test scores used in the value-added models are standardized within grade and subject, so the coefficient of 0.107 means that a standard deviation increase of 1 in a student’s grade 5 ELA score is associated with a 0.107 standard deviation increase in the student’s grade 6 math score. The larger coefficients indicate that proxy pre-tests can provide important information about students’ skills in a particular subject.

Although the larger coefficients on the proxy pre-tests may provide some compensation for the omitted information from the pre-test, the models omitting the pre-test may still be less credible for three reasons. First, when the same-subject pre-test score was included in the model, this score had the strongest relationship with current achievement in both the math and ELA models, suggesting it provides information about a student’s performance that is not fully captured by the proxy pre-tests. Second, differences in predicted current grade scores between students with high and low values of prior-grade scores become slightly smaller when excluding the pre-test for the ELA model, although not for math. To calculate these differences, we sum the coefficients on the prior-grade test scores in each value-added model and compare these sums across models. Each coefficient sum (reported in Table 2) measures how much larger a student’s predicted score is compared to another student who performed 1 standard deviation lower on each of the previous grade assessments included in the regression. In math, the difference between these two students’ predicted scores would be 0.867 standard deviations, whether or not previous-grade math tests were included. In ELA, the students’ predicted scores would differ by 0.912 standard deviations when including all three prior grade test scores, and slightly less (by 0.884) when omitting the prior grade ELA scores. Finally, the amount of variation in the outcome explained by the prior test scores and characteristics in the value-added model—the model *R*-squared reported in the last row of Table 2—falls in both subjects when excluding the pre-test.

Table 2. Relationships between previous test scores and achievement from math and ELA value-added models that include or exclude same-subject pre-test scores

	Coefficients from grade 6 math value-added models		Coefficients from grade 6 ELA value-added models	
Grade 5 subject	Model including grade 5 test scores from all three subjects	Model omitting grade 5 math scores	Model including grade 5 test scores from all three subjects	Model omitting grade 5 ELA scores
Math	0.611 (0.006)	n.a.	0.106 (0.005)	0.217 (0.006)
ELA	0.107 (0.007)	0.274 (0.008)	0.498 (0.006)	n.a.
Science	0.149 (0.011)	0.593 (0.010)	0.308 (0.009)	0.667 (0.008)
Sum of coefficients	0.867	0.867	0.912	0.884
R-squared from value-added model	0.72	0.62	0.70	0.64

Source: Mathematica calculations based on administrative data from OSDE.

Note: The table includes 1,026 math and 1,375 reading teachers in grade 6. Standard errors are in parentheses.

n.a. = not applicable.

In our second analysis, we examine the credibility of a science value-added model that includes math and reading proxy pre-tests, but no same-subject pre-test. In contrast to math and ELA, science is rarely tested in consecutive grades. The consequences of omitting a pre-test for science may be more severe if prior math and reading scores are only weakly related to science performance. Although we cannot directly examine how science value-added results are affected by excluding previous science scores, we can measure how well the available test scores predict subsequent science achievement. Because science scores are available in grades 5 and 8, for this exercise, we examine relationships between test scores in grades 4 and 5 and in grades 7 and 8.

Science value-added models that omit the pre-test may be less credible than either a value-added model for math or ELA that includes the pre-test, or a value-added model for math or ELA that excludes the pre-test. First, the math and reading scores that serve as proxy pre-tests in the science value-added model are stronger predictors of students' subsequent math or ELA scores than they are of subsequent science scores. For the grade 5 value-added models, the sum of the coefficients for the prior test scores in the two subjects is 0.849 in math and 0.847 in ELA, but only 0.554 in science (reported in Table 3). The pattern for the grade 8 value-added models is similar. Additionally, the amount of variation in the outcome explained by the test score and other student characteristics in the value-added model (given by the *R*-squared values in the last row of Table 3) is lowest for the grades 5 and 8 science value-added models.[8] The coefficient

[8] The *R*-squared value for math in grade 8 of 0.63 is more similar to the lower values for the science value-added models than to those for the other math and ELA models reported in Table 3. One possible explanation for this pattern is that the content tested on the grade 7 and grade 8 math tests differs more substantially than the content on the other pre- and post-tests. Another possible explanation is that this difference occurs only by chance. We observed this same pattern using data from the subsequent school year but not in the previous school year, which does not strongly support either of these possible explanations.

sums and variation explained for the grades 5 and 8 science value-added models are also smaller than those reported in Table 2 for the grade 6 math and ELA value-added models that exclude the pre-test. Taken together, these results suggest that a science value-added model that does not include science pre-test scores may not sufficiently account for student background and may be less credible than value-added models that do include pre-test scores.

Table 3. Relationships between previous test scores in math and ELA and achievement from math, ELA, and science value-added models

	Coefficients from grade 5 value-added models			Coefficients from grade 8 value-added models		
Previous grade subject	Grade 5 math scores	Grade 5 ELA scores	Grade 5 science scores	Grade 8 math scores	Grade 8 ELA scores	Grade 8 science scores
Math	0.694 (0.006)	0.158 (0.005)	0.202 (0.004)	0.669 (0.007)	0.197 (0.005)	0.262 (0.004)
ELA	0.155 (0.006)	0.689 (0.005)	0.352 (0.005)	0.160 (0.007)	0.615 (0.005)	0.271 (0.004)
Sum of coefficients	0.849	0.847	0.554	0.829	0.812	0.533
R-squared from value-added model	0.73	0.69	0.61	0.63	0.68	0.60

Source: Mathematica calculations based on administrative data from OSDE.

Note: For grade 5, the table includes data for 1,534 math teachers, 1,822 reading teachers, and 1,445 science teachers. For grade 8, the table includes data for 792 math teachers, 1,229 reading teachers, and 658 science teachers.

It is possible that the relationship between science scores in consecutive grades, which we cannot measure, would be weak. If so, including previous science scores may not greatly improve science value-added estimates compared to only including previous math and reading scores. Even if this possibility is true, the weak relationships suggest that the science value-added estimates may generally be less credible than math or ELA value-added estimates.

B. Omitting previous test scores in the same subject has larger consequences than changing some other features of value-added models and may lead to bias

We examine correlations to verify that the value-added estimates change when omitting same-subject pre-test scores, despite the offsetting larger relationships with the proxy pre-test scores shown in Table 2. Larger changes in value-added estimates suggest more significant concerns for the accuracy of the value-added estimates based on omitting the pre-test scores. As shown in Table 4, grade 6 math value-added estimates that do and do not omit grade 5 math scores have a correlation of 0.88. The correlation between grade 6 ELA value-added estimates that do and do not omit grade 5 reading scores is 0.93.

Compared to correlations obtained from omitting some other value-added model features, those for value-added estimates that do and do not omit pre-test scores are smaller, indicating larger differences. For example, Goldhaber et al. (2014) found that excluding characteristics of students' peers from the same classroom from a value-added model resulted in a correlation of 0.99, and Johnson et al. (2015) found this correlation to be between 0.95 and 0.97, depending on the grade and subject. Our estimated correlations of 0.88 and 0.93 are more similar to those obtained by replacing value-added estimates with growth percentiles from the Colorado Growth Model, which uses a different estimation method and omits background characteristics other than test scores in the same subject. Based on results in Goldhaber et al. (2014) and Walsh and Isenberg (2015), replacing value-added estimates with student growth percentiles results in correlations ranging from 0.83 to 0.93.

Table 4. How value-added estimates for grade 6 teachers change when grade 5 scores are excluded in a teacher's subject

Measure of similarity or difference	Math value added	ELA value added
Correlation	0.88	0.93
Average difference (standard deviations of teacher value added)	0.36	0.28
Largest difference (standard deviations of teacher value added)	1.5	1.8
Percentage of teachers identified among least-effective teachers who remain in that category	71	79
Percentage of teachers identified among most-effective teachers who remain in that category	70	77
Estimated bias in proxy pre-test value-added model, assuming model with all pre-tests is unbiased	11	12

Source: Mathematica calculations based on administrative data from OSDE.

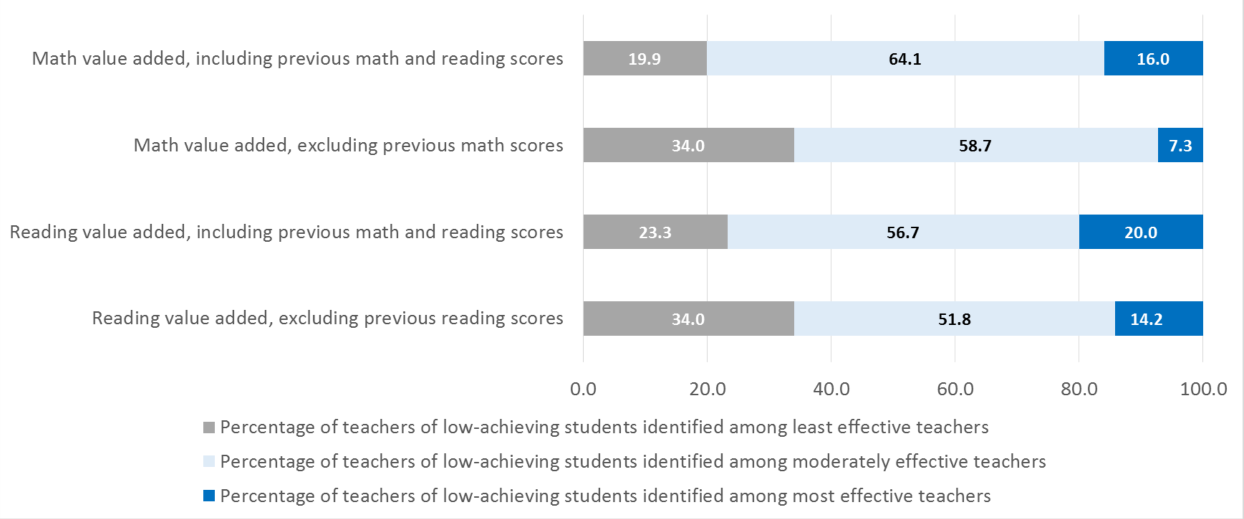
Note: The table includes grade 6 value-added estimates for 1,026 math teachers and 1,375 reading teachers.

Although the correlation is a useful index for comparing the magnitude of differences across contexts, even very high correlations can mask large differences for some individual teachers. Excluding previous same-subject scores from the value-added model led the average teacher's value added-estimate to change by 0.36 standard deviations in absolute value for math, and 0.28 standard deviations for ELA. The largest change for a teacher was 1.5 standard deviations in math and 1.8 standard deviations in ELA.

Next, we examined how excluding previous same-subject scores affected which teachers are identified as the most or least effective. Among the 20 percent of math teachers who are identified as the least effective in the state when including previous math scores, 71 percent would remain in that category. Similarly, 70 percent would remain in the most effective category, defined as the highest-ranked 20 percent of teachers. These percentages were 79 and 77 percent for ELA teachers remaining in the least effective and most effective categories, respectively.

Teachers of low-achieving students are ranked lower when excluding previous same-subject scores. When including previous math scores, 20 percent of teachers of low-achieving students are identified among the least effective teachers and 16 percent are identified among the most effective teachers. When excluding previous math scores, 34 percent are identified among the least effective teachers, and 7 percent are identified among the most effective teachers. The pattern is similar for ELA teachers (Figure 1).[9]

Figure 1. Consequences of excluding prior test scores in the same subject for teachers of low-achieving students in math and reading



Source: Mathematica calculations based on administrative data from OSDE.

Note: Teachers of low-achieving students are defined as the 20 percent of teachers who have the lowest-achieving students based on the same-subject previous assessment. This number includes 205 math teachers and 275 ELA teachers of grade 6.

Another way to measure the consequences of excluding pre-test scores is to directly measure the bias in the value-added estimates. We can measure bias in estimates from the value-added model with proxy pre-tests under the assumptions that the model that includes the full set of pre-tests is unbiased and that the difference in dispersion in estimates between the two models is due to imprecision and bias. Because value-added estimates from any model are likely to include some bias (Rothstein 2010, 2017), the first of these assumptions is unlikely to hold. However, the comparison arguably provides a useful starting point for understanding the magnitude of the bias. Under these assumptions, we find bias that accounts for 11 and 12 percent of the variance in value-added estimates for math and ELA, respectively.[10] In other words, when omitting a

[9] The same pattern does not extend to other student background characteristics in the value-added model, such as free and reduced-price lunch eligibility. For example, when including previous math scores in the math value-added model, 28 percent of teachers of disadvantaged students are identified among the least effective teachers, and 27 percent of these teachers are identified among the least effective teachers when excluding the math scores. These proportions of teachers of disadvantaged students are 24 and 25 percent, respectively, when including or excluding reading scores in the ELA value-added model. We define teachers of disadvantaged students as the 20 percent of teachers in the state who teach the largest proportions of students eligible for free or reduced-price lunch.

[10] Alternatively, this amount of bias reflects 13 and 15 percent of the variance in teacher effects for math and ELA, respectively, *after* accounting for imprecision. This bias is calculated by subtracting the variance due to imprecision from the total variance figure in the denominator. Estimates using this calculation may be more comparable to those from the quasi-experimental approach to estimating bias in value added developed by Chetty et al. (2014).

relevant pre-test score from the value-added model, more variation in achievement is attributed to teachers, rather than to factors outside of teachers' control. We provide additional details on the estimation of the variance components in value-added estimates in Appendix A.

C. Value-added estimates that do not account for previous test scores in the same subject may be slightly less precise

Omitting previous test scores could lead to less precise results because predicted scores are based on less information about student background. However, we find that value-added results become only slightly less precise when excluding previous test scores in the same subject. In both math and reading, the reliability of value-added estimates (measured on a scale of 0 to 1) falls by less than 0.01 when excluding the prior scores.

This small consequence for reliability suggests that the standard errors of the value-added estimates should also remain similar when excluding the prior scores. The average margin of error increases by 2 percent from 0.67 to 0.69 standard deviations of teacher value added when excluding previous math scores (Table 5). In ELA, the average margin of error increases by 1 percent when excluding previous ELA scores.[11] This difference suggests that although science value-added estimates are less precise than either math or ELA value-added estimates, including science test scores from the previous grade may not greatly increase precision.

Table 5. Precision of value-added results by subject and model

Value-added model	Reliability	Average standard error	Average margin of error (95 percent confidence)
Math grade 6	0.86	0.34	+/- 0.67
Math grade 6, no previous math scores	0.85	0.35	+/- 0.69
ELA grade 6	0.76	0.44	+/- 0.87
ELA grade 6, no previous reading scores	0.76	0.45	+/- 0.88
Science grade 5, no previous science scores	0.75	0.48	+/- 0.93
Science grade 8, no previous science scores	0.67	0.50	+/- 0.98

Source: Mathematica calculations based on administrative data from OSDE.

Note: The table includes value-added estimates for 1,026 math teachers and 1,375 ELA teachers of grade 6. The table also includes value-added estimates for 1,439 science teachers of grade 5 and 756 science teachers of grade 8. Values are rounded to the nearest one-hundredth. Standard errors and margins of error are reported in units of teacher-level standard deviations of value-added estimates.

[11] If we had not standardized the value-added estimates within each subject and grade, the consequences for the standard errors would have appeared much larger. The margin of error would have increased by 11 percent in math and 10 percent in ELA when excluding same-subject test scores. However, rather than measuring a real change in precision, these alternative increases largely reflect a larger variance across teachers in their measured contributions to student test scores when excluding the prior test scores. The standard deviation of the value-added estimates increased by 8 percent in math and 9 percent in reading when omitting the pre-test. The reliability estimates in Table 5 suggest that these increases in dispersion, which largely account for the increases in the unstandardized margins of error, are potentially driven by additional bias rather than less precision.

V. DISCUSSION

Although some policy makers might interpret our findings as firm evidence against using value-added estimates that rely on proxy pre-tests, this conclusion is too strong. The choice between different evaluation measures always involves tradeoffs, and alternatives to value-added estimates also have important limitations. For example, student learning objectives can be challenging to implement (Tennessee Department of Education 2015) and may not accurately reflect teachers' contributions (Gill et al. 2013), and classroom observations can also be biased (Steinberg and Garrett 2016; Campbell and Ronfeldt forthcoming). Furthermore, in an evaluation system that incorporates multiple measures of teacher effectiveness, policymakers do not need to choose between using a value-added model and using one of the available alternatives. Instead, policymakers may incorporate all available measures, giving each an appropriate weight. Doing so will improve the validity of the combined measure if each measure's flaws are offsetting rather than compounding.

Policymakers can reduce the weight given to value-added estimates from models that rely on proxy pre-tests relative to the weight given to those of other teachers in subjects with pre-tests. Formally, we recommend an approach that is similar to applying empirical Bayes shrinkage, which is commonly applied to value-added estimates to address error from imprecision. The same approach can be applied to address error from bias to limit the risk that teachers who receive value-added estimates believed to be biased are not incorrectly labeled as a high or low performer.

The additional shrinkage applied to address bias can be calculated by including both the portion of dispersion in the estimates due to bias and the portion due to imprecision as part of the error. In the case of the grade 6 math estimates, including the bias that we measured as 11 percent of the total variance in proxy pre-test value-added estimates (Table 6) as part of the estimation error, results in a reliability of 0.74 instead of 0.86. For example, a teacher with a pre-shrinkage value-added estimate that is 2 standard deviations above average (with average precision) would be shrunk to 1.7 when using a value-added model with pre-tests, or 1.5 when using only proxy pre-tests. While this difference is a substantial increase in the amount of shrinkage applied, a reliability near 0.7 is often tolerated for ELA value-added estimates, which are typically less precise than those for math. For comparison, the analogous ELA estimates using only proxy pre-tests would be shrunk by a factor of 64 percent instead of 76 percent.[12]

In practice, this additional shrinkage for bias must be done in a way that considers the broader evaluation system so that it has the intended effects of properly adjusting for bias and imprecision. The shrinkage adjustment for bias applies the same rate of shrinkage to all teachers

[12] This approach could be adapted to include another adjustment to further reduce systematic bias in the results, but the adjustment may not be practical. We found that value-added estimates with proxy pre-tests led more teachers of low-achieving students to be ranked among the least effective teachers. In theory, a state could remove this systematic bias by adjusting the value-added estimates of these teachers more toward the average teacher compared to the adjustment for teachers of students with pre-test scores that are closer to the average student (by shrinking to a value that depends on the average pre-test score of their students). However, the state is only in the position of needing to make an adjustment because it does not have a same-subject pre-test for these students, so such an adjustment is not feasible.

to account for the absence of a same-subject pre-test. As a consequence, it does not change the ranking of teachers. In contrast, shrinking to adjust for imprecision applies a different rate of shrinkage to each teacher based on the number of students taught and other factors that affect the teacher's standard error estimate, and does change the ranking of teachers. This difference can matter in an evaluation system that scales the adjusted value-added estimates by dividing by the teacher-level standard deviation (as in Oklahoma) or some other constant factor. In this case, shrinkage to adjust for bias is reversed by the scaling and has no effect. Instead, the constant amount of shrinkage to account for bias must be applied so that it affects the relative weight on value-added estimates when they are combined with other measures. Under this approach, proxy pre-test value-added estimates would receive less weight relative to other evaluation measures compared to the weight given to value-added estimates from a model with a same-subject pre-test.

Using this shrinkage-based approach to adjust the weight given to value-added estimates, a state or district can adopt value-added estimates from a grade and subject with only proxy pre-tests, but would continue to place significant weight on the other available evaluation measures.

Table 6. Consequences of omitting same-subject pre-tests for total error in value-added estimates

Value-added model	Standard deviation of teacher effects (standard deviations of student achievement)	Variance due to imprecision (percentage)	Variance due to bias (percentage)	Reliability when considering bias as a source of error
Math grade 6	0.31	14	0	0.86
Math grade 6, no previous math scores	0.33	15	11	0.74
ELA grade 6	0.22	24	0	0.76
ELA grade 6, no previous ELA scores	0.24	24	12	0.64

Source: Mathematica calculations based on administrative data from OSDE.

Note: The table includes value-added estimates for 1,026 math teachers and 1,375 reading teachers of grade 6.

VI. CONCLUSION

States and districts seeking to develop and implement high-quality teacher evaluations have made different decisions about the credibility of teacher evaluation measures in subjects and grades for which no same-subject pre-test is administered. By using these value-added models, states and districts can cover more teachers. Recent work has produced evidence that value-added models that account for same-subject pre-tests scores and other student background characteristics can provide accurate measures of teachers' contributions to student achievement. However, this work has not assessed the accuracy of value-added models without a same-subject pre-test.

Our results suggest that value-added estimates that must rely on previous scores in other subjects only may be less accurate, less credible, and less precise compared to those that include previous scores in the same subject. However, these important limitations do not necessarily mean that these results should not be used. Each measure, whether classroom observations, student learning objectives, or value-added estimates, has important limitations that affect its accuracy and precision. A state or district should consider whether the limitations of value-added estimates that do not account for previous same-subject test scores are larger or smaller than the limitations of the measures that would otherwise be used in their stead to evaluate teachers. Although flawed, it may be appropriate to assign some positive weight to these value-added measures, even if the weight is lower than weights given to other measures.

We identify two primary limitations with this study. First, in comparing results from value-added models that include different features, we can describe the extent of differences, but cannot conclusively measure bias. While it may be reasonable to speculate that any differences in value-added estimates that arise from excluding prior test scores represent additional bias, we cannot rule out the possibility that the changes offset another source of bias that we are unable to measure in this context. Second, we are only able to directly measure the consequences of excluding prior same-subject scores in value-added models for teachers of math and ELA, whereas states and districts will be most interested in the consequences for other subjects such as science that are not typically tested in every grade. Unfortunately, our data lack test scores from consecutive grades in other subjects. While we leave open the possibility that some consequences are larger or smaller in other subjects, we draw similar conclusions in both math and reading, suggesting our results may be broadly applicable to other subjects.

REFERENCES

American Institutes for Research. “Florida Comprehensive Assessment Test (FCAT) 2.0 Value-Added Model: Technical Report 2012–13.” Washington, DC: American Institutes for Research, November 2013.

American Institutes for Research. “Baltimore City Public Schools Value-Added Model Technical Report.” Washington, DC: American Institutes for Research, 2014.

Arellano, M. “Computing Robust Standard Errors for Within-Groups Estimators.” *Oxford Bulletin of Economics and Statistics,* vol. 49, no. 4, November 1987, pp. 431–434. doi:10.1111/j.1468-0084.1987.mp49004006.x.

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

Campbell, S. L., and M. Ronfeldt. “Observational Evaluation of Teachers: Measuring More Than We Bargained For?” *American Educational Research Journal,* forthcoming. doi:10.3102/0002831218776216.

Chetty, R., J. N. Friedman, and J. E. Rockoff. “Measuring the Impacts of Teachers I: Evaluating Bias in Teacher Value-Added Estimates.” *American Economic Review,* vol. 104, no. 9, September 2014, pp. 2593–2632. doi:10.1257/aer.104.9.2593.

CTB/McGraw Hill. *Oklahoma School Testing Program, Oklahoma Core Curriculum Tests, Grades 3 to 8 Assessments, 2012–2013 Technical Report.* Monterey, CA: CTB/McGraw Hill, 2013a.

CTB/McGraw Hill. *Oklahoma School Testing Program, Oklahoma Core Curriculum Tests, End-of-Instruction Assessments, 2012–2013 Technical Report.* Monterey, CA: CTB/McGraw Hill, 2013b.

Gill, B., J. Bruch, and K. Booker. “Using Alternative Student Growth Measures for Evaluating Teacher Performance: What the Literature Says” (REL 2013–002). Washington, DC: U.S. Department of Education, Institute of Education Sciences, National Center for Education Evaluation and Regional Assistance, Regional Educational Laboratory Mid-Atlantic, September 2013.

Goldhaber, D., J. Walch, and B. Gabele. “Does the Model Matter? Exploring the Relationship Between Different Student Achievement-Based Teacher Assessments.” *Statistics and Public Policy,* vol. 1, no. 1, November 2014, pp. 28–39. doi:10.1080/2330443X.2013.856169.

Herrmann, M., E. Walsh, and E. Isenberg. “Shrinkage of Value-Added Estimates and Characteristics of Students with Hard-to-Predict Achievement Levels.” *Statistics and Public Policy,* vol. 3, no. 1, May 2016, pp. 1–10. doi:10.1080/2330443X.2016.1182878.

Hock, H., and E. Isenberg. "Methods for Accounting for Co-Teaching in Value-Added Models." *Statistics and Public Policy*, vol. 4, no. 1, November 2016, pp. 1–11. doi:10.1080/2330443X.2016.1265473.

Isenberg, E., and E. Walsh. "Measuring Teacher Value Added in DC, 2013-2014 School Year." Washington, DC: Mathematica Policy Research, August 2014.

Isenberg, E., and E. Walsh. "Accounting for Co-Teaching: A Guide for Policymakers and Developers of Value-Added Models." *Journal of Research on Educational Effectiveness*, vol. 8, no. 1, January 2015, pp. 112–119. doi:10.1080/19345747.2014.974232.

Johnson, M. T., S. Lipscomb, and B. Gill. "Sensitivity of Teacher Value-Added Estimates to Student and Peer Control Variables." *Journal of Research on Educational Effectiveness*, vol. 8, no. 1, October 2015, pp. 60–83. doi:10.1080/19345747.2014.967898.

Liang, K.-Y., and S. L. Zeger. "Longitudinal Data Analysis Using Generalized Linear Models." *Biometrika*, vol. 73, no. 1, April 1986, pp. 13–22. doi:10.1093/biomet/73.1.13.

Louisiana's Value-Added Assessment Model as Specified in Act 54. "A Report to the Board of Elementary and Secondary Education, September 2013." Available at https://www.louisianabelieves.com/docs/teaching/2012-2013-value-added-report.pdf?sfvrsn=8. Accessed March 27, 2018.

Morris, C. N. "Parametric Empirical Bayes Inference: Theory and Applications." *Journal of the American Statistical Association*, vol. 78, no. 1, 1983, pp. 47–55. doi:10.1080/01621459.1983.10477920.

Resch, A., and J. Deutsch. "Measuring School and Teacher Value Added in Charleston County School District, 2014–2015 School Year." Washington, DC: Mathematica Policy Research, November 2015.

Rothstein, J. "Teacher Quality in Educational Production: Tracking, Decay, and Student Achievement." *Quarterly Journal of Economics*, vol. 125, no. 1, February 2010, pp. 175–214. doi:10.1162/qjec.2010.125.1.175.

Rothstein, J. "Measuring the Impact of Teachers: Comment." *American Economic Review*, vol. 107, no. 6, June 2017, pp. 1656–1684. doi:10.1257/aer.20141440.

Rotz, D., M. Johnson, and B. Gill. "Value-Added Models for the Pittsburgh Public Schools, 2012-13 School Year." Cambridge, MA: Mathematica Policy Research, May 2014.

Steinberg, M. P., and R. Garrett. "Classroom Composition and Measured Teacher Performance: What Do Teacher Observation Scores Really Measure?" *Educational Evaluation and Policy Analysis*, vol. 38, no. 2, June 2016, pp. 293–317. doi:10.3102/0162373715616249.

Tennessee Department of Education. "Teacher and Administrator Evaluation in Tennessee: A Report on Year 3 Implementation." Nashville, TN: Tennessee Department of Education, April 2015.

Value-Added Research Center. “NYC Teacher Data Initiative: Technical Report on the NYC Value-Added Model 2010.” Madison, WI: Value-Added Research Center at Wisconsin Center for Education Research, University of Wisconsin-Madison, 2010.

Value-Added Research Center. “Academic Growth over Time: Technical Report on the LAUSD Teacher-Level Model Academic Year 2012–2013.” Madison, WI: Value-Added Research Center at Wisconsin Center for Education Research, University of Wisconsin-Madison, 2013a.

Value-Added Research Center. “Value-Added Analysis Technical Report for Oklahoma Gear Up, 2012–2013.” Madison, WI: Value-Added Research Center at Wisconsin Center for Education Research, University of Wisconsin-Madison, 2013b.

Value-Added Research Center. “Hillsborough County Public Schools Value-Added Project: Year 3 Technical Report HCPS Value-Added Models 2013.” Madison, WI: Value-Added Research Center at Wisconsin Center for Education Research, University of Wisconsin-Madison, 2014.

Walsh, E., and D. Dotter. “Longitudinal Analysis of the Effectiveness of DCPS Teachers.” Washington, DC: Mathematica Policy Research, July 2014.

Walsh E., and E. Isenberg. “How Does Value Added Compare to Student Growth Percentiles?” *Statistics and Public Policy,* vol. 2, no. 1, April 2015, pp. 1–13. doi:10.1080/2330443X.2015.1034390.

Walsh, E., A. Y. Liu, and D. Dotter. “Measuring Teacher and School Value Added in Oklahoma, 2013–2014 School Year.” Washington, DC: Mathematica Policy Research, February 2015.

APPENDIX A. ADDITIONAL DETAILS OF THE DECOMPOSITION OF VARIANCE IN VALUE-ADDED ESTIMATES

To measure the reliability and accuracy of value-added estimates, we decomposed the variance in each set of value-added estimates we produced. We estimated components for the true dispersion in value added, bias, and estimation error, as shown in Equations (3) and (4). To prepare this estimate, we used value-added estimates that were not subject to empirical Bayes shrinkage or standardization so that the variances could be compared across value-added models in units of student achievement. These value-added estimates are measured in units of student-level standard deviations. In Table A.1, we report the estimated variance components for these value-added estimates. Omitting the pre-tests results in larger total and error (imprecision) variances in each subject. We attribute the remaining variance in the value-added models that include the pre-test to true differences in effectiveness by assuming these estimates contain no bias. By applying these same estimates of the variance in true effectiveness to the models that omit the pre-tests, we obtain estimates of the variance due to bias.

Table. A.1.Estimated components of variance in each value-added model

Value-added model	Total variance	Variance due to true differences in effectiveness	Variance due to imprecision	Variance due to bias
Math grade 6	0.094	0.081	0.014	0.000
Math grade 6, no previous math scores	0.109	0.081	0.016	0.012
ELA grade 6	0.048	0.036	0.011	0.000
ELA grade 6, no previous ELA scores	0.057	0.036	0.014	0.007

Source: Mathematica calculations based on administrative data from OSDE.

Note: The table includes value-added estimates for 1,026 math teachers and 1,375 reading teachers of grade 6.

.

ABOUT THE SERIES

Policymakers and researchers require timely, accurate, evidence-based research as soon as it's available. Further, statistical agencies need information about statistical techniques and survey practices that yield valid and reliable data. To meet these needs, Mathematica's working paper series offers access to our most current work.

For more information about this paper, contact Elias Walsh, Researcher, at EWalsh@mathematica-mpr.com.

Suggested citation: Walsh, Elias, Dallas Dotter, and Albert Y. Liu. "Can More Teachers Be Covered? The Accuracy, Credibility, and Precision of Value-Added Estimates with Proxy Pre-Tests." Working Paper 64. Princeton, NJ: Mathematica Policy Research, August 2018.

www.mathematica-mpr.com

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

PRINCETON, NJ ■ ANN ARBOR, MI ■ CAMBRIDGE, MA ■ CHICAGO, IL ■ OAKLAND, CA ■ SEATTLE, WA ■ TUCSON, AZ ■ WASHINGTON, DC ■ WOODLAWN, MD

M50

MATHEMATICA
Policy Research

Mathematica® is a registered trademark
of Mathematica Policy Research, Inc.