Different Tests, Different Answers: The Stability of Teacher Value-Added Estimates Across Outcome Measures

John P. Papay

Am Educ Res J 2011 48: 163 originally published online 19 April 2010
DOI: 10.3102/0002831210362589

American Educational Research Journal
February 2011, Vol. 48, No. 1, pp. 163–193
DOI: 10.3102/0002831210362589

Different Tests, Different Answers: The Stability of Teacher Value-Added Estimates Across Outcome Measures

John P. Papay
Harvard University Graduate School of Education

Recently, educational researchers and practitioners have turned to value-added models to evaluate teacher performance. Although value-added estimates depend on the assessment used to measure student achievement, the importance of outcome selection has received scant attention in the literature. Using data from a large, urban school district, I examine whether value-added estimates from three separate reading achievement tests provide similar answers about teacher performance. I find moderate-sized rank correlations, ranging from 0.15 to 0.58, between the estimates derived from different tests. Although the tests vary to some degree in content, scaling, and sample of students, these factors do not explain the differences in teacher effects. Instead, test timing and measurement error contribute substantially to the instability of value-added estimates across tests.

KEYWORDS: teacher research, school/teacher effectiveness, teacher assessment, educational policy

In the last two decades, educational researchers and practitioners have increasingly adopted value-added methods to evaluate student, teacher, and school performance. By examining test scores over time, value-added models purport to isolate the contributions of individual teachers or schools to student achievement. Researchers have implemented these models to explore many important topics in the economics of education. At the same time, states and districts have begun using them to identify and reward high-performing teachers. As value-added models have become increasingly widespread and carry higher stakes, questions concerning the validity and reliability of their results have grown more important.

JOHN P. PAPAY is an advanced doctoral student at Harvard Graduate School of Education, Appian Way, Cambridge, MA 02138; e-mail: *john_papay@mail.harvard.edu*. His research interests include teacher policy, the economics of education, teacher labor markets, and teachers unions.

Researchers almost unanimously acknowledge that value-added methods represent a substantial improvement over traditional analyses based on test score levels. Because students selectively sort into schools and classes, simply comparing average test scores provides only limited information about teacher (or school) performance. Instead, isolating student progress over the course of a year can offer a more realistic assessment. The research community, however, remains divided over the ultimate usefulness of value-added estimates for policy. Some argue that the methodologies support causal claims that specific teachers[1] increase student achievement; as such, compensation and accountability policies based on these estimates are justified (Sanders, 2000; Sanders & Horn, 1994). Others, however, assert that the many assumptions underlying these models make such claims tenuous at best (Koretz, 2002; Raudenbush, 2004; Rothstein, 2007; Rubin, Stuart, & Zanutto, 2004) and that low-stakes uses may be most appropriate.

Given their growing prominence, value-added models have recently come under increased scrutiny from the research community, which has explored the sensitivity of estimates to a variety of modeling choices. Several different types of value-added models exist, and analysts must choose a model and decide how to specify it. These decisions can affect estimates of teacher performance. For example, the decision whether or not to include school fixed effects is important because teachers may look very effective when compared with other teachers in their school, but not compared to all teachers in the district. Some of these modeling choices, such as whether to make within-school or across-school comparisons, should be driven by the specific inference of interest. However, other decisions, including how to model cross-grade correlations or persistence of teacher effects over time, depend on larger—and less transparent—assumptions.

Similarly, value-added models all depend on the specific assessment used to measure student achievement. The importance of choices about which outcome measure to use, though, has received much less attention in the research literature. In theory, policymakers design state and local tests to assess the specific content standards deemed important. However, most observers would hope that student progress on one achievement test in a certain domain would translate to progress on other, similar tests. In other words, students deemed excellent readers on one test should also do well on another test of reading. Similarly, if a district wanted to reward its teachers based on their ability to improve students' reading performance, officials would hope that the best teachers would raise student scores not only on the state test but also on other reading assessments. Conflicting evidence from two similar tests would raise important questions about the validity of any teacher effectiveness estimates.

Scholars have compared results obtained from different value-added models and specifications, often finding important differences (Ballou, Sanders, & Wright, 2004; Harris & Sass, 2006; Lockwood et al., 2007;

Tekwe et al., 2004). Few studies to date, however, have explored the issue of outcome choice on teacher value-added estimates. Sass (2008) reported correlations of 0.48 between value-added measures derived from the high-stakes and low-stakes tests in Florida. In a more comprehensive study, J. R. Lockwood and colleagues (2007) examined teacher effectiveness estimates derived from two different subscales of the mathematics Stanford Achievement Test (SAT). They found that the choice of outcome measure contributes more substantially to variation in these estimates than decisions about model specification.

Using a longitudinal data set with 6 years of linked teacher-student records from a large, urban school district in the Northeast, I first replicate Lockwood et al.'s primary analyses using two mathematics subscales of the SAT. My results generally support their earlier conclusions: Much more variation in teacher value-added estimates arises from the choice of outcome than the model specification. I then extend Lockwood et al.'s analysis in several ways. Most importantly, while Lockwood et al. examine different subtests of the same assessment, I use three complete reading achievement tests—the state test, the SAT, and the Scholastic Reading Inventory (SRI)—to explore how estimates of teacher performance vary across outcome measures. The district currently uses all three of these tests, formally and informally, to measure student progress towards district and state goals.

My primary results compare the relative rankings of teacher effectiveness across these three reading outcome measures. I find moderate-sized rank correlations, ranging from 0.15 to 0.58, between the estimates derived from these different tests. In all cases, these correlations are statistically significant and indicate that teachers who generate substantial student achievement growth on one measure also tend to perform well on others. However, these estimates are not sufficiently comparable to rank consistently the same individual teachers as high- or low-performing. In fact, if this district implemented a high-stakes pay-for-performance program similar to the one currently operating in Houston, Texas, simply switching the outcome measure would affect the performance bonuses for nearly half of all teachers and the average teacher's salary would change by more than $2,000.

I extend this analysis by exploring several possible explanations for the differences in teacher effects across outcomes. First, while all three tests purport to measure English and reading achievement, they may in fact assess somewhat different content and skills, and any differences in teacher value-added estimates may merely reflect this variation in content coverage or test coaching. Second, all three tests have different scales. Third, even within the same classroom, a different sample of students takes each test because of absenteeism or mobility throughout the school year. Fourth, test timing itself may affect value-added estimates. The district gives students the SRI in both September and May, while students take the SAT in October and the state test in March. Differences in the specific test date (e.g., March

vs. May) and the baseline-outcome test combination (i.e., Fall-Fall, Fall-Spring, or Spring-Spring) may affect value-added estimates. Finally, all tests are noisy measures of latent reading achievement, and this error may produce instability in teacher effects.

In short, using different achievement tests produces substantially different estimates of individual teacher effectiveness. The variation in teacher value-added estimates that arises from using different outcomes far exceeds the variation introduced by implementing different model specifications. Although the three tests do vary to some degree in their content, item format, scaling, and sample of students, these factors do not appear to explain the differences in teacher effects. Instead, my results suggest that test timing and inconsistency, such as measurement error, play a much greater role. In particular, the finding that the timing of the test alone may produce substantial variation in teacher productivity estimates across outcome measures raises important questions for teacher accountability policies. Importantly, though, the study design does not support definitive conclusions about the relative contributions of these different factors, and these results must be seen as tentative.

In the next section, I examine the importance, to both policy and research, of teacher value-added models and present a theoretical model for understanding the possible sources of inconsistency in teacher estimates arising from outcome selection. I then describe my research design and the data used. I revisit Lockwood et al.'s (2007) work on mathematics achievement subtests and compare results from three complete reading assessments. I then present findings concerning possible reasons for these differences in teacher value-added estimates. Finally, I conclude and offer some implications for further research and practice.

Value-Added Models in Practice

Value-added methodologies for evaluating teacher performance first appeared in the education research literature in the 1970s (Boardman & Murnane, 1979; Hanushek, 1971) and grew in prominence in the mid-1990s, as William Sanders's evaluations of Tennessee test scores drew the attention of policymakers and practitioners (Sanders, 2000). Educational researchers have used these models to examine the importance of teachers in raising student achievement (Rivkin, Hanushek, & Kain, 2005), teacher attrition patterns (Boyd, Grossman, Lankford, Loeb, & Wycoff, 2007), differences in performance by certification pathway (Boyd, Grossman, Lankford, Loeb, & Wycoff, 2006; Clotfelter, Ladd, & Vigdor, 2007; Gordon, Kane, & Staiger, 2006; Kane, Rockoff, & Staiger, 2008; Murnane, 1984), teacher evaluation (Goldhaber & Anthony, 2007; Jacob & Lefgren, 2005), and the returns to teacher experience (Rockoff, 2004).

Beyond the research community, many educational policymakers have also come to see value-added models as a promising method to reform

teacher evaluation and offer pay-for-performance. In a widely publicized announcement, New York City recently unveiled publicly reported grades for all schools based largely on value-added test scores (Gewertz, 2007). Several states, including Florida, Texas, and North Carolina, and large school districts have created performance bonuses based on some form of teacher value-added scores (Olson, 2007). Recently, the U.S. Department of Education called for states to implement measures of teacher performance as part of the "Race to the Top" guidelines (McNeil, 2009).

Estimating Value-Added Models

Value-added models attempt to estimate the causal contribution of a teacher to his or her students. Analysts attribute to the teacher any persistent differences between each student's actual performance and the student's hypothetical (or counterfactual) performance if he or she had had an "average" teacher, as follows:

$$Y_{it} - \tilde{Y}_{i\bar{t}} = T_{it}' \, \delta + \varepsilon_{it}, \;\; \varepsilon_{it} \perp T_{it} \qquad (1)$$

where Y_{it} represents a student's standardized test score, $\tilde{Y}_{i\bar{t}}$ represents the counterfactual score if the student had had an average teacher, and T_{it} is a full set of teacher indicator variables for student i in year t. Here, δ represents the estimates of teacher effectiveness. To predict this counterfactual performance, analysts attempt to estimate educational production functions in a tractable manner by making assumptions about how past educational and family inputs affect a student's future academic performance. In general, these models rely on some form of a student's past test performance as a sufficient statistic for the full range of these previous inputs (McCaffrey, Lockwood, Koretz, & Hamilton, 2003; McCaffrey, Lockwood, Koretz, Louis, & Hamilton, 2004; Todd & Wolpin, 2003). Several scholars have recently questioned whether these approaches accurately identify causal teacher effects (Rothstein, 2007; Rubin et al., 2004); this article leaves aside that important question and focuses instead on value-added models as they have been implemented.

Analysts have developed a variety of value-added approaches (see McCaffrey et al., 2004, for a more complete discussion). These models differ both in their level of complexity and in the assumptions they make about the persistence of teacher effects and the correlations in student test scores over time and within classes. Despite these differences, McCaffrey et al. (2004) argue that the different models "provide reasonably similar estimates of individual teacher's effects" (p. 91). Other comparisons of value-added models have reached similar conclusions (Harris & Sass, 2006; Lockwood et al., 2007; Tekwe et al., 2004).

Other choices about model specification depend on how districts or researchers plan to use these estimates. For example, some models take a teacher's estimated effect as given (essentially using the teacher's estimated

"fixed effect"), while others "shrink" teacher estimates back toward the mean, with the amount of attenuation depending on the available information concerning teacher performance (e.g., the sample size and variability of student test scores). With unshrunken estimates, extreme results can occur both from true differences in teacher performance and from sampling variation. Some evidence suggests that the decision may not make a substantive difference in relative teacher rankings (Tekwe et al., 2004), although the question remains open (see Harris & Sass, 2006; McCaffrey et al., 2003).

Analysts must also choose how to estimate a student's counterfactual performance, deciding whether to control for school fixed effects as well as individual or classroom-level covariates to account for differences in family background or classroom peer groups (see McCaffrey et al., 2004, and McCaffrey et al., 2003, for a more complete discussion). Including covariates implicitly compares teachers in the same schools or with the same mix of students, requiring analysts to make substantially different inferences. Several researchers have found that including student-level demographic covariates produces only small changes in teacher effects (Ballou et al., 2004; Harris & Sass, 2006; Lockwood et al., 2007), but Harris and Sass (2006) found that teacher estimates are much more sensitive to the inclusion of school effects. Thus, the decision to compare teachers within or across schools can have a substantial impact on teacher value-added estimates.

Research Questions

Although a growing literature surrounds the effect of model choice and specification on teacher value-added estimates, the impact of outcome selection has been largely ignored. Lockwood and colleagues (2007) assert that researchers "have not directly compared VAM [value-added modeling] teacher effects obtained with different measures of the same broad content area" (p. 48). Their analysis represents the first attempt to quantify these differences. My study extends their important work to examine the sensitivity of teacher value-added models to outcome selection along several dimensions.

All value-added models rely on the assumption that teacher effectiveness can be estimated reliably and validly through student achievement tests. Here, both the choice of achievement measure and the properties of that assessment play a substantial role in the accurate estimation of teacher effects. I structure my analysis around three primary research questions:

Research Question 1: Do teacher value-added effects constructed from different assessments produce different estimates of teacher effectiveness?

Research Question 2: Do these differences, if any, have practical significance for teacher accountability?

Research Question 3: What accounts for any differences in teacher estimates across outcomes?

A Model of the Importance of Outcome Selection

A simple theoretical model, building on equation (1), helps clarify this discussion. Here, instead of examining test scores, I model the difference in a student's actual and counterfactual latent achievement, A^*, as a function of a set of indicators, T^*, that denote each student's teacher:

$$A^*_{it} - \tilde{A}^*_{i\bar{t}} = T^*_{it}{}'\delta + \varepsilon_{it},\ \varepsilon_{it} \perp T^*_{it} \tag{2}$$

This relationship provides the logic for a model of "true" teacher effects. Unfortunately, a student's latent achievement is not observable. Instead, a test represents a noisy measure of some aspects of that achievement. Variation in student achievement across tests can come from a variety of sources: test-level sampling of domains assessed (η_t), school-level or classroom-level shocks that affect all students taking the same test at the same time (υ_t), and individual-level variation from idiosyncratic shocks or sampling of test items (μ_{it}).

We can explore these sources of variation beginning with a Classical Test Theory framework. Here, a student's observed test score represents his or her true score plus individual-level mean zero error:

$$Y_{it} = Y^*_{it} + \mu_{it},\ \mu_{it} \perp Y^*_{it} \tag{3}$$

Assessments sample items (both content and format) that attempt to represent an entire domain of knowledge. Because some students may perform better on any individual item than on others, this sampling of items leads to individual variation in test scores. Furthermore, individual-level error arises from idiosyncratic shocks, such as a good night's sleep the day before the test or skipping breakfast that morning. Over repeated instances of measurement on an unbiased test, these sources of error wash out. From the perspective of estimating teacher effectiveness, increasing the sample of students reduces the inconsistency produced by these errors.

Beyond individual-level error, though, school-level or classroom-level shocks may also cause a student's actual score to differ from his or her true score on the specific test:

$$Y_{it} = Y^*_{it} + \upsilon_t + \mu_{it},\ (\upsilon_t + \mu_{it}) \perp Y^*_{it} \tag{4}$$

This classroom-level error, υ_t, can arise for a variety of reasons, such as a disruptive student in the class on test day, a teacher who inadvertently gives the students an extra 5 minutes on a section, or, as suggested by Kane and Staiger (2002), a dog barking in the parking lot. These events may affect scores for many students in the classroom. These random aggregate shocks also wash out over repeated instances of measurement. However, as many elementary school teachers teach one class per year, these classroom-level shocks can produce serious challenges for estimating teacher effects. For

example, in accountability systems that attempt to estimate a teacher's effectiveness using only 1 year of data, the value-added estimates for these teachers would be conflated with the classroom-level error in that year.

Individual-level and aggregate-level errors make a student's actual score on a specific test (Y_{it}^{1}) a noisy measure of his or her true score on that test. Furthermore, teacher value-added models aim to estimate teacher contributions to latent student achievement ($A^{*}{}_{it}$), not just to performance on a single test. A student's true score on one test (Y_{it}^{1*}) is a noisy measure of latent achievement—the test's ability to represent latent student achievement depends on selection of the content domains tested. These choices will lead to different inferences about a student's achievement, producing different true test scores on two assessments. Here, η_t captures this variation[2]:

$$Y_{it}^{1*} = A^{*}{}_{it} + \eta_t^1,\ \eta_t^1 \perp A^{*}{}_{it} \tag{5}$$

For example, a student's true score, purged of measurement error, on two tests may be different simply because they measure different latent constructs or because one focuses more on reading comprehension while the other includes more vocabulary items. These differences in student scores can obviously produce variation in teacher effectiveness estimates: A teacher who focuses instruction on vocabulary knowledge will look more effective using the latter test. Thus, two teachers may be of equal quality, but their measured performance will differ depending on which test is used. This concern may be particularly important when one of the tests carries high stakes for schools or teachers, as with a state test for accountability. Here, teachers face incentives to tailor their instruction to the types of content and item formats found most frequently on the state test. Coaching for a specific test would improve student performance on the domains that make up that test, so student scores would increase more on that test than on other tests that exclude or weight differently these domains. As a result, inferences about student performance may contain not only measurement error but also bias that would not wash out over repeated instances of measurement.

Student scores on an individual assessment are thus noisy measures of their latent achievement because of inconsistencies—error and bias—at the individual, aggregate, and test levels, as follows:

$$Y_{it}^{1} = A^{*}{}_{it} + \eta_t^1 + \upsilon_t^1 + \mu_{it}^1 \tag{6}$$

As value-added models essentially involve estimating gains in student-level achievement, these sources of error become magnified. In a pure gain score model, the object of interest is the difference between two noisy measures of student performance. Even if the two measures themselves are highly reliable, the reliability of the difference will, in practice, be lower because any persistent elements of student performance are differenced out, greatly reducing true score variance.[3] While covariate-adjustment and other more

complicated value-added models do not preserve this precise relationship found in gain score models, the logic follows; in practice, the reliability of student achievement growth is lower than that of the individual tests themselves.

These sources of variation all lead to different estimates of student achievement growth using different outcome measures. Aggregating student-level test scores to estimate teacher effects attenuates individual-level error but compounds other inconsistencies, such as classroom-level error, that do not disappear. Similarly, school-level value-added estimates would substantially limit issues arising from individual-level errors, but if school-level errors are large—particularly if the true variation among schools is relatively small—the reliability of school-level value-added estimates could be similarly low even at a higher level of aggregation.

Additional variation in teacher estimates arises from the nature of testing. Students take tests on different days and at different times of the year. Because students, particularly those in urban schools, have relatively high absenteeism and mobility, the students present to take each test may vary substantially. Thus, teacher value-added estimates may vary across outcomes in part because different samples of students take each test.

Furthermore, measuring teacher effects introduces other considerations surrounding the identification of a student's actual teacher (T^*) and the timing of outcome and baseline tests. In theory, a teacher should only be assessed for student learning that occurs when the student is in that teacher's class. Thus, a "pure" value-added measure would compare a student's performance on the first day of the school year to his or her performance on the very last day, although such testing patterns are not practically tenable. However, tests rarely occur at the very beginning or end of the school year, leading a portion of one teacher's instruction to be attributed to another teacher.[4] For example, using a March test, like the state test, raises two issues about estimating teacher effects. First, 30% of the progress attributable to an individual teacher may actually have come from the previous year's teacher. Second, the estimates do not account for the teacher's contribution to student learning during the last 3 months of the school year. Thus, a March test will likely provide different estimates of teacher productivity than a May test simply because of test timing.

A further complication is that even within a school year it is often difficult to attribute students' learning only to one teacher. This issue arises both because of mobility—within and across schools—and because students often learn skills from several teachers at the same time (Croninger & Valli, 2009). For example, students may learn reading skills not only from their English teacher but also in science or social studies classes. This challenge is particularly problematic in middle and high schools where students often have different teachers for many of their subjects.

Finally, most value-added models use a previous year's test score as a measure of past performance.[5] Thus, any learning loss that occurs over the summer—or any gain from summer enrichment programs—becomes conflated with value-added estimates. If some teachers have classes that are disproportionately filled with students whose skills eroded over the summer, those teachers may actually have greater impacts on student learning than teachers with similar estimated effects whose students' skills did not diminish during vacation.[6] A fall-to-fall analysis could produce the same difficulty, only with a different summer's loss included. Thus, even across the same test, we would expect spring-to-spring, fall-to-spring, and fall-to-fall estimates to differ because summer learning loss (or gain) may not be distributed randomly across teachers.

Research Design

Data

This article uses a comprehensive administrative dataset from a large, urban school district in the Northeast U.S. that includes student, test, and teacher records from fall 2001–2002 to fall 2007–2008. This district has approximately 55,000 students and 5,000 teachers. Student data include demographic information and teacher identifiers for each subject. The district tests its students using several different outcomes, including the state test—which is used for school-level accountability—and the low-stakes Stanford Achievement Test (SAT) and Scholastic Reading Inventory (SRI). The dataset includes separate files with testing data for each of these assessments. I merged testing data to student records using school year and unique student identifiers.

Appropriate identification of value-added estimates requires both baseline and outcome test data. Because most of the possible value-added estimates come from late elementary school grades, I focus my analysis on students in Grades 3 through 5. Depending on the outcome, the data available for value-added analysis includes 20,000 to 32,000 student-year records and produces estimates for 526 to 762 unique teachers.

Measures

Baseline and outcome test measures come directly from the district's administrative data. These tests vary in the grades and timing of their administration: The state test is given in the spring of each academic year, the SAT in the fall, and the SRI in both the fall and spring. They all measure reading proficiency, although the SRI is used as a formative assessment designed to evaluate student progress through the year while the state test is a summative assessment mapped directly to the state's content standards. Both the SAT and SRI are vertically equated while the state test is scaled separately within

each grade. I use the raw scores for the state examination because the state does not provide scaled scores for third grade students, an important year of baseline data. For consistency, I also use raw scores for the SRI and SAT, although I present evidence that the results are quite insensitive to decisions about test scaling. For each test, I standardize the appropriate score within grade and school year by subtracting the mean and dividing by the standard deviation to put all tests on a comparable scale with a mean of zero and a standard deviation of one.

The administrative dataset is quite extensive, allowing me to use a wide range of individual-level control variables (X_{it}), including indicators of the student's race, gender, federal free and reduced-price lunch status, language proficiency, and special education status and whether the student was enrolled in a gifted and talented program. For each student, I also generate a vector of class-level means to control for peer and classroom composition effects ($\bar{X}_{jt}$). These variables include average student race, gender, poverty status, language proficiency, and special education status as well as the class size and average student baseline test scores for each subject. Finally, I create a full set of mutually exclusive grade, year, and school indicators and, most importantly, a set of teacher indicator variables that uniquely identify a student's teacher in a specific subject and year.

Data Analysis

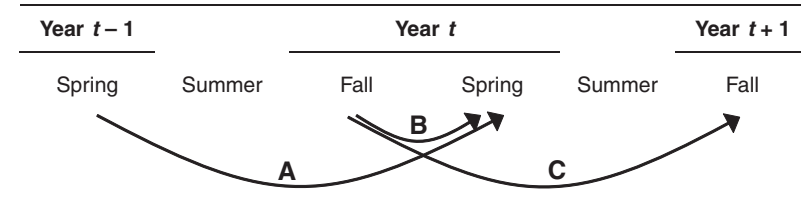
As discussed earlier, analysts have recommended and adopted a variety of different value-added models. I use a covariate-adjustment model that includes a baseline test score as a right-hand-side covariate in predicting the outcome test score. While I believe this approach provides a good balance between clarity and complexity, I use it here largely because of its popularity: it has a long history in the education research literature (Boardman & Murnane, 1979; Boyd et al., 2007; Cantrell, Fullerton, Kane, & Staiger, 2007; Gordon et al., 2006).

For each subject (mathematics and English language arts [ELA]) and each assessment, I fit a mixed model that represents the relationship between a student's standardized test score and a variety of predictors, including up to a fifth-order polynomial of their previous year's test score.[7] I estimate the following basic model:

$$Y_{ijt} = \alpha_g * f(Y_{i,t-1}) + X_{it}' \gamma + \overline{X}_{jt}' \varsigma + \delta_j + \varphi_k + \theta_g + \lambda_t + \varepsilon_{ijt} \qquad (7)$$

for student i with teacher j in school k, grade g, and year t. I allow the effects of the baseline test score to vary by the student's grade. I include school (φ_k), grade (θ_g), and year (λ_t) fixed effects. The objects of interest are the teacher effects (δ_j).

This equation represents the appropriate specification for the state test and the spring SRI. However, this pattern changes depending on the timing of the pre-test and the outcome. In Figure 1, I highlight the three possible



Option	Timing	Examples
A	Spring$_{t-1}$ to Spring$_t$	$\delta_{\text{STATE-SS}}$ and $\delta_{\text{SRI-SS}}$
B	Fall$_t$ to Spring$_t$	$\delta_{\text{SRI-FS}}$
C	Fall$_t$ to Fall$_{t+1}$	$\delta_{\text{SAT-FF}}$ and $\delta_{\text{SRI-FF}}$

Figure 1. **Diagram showing possible combinations of pretest and outcome tests to estimate the effect of a teacher in year *t*.**

pairs of tests that could be used to evaluate a teacher. For example, for value-added estimation using fall tests, like the SAT, the following fall's test score (SAT_{t+1}) is the outcome and the current year's fall test score (SAT_t) is the baseline test. I estimate effects in separate equations for each of the assessments, producing separate indicators of teacher effectiveness for each test. I estimate an effect for each teacher but omit the subscript *j* for clarity. Instead, I label these indicators with both the test name and the timing of baseline and outcome test administration—fall (F) or spring (S)—as subscripts. Thus, $\hat{\delta}_{SRI-SS}$ represents the set of teacher value-added estimates derived using the spring SRI test as the outcome and the previous spring SRI test as the baseline assessment.

For each subject, I include only students with non-missing baseline and outcome test data. I also exclude student-year records with missing teacher links and with more than one recorded teacher in the subject (approximately 3% of the sample). In each year, I drop students in very small or large classes as well as students in classes with large numbers of special needs or limited English proficient students.[8] Thus, my final sample includes teachers of traditional mathematics and ELA classes. In order to create the most precise estimates possible, I use all available data to estimate one set of teacher estimates for each outcome, rather than estimating teacher effects separately year by year. The substantive findings using only 1 year of data are quite similar to those using data pooled across years.

In Table 1, I present sample sizes by year and grade for each of the value-added measures. The top panel includes the total number of students in the district, while the subsequent panels present the student sample sizes

used in the value-added calculations. Given the need to have baseline and outcome test information, the grades in which value-added models can be estimated vary across the assessments. In general, my final models include 50% to 70% of all students in the tested grades.[9] In Table 2, I summarize this information, showing the total number of teachers for whom I create value-added estimates and the average number of students per teacher. In all cases, I use at least 20,088 student-level records and estimate effects for at least 526 teachers. These estimates rely on average sample sizes per teacher ranging from 33 to 42 students, with more than 2 years of data for the average teacher. For my primary analyses, I compare teacher effects derived from these full samples because they represent the estimates of teacher performance that a school district would use. At the end of the article, I explore whether these estimates vary because of the different samples of students who take each test.

This general model frames the remaining analyses. To assess the robustness of my findings to model specification, I create nine separate models that include different combinations of individual-level, classroom-level, and school-level controls; in some models, I include school fixed effects, while others contain school-level averages. These nine models essentially represent different comparison groups for teachers in the district. I also model teacher effects, δ_j, as both shrunken and unshrunken. Given the prominence of random effects models that produce empirical Bayes shrinkage estimators, I report my primary results using these random (shrunken) teacher effects. In these models, I include teacher-year effects to account for transitory differences in teacher performance. In general, the unshrunken fixed effects models produce less stable estimates and reduce the correlations across measures substantially in several of the specifications. However, the main patterns remain unchanged.

Throughout most of the article, I report estimates using the full model, including the school fixed effects and student and classroom demographic characteristics shown in Equation (7), which I call model M1. I find that estimates of the variability of teacher effects differ across the tests. The standard deviation of teacher effects from the state test ($\delta_{\text{STATE-SS}}$) is 0.19 and from the spring SRI ($\delta_{\text{SRI-SS}}$) is 0.21, both of which are quite similar to the results from prior studies (e.g., Boyd et al., 2007; Gordon et al., 2006; Rivkin et al., 2005; Rockoff, 2004). The standard deviation of teacher effects derived from the fall tests is smaller, 0.12 for $\delta_{\text{SRI-FS}}$, 0.13 for $\delta_{\text{SRI-FF}}$, and 0.05 for $\delta_{\text{SAT-FF}}$.[10] I return to this finding later in discussing the issue of test timing. Given the different amounts of variation in these teacher effects and that the goal of most accountability systems involves some form of ranking teachers, I present comparisons between estimates using Spearman's rank correlations in most instances. However, the findings do not differ substantively with Pearson's correlations.

Table 1

Total Number of Students in the District, by Grade and Year (top panel), With Sample Sizes Used to Construct Value-Added Estimates for Each Outcome (subsequent panels)

	School Year					
	2002–2003	2003–2004	2004–2005	2005–2006	2006–2007	Total
District total						
Grade 3	3,742	4,699	4,182	4,066	4,195	20,884
Grade 4	3,892	4,596	4,435	4,138	3,985	21,046
Grade 5	3,498	4,725	4,276	4,160	3,946	20,605
Total	11,132	14,020	12,893	12,364	12,126	62,535
SAT math (FF)						
Grade 3	2,439	3,062	2,676	2,606	2,810	13,593
Grade 4	2,599	3,006	2,916	2,605	2,512	13,638
Grade 5	0	0	0	0	0	0
Total	5,038	6,068	5,592	5,211	5,322	27,231
State ELA (SS)						
Grade 3	0	0	0	0	0	0
Grade 4	2,807	3,272	3,172	2,699	2,658	14,608
Grade 5	0	0	0	2,782	2,698	5,480
Total	2,807	3,272	3,172	5,481	5,356	20,088
SAT reading (FF)						
Grade 3	2,440	3,061	2,660	2,603	2,801	13,565
Grade 4	2,628	3,015	2,928	2,602	2,509	13,682
Grade 5	0	0	0	0	0	0
Total	5,068	6,076	5,588	5,205	5,310	27,247
SRI reading (SS)						
Grade 3	394	1,377	1,606	1,234	1,044	5,655
Grade 4	1,034	1,483	2,116	1,562	1,136	7,331
Grade 5	2,011	2,700	2,651	1,897	1,420	10,679
Total	3,439	5,560	6,373	4,693	3,600	23,665
SRI reading (FS)						
Grade 3	1,542	2,106	1,821	1,489	1,183	8,141
Grade 4	2,547	2,998	2,978	2,081	1,531	12,135
Grade 5	2,159	2,943	2,886	1,968	1,470	11,426
Total	6,248	8,047	7,685	5,538	4,184	31,702
SRI reading (FF)						
Grade 3	1,487	2,000	1,549	1,107	1,188	7,331
Grade 4	2,392	2,782	2,265	1,378	1,220	10,037
Grade 5	1,850	2,540	2,468	0	0	6,858
Total	5,729	7,322	6,282	2,485	2,408	24,226

Note. SAT = Stanford Achievement Test; SRI = Scholastic Reading Inventory; FF = fall-to-fall; SS = spring-to-spring; FS = fall-to-spring; ELA = English language arts.

Table 2

Total Student-Years, Teachers, and Teacher-Years Included in the Final, Full Sample Value-Added Models, From 2002–2003 to 2006–2007 for Teachers in Grades 3 Through 5

Value-Added Test	Student-Year Records	Unique Teachers	Teacher-Years	Average Students per Teacher	Average Students per Teacher-Year
SAT math (FF)	27,231	663	1,625	41.1	16.8
State ELA (SS)	20,088	526	1,148	38.2	17.5
SAT reading (FF)	27,247	663	1,623	41.1	16.8
SRI (SS)	23,665	702	1,513	33.7	15.6
SRI (FS)	31,702	762	1,786	41.6	17.8
SRI (FF)	24,226	738	1,558	32.8	15.5

Note. SAT = Stanford Achievement Test; SRI = Scholastic Reading Inventory; FF = fall-to-fall; SS = spring-to-spring; FS = fall-to-spring; ELA = English language arts.

Replication of Lockwood et al.'s Analysis of Mathematics Results

My main mathematics results confirm the analysis of Lockwood and colleagues (2007), who compared teacher effects estimated using different subscales of the mathematics SAT. Lockwood et al. report on one cohort of 2,855 students in Grades 6, 7, and 8 to estimate effects for 71 teachers across 2 years. They compare four types of models with several different specifications of demographic controls and find that "the sensitivity of the estimates to MODEL [model selection] and CONTROLS [choice of demographic controls] is only slight compared to their sensitivity to the achievement outcome" (p. 55). My analysis uses a larger dataset and incorporates a wider range of statistical controls. It also extends this work to earlier grades in a different setting. I estimate effects for 663 teachers using nearly 10 times as many students over a longer time period. I include combinations of individual-level demographic controls, classroom-level covariates, and school fixed effects or school-level covariates, resulting in nine different specifications for the covariate-adjustment model.

Like Lockwood et al., I find that teacher estimates are more sensitive to the subtest choice than the model specification. I find a minimum Spearman's rank correlation of 0.77 across the different model specifications, with the greatest difference arising from the decision to include or exclude school fixed effects. However, correlations of estimates using the two different subtests (from Table 3) range from 0.52 to 0.65, depending on the specification. Thus, the choice of subtest makes a greater difference in teacher rankings than the inclusion of school-, classroom-, or individual-level covariates. This finding does not imply that covariate choice has little impact; on the contrary, a correlation of 0.77 between a model with a full set of

Table 3
Spearman Rank Correlations Between Teacher Value-Added Estimates From the SAT Mathematics Subtests, by Model

	Controls				
Model	Student	Class	School Effects	School Means	Corr($\delta_{SAT\_1\text{-}FF}$, $\delta_{SAT\_2\text{-}FF}$)
M1	x	x	x		0.52
M2		x	x		0.52
M3	x		x		0.53
M4			x		0.58
M5	x	x		x	0.55
M6		x		x	0.54
M7	x			x	0.56
M8				x	0.59
M9					0.65

Note. Estimates represent teacher random effects derived from different specifications of Equation (7) (n = 663 teachers). SAT = Stanford Achievement Test; FF = fall-to-fall; SS = spring-to-spring.

demographic controls and one with none suggests that the choice of a comparison group makes an important difference in teacher rankings. However, in all cases, the choice of subtest has an even greater impact.

Examining the sensitivity of teacher effectiveness estimates to outcome choice using these subscales proves interesting and useful for a variety of reasons. Because the subsections come from the same test given at the same time, many of the sources of error that could produce variation between two outcomes are eliminated. These subtests have the same scale and a wide range of issues, such as item format, are likely constant across the sections. Furthermore, because teachers would likely not face incentives to coach towards only one test section, differential instructional responses to these individual subtests are less likely to drive any differences. Finally, the effects of some classroom-level shocks in student performance may be eliminated because students take both sections at the same time under the same testing conditions. Thus, we can use these subtest analyses to essentially isolate the effects of test inconsistency and different test content on teacher value-added estimates, and we find that these two issues contribute to substantial variation in estimated effectiveness.

On the other hand, these subscales clearly—and intentionally—assess different content dimensions: The Procedures and Problem Solving subsections call on students to use different skills and to know different things. The assumption that these tests measure a unidimensional "mathematics" construct is particularly tenuous. Thus, using complete assessments that attempt to cover broader domains of knowledge may provide a comparison

Table 4

Spearman Rank Correlations Between Teacher Value-Added Estimates From the SAT, State Test, and SRI Reading Assessments, by Model

	Controls						
Model	Student	Class	School Effects	School Means	Corr($\delta_{STATE\text{-}SS}$, $\delta_{SAT\text{-}FF}$)	Corr($\delta_{SAT\text{-}FF}$, $\delta_{SRI\text{-}FF}$)	Corr($\delta_{STATE\text{-}SS}$, $\delta_{SRI\text{-}SS}$)
M1	x	x	x		0.16	0.27	0.44
M2		x	x		0.18	0.23	0.45
M3	x		x		0.15	0.26	0.46
M4			x		0.21	0.28	0.51
M5	x	x		x	0.20	0.27	0.45
M6		x		x	0.21	0.23	0.45
M7	x			x	0.20	0.27	0.46
M8				x	0.25	0.28	0.51
M9					0.36	0.40	0.58
Teachers					395	541	429

Note. Estimates represent teacher random effects derived from different specifications of Equation (7). SAT = Stanford Achievement Test; SRI = Scholastic Reading Inventory; FF = fall-to-fall; SS = spring-to-spring.

of tests with more similar content. It also represents a more policy-relevant comparison demonstrating the impact of outcome selection.

Comparison of Teacher Value-Added Estimates From Three Complete Reading Tests

The analysis of estimated teacher effects using three different reading achievement measures closely matches the results for the mathematics subtests above. As Table 4 shows, depending on the specification, the Spearman correlations between $\delta_{STATE\text{-}SS}$ and $\delta_{SAT\text{-}FF}$ range from 0.15 to 0.36; between $\delta_{STATE\text{-}SS}$ and $\delta_{SRI\text{-}SS}$ from 0.44 to 0.58; and between $\delta_{SAT\text{-}FF}$ and $\delta_{SRI\text{-}FF}$ from 0.23 to 0.40.[11] These correlations are all statistically significant and most are moderately sized, suggesting that, on average, teachers whose students perform well on one test tend to perform well on other tests. However, they are sufficiently low that they produce substantially different classifications of many individual teachers. Thus, interpreting these correlations depends in large part on the relevant inferences to be drawn from them.

To explore this issue in more detail, I examine the distributions of teacher effects from these tests rather than a single correlation. In Figure 2, I present a scatterplot showing the relationship between $\delta_{STATE\text{-}SS}$ and $\delta_{SRI\text{-}SS}$, the two estimates with the greatest correlation, with a linear best fit overlaid. This figure shows clearly the positive relationship between these estimates but also

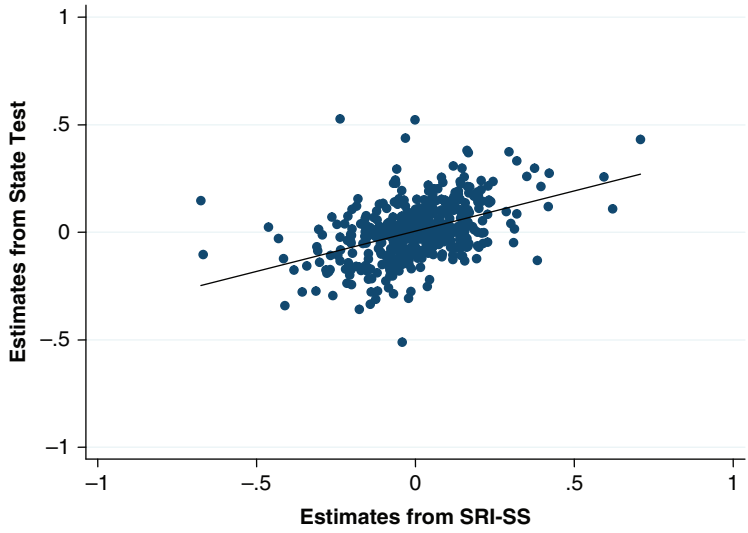


Figure 2. **Scatterplot showing the relationship between $\delta_{STATE\text{-}SS}$ and $\delta_{SRI\text{-}SS}$, with a linear best fit overlaid. Estimates represent teacher random effects derived from Equation (7), model M1, with school fixed effects and individual and classroom-level controls (n = 429).**

illustrates that individual teachers may rank quite differently depending on the outcome.

These inconsistent classifications would have substantial consequences for any policy that rewards teachers based on their value-added scores. For example, Houston Independent School District has developed a pay-for-performance program, called ASPIRE, that offers teachers bonuses for student test performance. In 2008–2009, Houston teachers could earn up to $10,300 in school-wide and individual bonus pay for both student achievement levels and progress. For core subject teachers in Grades 3 through 8, by far the largest component of the reward program is the Teacher Progress Award, which provides a bonus of $7,000 to teachers whose value-added scores rank them in the top 25% in the district and $3,500 for teachers in the next 25%. In Table 5, I present a transition matrix that demonstrates how this policy would play out with two different outcome measures, the state test and the SRI. The top panel of Table 5 arrays teachers by quartiles on both $\delta_{STATE\text{-}SS}$ and $\delta_{SRI\text{-}SS}$, while the bottom panel demonstrates how much bonus pay teachers would receive.

As seen in Table 5, approximately half of the teachers who would earn a $7,000 bonus using the state test would lose money if the district used the SRI instead. In fact, one in four of these teachers would lose their entire

Table 5

Transition Matrices Showing Row Percentages of Teacher Effectiveness Estimate Quartiles (top panel) and Hypothetical Bonus Amounts (bottom panel) Derived From the State Test and the SRI

Quartiles of $\delta_{STATE\text{-}SS}$ vs. $\delta_{SRI\text{-}SS}$

	$\delta_{SRI\text{-}SS}$ Quartiles				
$\delta_{STATE\text{-}SS}$ Quartiles	Top	Q2	Q3	Bottom	Total
Top	53.1%	22.1%	16.8%	8.0%	113
Q2	21.8%	26.4%	30.0%	21.8%	110
Q3	22.9%	32.4%	21.0%	23.8%	105
Bottom	6.9%	8.9%	32.7%	51.5%	101
Total	115	97	107	110	429

Bonus Amounts for $\delta_{STATE\text{-}SS}$ vs. $\delta_{SRI\text{-}SS}$

	Bonus Using SRI			
Bonus Using State Test	$7,000	$3,500	None	Total
$7,000	53.1%	22.1%	24.8%	113
$3,500	21.8%	26.4%	51.8%	110
None	15.0%	20.9%	64.1%	206
Total	115	97	217	429

Note. Quartiles may be different sizes because different teachers have estimates for different combinations of tests. Estimates represent teacher random effects derived from Equation (7), Model M1. SRI = Scholastic Reading Inventory; SS = spring-to-spring.

bonus. Similarly, 36% of teachers ranked in the bottom half on the state test—teachers whose students made below-average progress during the year—would have earned a bonus using the SRI. In general, about half of all teachers would receive the same bonus using either measure, while 25% would earn more and 25% would earn less depending on the test chosen. The average teacher in the district would see his or her pay changed by $2,178 simply by switching outcome measures. Interestingly, the instability in teacher estimates across outcome measures is much greater for teachers in the middle two quartiles. Importantly, I compare $\delta_{STATE\text{-}SS}$ and $\delta_{SRI\text{-}SS}$, the estimates with the greatest correlation in my analysis. Using $\delta_{STATE\text{-}SS}$ and $\delta_{SAT\text{-}FF}$ instead produces even more instability, with nearly 60% of teachers earning a different bonus and an average change in compensation of more than $2,700.

Table 6

Spearman Rank Correlations Between Teacher Value-Added Estimates From the State Test Across Model Specifications With Different Statistical Controls

Model	M1	M2	M3	M4	M5	M6	M7	M8	M9
M1	1.00								
M2	0.99	1.00							
M3	0.99	0.99	1.00						
M4	0.95	0.97	0.97	1.00					
M5	0.89	0.89	0.88	0.86	1.00				
M6	0.89	0.90	0.89	0.88	0.99	1.00			
M7	0.88	0.88	0.89	0.87	0.99	0.99	1.00		
M8	0.86	0.88	0.88	0.90	0.96	0.97	0.97	1.00	
M9	0.79	0.80	0.81	0.84	0.88	0.88	0.89	0.91	1.00
Student controls	x		x		x		x		
Classroom controls	x	x			x	x			
School effects	x	x	x	x					
School averages					x	x	x	x	

Note. Estimates represent teacher random effects derived from different specifications of Equation (7) (n = 526 teachers).

The Effects of Model Specification

As in the mathematics analysis, the differences in teacher value-added estimates in reading that arise from model specification are quite small compared to those produced by different outcome measures. In Table 6, I present a correlation matrix for $\delta_{\text{STATE-SS}}$ from the different specifications. The minimum correlation across these models exceeds 0.79. Again, the most important difference across models comes from the inclusion/exclusion of school fixed effects, a result that aligns with Harris and Sass's (2006) earlier analysis and that makes intuitive sense—it is not surprising that comparing teachers within schools provides somewhat different answers about performance than comparing teachers across the district. Among models that control for school fixed effects, the inclusion of student and/or classroom-level covariates makes little difference, with correlations ranging from 0.95 to 0.99. If school fixed effects are omitted, the decision to include student and classroom-level controls has a similarly minor effect. By contrast, the correlations between estimates from models that include and exclude school fixed effects are smaller. For example, simply replacing school fixed effects (model M1) with school-level covariates (model M5) reduces the correlation between models to 0.89. In all cases, though, the level of instability in teacher effectiveness estimates that results from outcome choice far outweighs the instability produced by model selection.

Accounting for Differences in Teacher Effects Across Outcomes

That different measures of reading achievement produce substantively different estimates of individual teacher performance raises important questions about why such variation arises. Tests are noisy measures of latent student achievement in a variety of ways. As discussed earlier, tests differ in their content, scaling, samples of students who take them, and timing, all of which can introduce inconsistency into estimates of true teacher effectiveness. Furthermore, measurement error itself can contribute substantially to this instability. Understanding which of these components produces the differences across outcome measures is clearly important for policy decisions. For example, if these estimates are simply different because tests measure different content, policymakers can choose the outcome measure that aligns most closely with their intended goals. However, if measurement error produces these differences, the variation across estimates from different outcome measures is likely a source of real concern if school officials rely on these measures to reward or sanction teachers.

While precisely disentangling the relative contributions of these sources proves difficult, the presence of each provides different, and testable, empirical implications. I find that differences in test content and scaling do not appear to explain the variation in teacher effects across outcomes in this district. The different samples of students who take each of the tests contribute somewhat, but they do not account for most of the differences. Test timing appears to play a greater role in producing these differences. Nonetheless, it does not explain all of the variation, suggesting that measurement error also contributes to the instability in teacher rankings.

Test Content

Teacher effectiveness estimates could differ by outcome measure if the two tests assess distinct domains of reading achievement or weight several domains differently. Although there is no simple test to determine if two tests measure the same construct (i.e., if $\eta = 0$), I can examine the influence of test content in several ways. Most importantly, all three tests have similar stated goals about measuring students' reading competency. Furthermore, correlations in student-level test scores between these measures, disattenuated for measurement error, are approximately 0.80 across the three tests.[12] This correlation likely understates the true relationship for two reasons. First, traditional disattenuation uses internal consistency reliability as an estimate of measurement error; this approach ignores other idiosyncratic individual- and classroom-level sources of error, such as a disruptive student in the class on test day, that are not reflected in inconsistent performance on different test items. Second, cross-sectional correlations may mask variation over time because of coaching towards specific test components.

Table 7

Student-Level Raw Pearson's and Spearman's Correlations Between Tests Over Time, Across All Years and Grades (not disattenuated for measurement error)

Comparison	Pearson	Spearman	*n*
STATE-S_t and STATE-S_{t+1}	0.71	0.72	25,504
STATE-S_t and SRI-S_{t+1}	0.67	0.68	18,846
STATE-S_t and SAT-F_{t+2}	0.67	0.70	16,387
SAT-F_t and SAT-F_{t+1}	0.74	0.76	33,032
SAT-F_t and STATE-S_t	0.73	0.74	42,657
SAT-F_t and SRI-S_t	0.73	0.74	28,126
SRI-S_t and SRI-S_{t+1}	0.57	0.55	35,854
SRI-S_t and STATE-S_{t+1}	0.57	0.56	21,996
SRI-S_t and SAT-F_{t+2}	0.51	0.52	15,641

Note. Because the SAT is administered in the fall, the time period $t+1$ corresponds to the students' performance at the end of year t and period $t+2$ corresponds to performance at the end of year $t+1$. SAT = Stanford Achievement Test; SRI = Scholastic Reading Inventory.

As a result, I look at correlation over time between pairs of tests: If two tests measure different content, the correlation over time between student performance on one test (test A) should exceed the correlation between two different tests (A and B). In other words, Corr(A_t, A_{t+1}) > Corr (A_t, B_{t+1}). I find little evidence that the three assessments measure substantially different domains of reading achievement. In Table 7, I present raw Pearson's and Spearman's correlations between tests over time, not disattenuated for measurement error; in all cases, correlations between the same test do not systematically or substantially exceed correlations between two different tests. For example, the Pearson's correlation in student-level test scores between the state test across 2 years, 0.71, is quite similar to the correlation over time between the state test and the SRI (0.67) or the SAT (0.67). Similar patterns arise when examining the other assessments. Again, this analysis does not provide conclusive evidence that differences in test content are not producing the inconsistency in value-added estimates across assessments, but the relative stability of these student-level correlations across tests both in the same time period and over time suggests that content differences appear to be a second-order concern.

Test Scaling

Because the underlying relationship between student test performance and true proficiency is unknown, decisions about test scaling remain rather arbitrary. If estimates were to depend substantially on test scaling, that

Table 8
Spearman's Rank Correlations Across Teacher Effectiveness Measures, in the Full and Comparison Samples From Different Outcomes

Comparison	Correlation (full sample)	Correlation (comp. sample)	Comparison Sample Sizes: Students	Comparison Sample Sizes: Teachers
$\delta_{STATE\text{-}SS}$ and $\delta_{SAT\text{-}FF}$	0.16	0.15	12,850	388
$\delta_{STATE\text{-}SS}$ and $\delta_{SRI\text{-}SS}$	0.44	0.54	10,450	388
$\delta_{SAT\text{-}FF}$ and $\delta_{SRI\text{-}FF}$	0.27	0.32	16,658	521

Note. Estimates represent teacher random effects derived from Equation (7), model M1, with school fixed effects and individual- and classroom-level controls. SAT = Stanford Achievement Test; SRI = Scholastic Reading Inventory; FF = fall-to-fall; SS = spring-to-spring.

would raise concern that arbitrary choices may be driving teacher effects. For both the SAT and the SRI, though, test scaling affects the rankings of teacher estimates seen here only minimally. Correlations between teacher effects using raw and scaled scores exceed 0.98 for these tests. Such high correlations mirror the results that Briggs, Weeks, and Wiley (2008) found in comparing eight different scales on a single assessment in Colorado for school-level value-added estimates.

Sample of Students

Because schools administer tests on different days, student absenteeism and mobility can create variation in the sample of students who take each assessment. Most value-added measures require scores from two school years. For example, estimates of $\delta_{STATE\text{-}SS}$ for 2006–2007 include all students with valid test scores in March 2006 and March 2007, while estimates of $\delta_{SAT\text{-}FF}$ in the same year include students with valid October 2006 and October 2007 scores. The sample of students used to estimate $\delta_{STATE\text{-}SS}$ might be quite different from that of $\delta_{SAT\text{-}FF}$ because students leave the district or are absent on one of the test days. As a result, teacher estimates may differ.

To examine this issue, I develop a different "comparison sample" for each pair of teacher estimates. Each sample includes only students for whom I have a valid baseline and outcome measure on both sets of assessments. These comparison samples are generally much smaller than the full samples used above. However, examination of student covariates reveals only minor differences across these samples, suggesting that students in the comparison sample are not systematically different on observable characteristics.

As Table 8 shows, teacher estimates derived from these "comparison samples" tell much the same story as the earlier analysis in Table 4. In general, teacher value-added estimates that use a common sample of students are

somewhat greater than those including all students, suggesting that the students who take the tests matter. For example, using the comparison sample increases the Spearman rank correlation between $\delta_{STATE\text{-}SS}$ and $\delta_{SRI\text{-}SS}$ approximately 20%, from 0.44 to 0.54. In general, though, the different samples of students taking each assessment do not explain the substantial variation in teacher effectiveness estimates. Even a correlation of 0.54 between measures derived from two tests represents a great deal of inconsistency for evaluating teacher performance.

Test Timing

Finally, differences in value-added estimates for teachers may arise from the timing of achievement tests. As described earlier, test timing can affect teacher value-added estimates in several ways. First, the amount of time a student has to learn with the teacher of record varies depending on the test date. Second, the specific baseline-outcome test combination used determines whether summer learning loss (or gain) factors into the estimates. The impact of test timing on teacher effectiveness estimates is a very important but understudied area in value-added research. McCaffrey et al. (2004) report on a small simulation which suggests that test timing may matter little. The unique testing patterns in my data allow me to examine this issue more closely, and I find more reason for concern.

Importantly, as described above, I find that estimates of the variability of teacher effects are smaller for fall assessments than for spring tests. This pattern could suggest that value-added estimates derived from these fall assessments may not be reliable indicators of teacher performance, perhaps because the underlying student test scores do not sufficiently distinguish between students because of summer learning loss.[13] However, because the district gives most students the SRI in both the fall and spring, I can compare teacher value-added estimates using the same test but different time periods (i.e., fall-to-fall, fall-to-spring, and spring-to-spring comparisons).

If test timing affects teacher estimates, correlations between models that use only fall scores or only spring scores should be higher than correlations comparing spring and fall estimates. As seen in Table 9, this prediction plays out. Here, the rank correlation between $\delta_{SAT\text{-}FF}$ and $\delta_{SRI\text{-}FF}$ (0.27) far exceeds that between $\delta_{SAT\text{-}FF}$ and $\delta_{SRI\text{-}SS}$ (0.12). Similar patterns emerge on the state test, with the correlation between $\delta_{STATE\text{-}SS}$ and $\delta_{SRI\text{-}SS}$ (0.44) much greater than between $\delta_{STATE\text{-}SS}$ and $\delta_{SRI\text{-}FF}$ (0.15).

Focusing only on the SRI, I also find that teacher effectiveness estimates derived from the fall and spring tests vary substantially. Because the SRI is scaled consistently and is designed to assess similar content across grades, these differences reflect primarily the effects of test timing and inconsistency. In the bottom panel of Table 9, I present Spearman rank correlations for estimates across time periods. Holding the sampling of students constant by looking at the comparison sample, I find that while the correlation between

Table 9

Spearman's Rank Correlations Across Teacher Effectiveness Measures, in the Full and Comparison Samples From Different Outcomes at Different Times

			Comparison Sample Sizes	
Comparison	Correlation (full sample)	Correlation (comp. sample)	Students	Teachers
$\delta_{STATE\text{-}SS}$ and $\delta_{SRI\text{-}SS}$	0.44	0.54	10,450	388
$\delta_{STATE\text{-}SS}$ and $\delta_{SRI\text{-}FF}$	0.15	a	a	a
$\delta_{SAT\text{-}FF}$ and $\delta_{SRI\text{-}FF}$	0.27	0.32	16,658	521
$\delta_{SAT\text{-}FF}$ and $\delta_{SRI\text{-}SS}$	0.12	0.16	11,510	452
$\delta_{SRI\text{-}SS}$ and $\delta_{SRI\text{-}FS}$	0.66	0.71	22,536	684
$\delta_{SRI\text{-}SS}$ and $\delta_{SRI\text{-}FF}$	–0.10	–0.06	15,074	624
$\delta_{SRI\text{-}FS}$ and $\delta_{SRI\text{-}FF}$	0.19	0.21	22,313	713

Note. Estimates represent teacher random effects derived from Equation (7), model M1, with school fixed effects and individual and classroom-level controls. FF = fall-to-fall; SS = spring-to-spring; FS = fall-to-spring.
[a]The comparison size is too small to estimate teacher effects using the full model.

$\delta_{SRI\text{-}SS}$ and $\delta_{SRI\text{-}FS}$ is 0.71, other correlations are substantially lower. For example, in the comparison sample, the correlation between $\delta_{SRI\text{-}FS}$ and $\delta_{SRI\text{-}FF}$ is 0.21. Interestingly, the correlation between $\delta_{SRI\text{-}SS}$ and $\delta_{SRI\text{-}FF}$ is –0.06: These two teacher estimates are essentially orthogonal. These comparisons suggest that summer learning loss (or gain) may produce important differences in teacher effects. Here, the fall-to-fall estimates attribute one summer's learning loss to the teacher, while the spring-to-spring estimates attribute a different summer's loss. Thus, the fact that the fall-to-fall and spring-to-spring estimates produce substantially different answers likely reflects, in part, the inclusion of a different summer in each estimate.

Conclusions and Implications

The analyses presented above suggest that the correlations between teacher value-added estimates derived from three separate reading tests—the state test, SRI, and SAT—range from 0.15 to 0.58 across a wide range of model specifications. Although these correlations are moderately high, these assessments produce substantially different answers about individual teacher performance and do not rank individual teachers consistently. Even using the same test but varying the timing of the baseline and outcome measure introduces a great deal of instability to teacher rankings. Therefore, if a school district were to reward teachers for their performance, it would identify a quite different set of teachers as the best performers depending simply on the specific reading assessment used.

These tests provide different answers about individual teacher performance for a variety of reasons. My results suggest that variation in test content and scaling across tests may play a small role in producing these differences, but I find little evidence that they contribute substantially, at least in the grades examined. In some cases, the specific sample of students that takes each test appears to make a greater difference, but teacher effectiveness estimates vary substantially even when derived from the same sample of students on different tests. The evidence presented here suggests that the real forces driving the variation in teacher effects across outcome measures appear to be the timing of the assessments and measurement error. A consistent finding in the value-added literature is that individual measurement error on student achievement tests and the presence of idiosyncratic classroom-level and school-level shocks to performance produce quite noisy estimates of teacher effectiveness. My results suggest that test timing also contributes substantially to differences in teacher effectiveness estimates across outcome measures. This is an important finding that merits further study.

These findings raise several important implications and questions for policy and practice. Most obviously, teacher value-added estimates are sensitive to many characteristics of the tests on which they are based. Thus, policymakers and practitioners who wish to use these estimates to make high-stakes decisions must think carefully about the consequences of these differences, recognizing that even decisions seemingly as arbitrary as when to schedule the test within the school year will likely produce variation in teacher effectiveness estimates. While local administrators and policymakers must decide their tolerance for misclassification, a system that would identify nearly half of all teachers as "high performing" on one test but not the other, as even the examinations with the most consistent estimates do in my data, is likely not sufficiently precise for rewarding or sanctioning teachers. If policymakers do hope to use pay-for-performance programs to motivate teachers, they must recognize that any incentive effects will be attenuated substantially by the amount of inconsistency in teacher value-added estimates.

However, this variation in estimates across different outcome measures does not suggest that value-added models should be abandoned entirely. On the contrary, the moderately high correlations between teacher effects across these outcomes suggest that value-added estimates can play an important role in understanding teacher performance. On average, teachers whose students perform well using one assessment also perform well using alternate tests. Thus, particularly in combination with other measures, value-added estimates may offer useful information about teacher quality. In particular, value-added estimates may contribute both to program evaluation research that combines estimates across a wider sample of teachers and to formative uses for improving teacher performance in schools. For example, although value-added models may not be sufficiently stable to reward individual teachers, they are likely more consistent in answering policy questions

that compare groups of teachers.[14] They may also be quite useful in more formative ways, helping principals and teachers identify areas for improvement within their schools or classes.

If policymakers intend to continue using value-added measures to make high-stakes decisions about teacher performance, more attention should be paid to the tests themselves. Currently, all value-added estimates of teacher effectiveness use tests designed to measure student, not teacher, performance. The ideal properties of tests designed to identify a district's best teachers may well differ from those designed to assess student proficiency. Furthermore, the timing of tests must be considered more carefully. For example, the practice of giving high-stakes tests in early spring may not matter much for inferences about student performance in the district—having an assessment of student skills in February may be just as useful as one in May. However, decisions about timing have substantial implications for teacher value-added estimation.

These findings also raise the question of whether using multiple assessments to generate teacher value-added effects may be of particular use to policymakers. For example, information about student achievement growth on several related tests could be combined to produce a more robust measure of teacher effectiveness, which would have several important advantages. First, it would increase the effective sample size by using information from three different tests. While it would not affect the sample of students in a particular teacher's classroom, it could help to eliminate measurement error arising from each specific assessment. It could also attenuate issues of test timing and sampling of students taking each test. Finally, and perhaps most importantly, it could encourage representation in the classroom of content across all tests, perhaps helping to reverse the narrowing of curriculum towards one assessment. Of course, any student assessments that will become part of teacher accountability programs should reflect valued skills that policymakers want students to learn over the course of the year. Domains that are represented on multiple assessments will draw additional instructional attention.

Clearly, even an accountability system based on several assessments would provide incentives for coaching and instructional responses on all tests in that subject, perhaps to the detriment of other instructional activities. Furthermore, while the ability to raise student test scores remains an important aspect of teacher quality, it represents only one of many key dimensions. Combining this test-based accountability with other performance measures, such as peer or principal evaluations, could further address these issues. Given the amount of inaccuracy in any single assessment of teacher performance—whether based on test scores or observations—combining multiple sources of information could provide schools and teachers with a better sense of their performance on a wider range of domains.

While multiple measures may provide a more robust assessment of teacher performance and may mitigate the effects of measurement error from using any single test, policymakers and district officials must take

care in deciding how to combine measures. Douglas (2007) found that using multiple assessments increases evaluation reliability when the measures are highly related, but this result is not consistent with less correlated measures. In a recent special issue of *Educational Measurement* on multiple measures, Chester (2003) agreed, arguing that the methods for combining the measures determine the reliability and validity of any decisions based on them. Thus, policymakers should carefully consider the specific approaches used to combine measures. Furthermore, officials must develop clear plans for addressing situations where different measures provide conflicting evidence about teacher effectiveness.

Importantly, additional research is needed into the different implications of high- and low-stakes tests for estimating teacher effects. Teachers who appear to perform well using a high-stakes examination but not well with a low-stakes test may be effectively teaching state standards or may be engaged in inappropriate coaching. No apparent patterns appear to distinguish the high-stakes state test as being substantially different from the low-stakes SAT or SRI in my analysis, but the role of stakes in driving any differences in teacher effects deserves further attention.

Many open questions remain concerning the reliability and validity of teacher value-added effects as causal estimates of a teacher's productivity. This article confirms earlier results and provides new evidence that the specific outcome used matters a great deal. Outcome choice produces substantially more variation in teacher effects than decisions about model specification. Importantly, the existing research concerning value-added models has focused largely on decisions about model specification. This article argues for further attention to the measures themselves.

Notes

The author thanks Daniel Koretz, Lawrence Katz, Richard Murnane, and Susan Moore Johnson for their helpful comments and suggestions on an earlier draft, as well as Robert Croninger and three anonymous reviewers for their thoughtful feedback. Financial support was provided through the Harvard Graduate School of Education Dean's Summer Fellowship.

[1]Although analysts also compute value-added estimates for individual schools, this article focuses on teacher-level estimates.

[2]This presentation implies that differences in test selection are random, or independent of true achievement. While this assumption may be tenuous, it is helpful for explication and does not underpin the empirical analyses here.

[3]Rogosa and Willett (1983) demonstrate that in extreme cases, particularly with large year-to-year gains, this pattern does not hold. In practice, though, and in the tests presented here, student gains from year-to-year are sufficiently small that the reliability of each individual test score is substantially greater than the reliability of the difference between them.

[4]A simple intuitive model involves decomposing the student's "teacher" into a sum of his or her teachers over the period, weighted by the number of days of instruction: $T^*_{it} = \sum_t \sum_d T_t D_d$, where T_t is the student's teacher in year t and D_d indicates whether the student was enrolled in a classroom with that teacher on a specific day d.

[5]This practice may also cause challenges if states change their test substantially between years.

[6]McCaffrey et al. (2003) argue that summer learning loss is correlated with observable student characteristics. Because students sort across schools and teachers, learning loss will affect some teacher estimates more than others.

[7]Some analysts recommend fitting more complicated models that use multiple test scores (e.g., both mathematics and reading tests) as predictors. Given the pattern of testing in the district, such specifications are only possible for the SAT, but the teacher value-added estimates that result from this more complicated specification are nearly identical to those from Equation (7). Importantly, I do not include teacher characteristics, such as years of service, because districts have begun using value-added approaches to understand—and reward—teacher performance regardless of experience.

[8]Here, I eliminate students outside the 1st and 99th percentiles in class size, or classes with fewer than 6 or more than 28 students, removing 1% of the sample. I drop classes with more than 50% special needs students and more than 75% LEP (limited English proficiency) students based on visual analysis of kernel densities, removing an additional 15% of the sample. Finally, for each assessment, I drop students in classes where fewer than 5 students had sufficient data to compute a value-added estimate for the teacher, excluding just 0.3% of the remaining sample.

[9]Missing data could affect teacher value-added estimates in several ways. Most importantly, to the extent that students with missing data are not missing completely at random, estimates could be biased. However, for this study, missing data concerns are less important because I mirror the practices that a district would use in constructing value-added estimates. A district would use available information for its teachers, as I do here. Furthermore, I assess the robustness of my results to sampling by using common samples of students, as described later in the article.

[10]These standard deviations are the square roots of the estimated variance components from my model. The sample standard deviations of the empirical Bayes shrinkage estimates are somewhat smaller, but the general pattern remains.

[11]I find nearly identical results when I compare teachers who appear in the data for at least 3 years, suggesting that the differences in correlations are not driven largely by teachers with little data.

[12]Across all years and grades, item-level analyses indicate that internal consistency reliability (measured by Cronbach's alpha) ranges from 0.91 to 0.95 for the state examination and from 0.95 to 0.96 for the SAT. I use average reliabilities of 0.925 for the state test and 0.955 for the SAT to disattenuate correlations for measurement error. These reliabilities compare favorably to figures provided by the test publishers. Because I do not have item-level data for the SRI, I assume a reliability of 0.90.

[13]Given the lower variability in fall-to-fall value-added estimates, restriction of range could also contribute to lower correlations that include fall-to-fall measures.

[14]McCaffrey, Sass, and Lockwood (2008) find relatively similar results concerning the intertemporal stability of value-added estimates using two different achievement measures. However, Harris and Sass (2009) find some more substantial differences in their estimated effect of National Board for Professional Teaching Standards certification using two different tests in Florida.

References

Ballou, D., Sanders, W., & Wright, P. (2004). Controlling for student background in value-added assessment of teachers. *Journal of Educational and Behavioral Statistics*, *29*(1), 37–65.

Boardman, A. E., & Murnane, R. J. (1979). Using panel data to improve estimates of the determinants of educational achievement. *Sociology of Education*, *52*(2), 113–121.

Boyd, D., Grossman, P., Lankford, H., Loeb, S., & Wycoff, J. (2006). How changes in entry requirements alter the teacher workforce and affect student achievement. *Education Finance and Policy*, *1*(2), 176–216.

Boyd, D., Grossman, P., Lankford, H., Loeb, S., & Wycoff, J. (2007). *Who leaves? Teacher attrition and student achievement*. Unpublished manuscript.

Briggs, D. C., Weeks, J. P., & Wiley, E. (2008). *The sensitivity of value-added modeling to the creation of a vertical score scale* (Working paper for the National Conference on Value-Added Modeling). Retrieved from http://www.wcer.wisc.edu/news/events/natConf_papers.php

Cantrell, S., Fullerton, J., Kane, T. J., & Staiger, D. O. (2007). *National Board Certification and teacher effectiveness: Evidence from a random assignment experiment*. Unpublished manuscript.

Chester, M. D. (2003). Multiple measures and high-stakes decisions: A framework for combining measures. *Educational Measurement: Issues and Practice*, *22*(2), 32–41.

Clotfelter, C. T., Ladd, H. F., & Vigdor, J. L. (2007). *How and why do teacher credentials matter for student achievement?* (NBER Working Paper No. 12828). Cambridge, MA: National Bureau of Economic Research.

Croninger, R. G., & Valli, L. (2009). "Where is the action?" Challenges to studying the teaching of reading in elementary classrooms. *Educational Researcher*, *38*(2), 100–108.

Douglas, K. M. (2007). *A general method for estimating the classification reliability of complex decisions based on configural combinations of multiple assessment scores*. Unpublished doctoral dissertation, University of Maryland, College Park, MD.

Gewertz, C. (2007). N.Y.C. district issues "value added" grades for schools. *Education Week*, *27*(12), 6.

Goldhaber, D., & Anthony, E. (2007). Can teacher quality be effectively assessed? National Board Certification as a signal of effective teaching. *Review of Economics and Statistics*, *89*(1), 134–150.

Gordon, R., Kane, T. J., & Staiger, D. O. (2006). *Identifying effective teachers using performance on the job*. Washington, DC: Brookings Institution.

Hanushek, E. (1971). Teacher characteristics and gains in student achievement: Estimation using micro data. *The American Economic Review*, *61*(2), 280–288.

Harris, D. N., & Sass, T. R. (2006). *Value-added models and the measurement of teacher quality*. Unpublished manuscript.

Harris, D. N., & Sass, T. R. (2009). The effects of NBPTS-certified teachers on student achievement. *Journal of Policy Analysis and Management*, *28*(1), 55–80.

Jacob, B. A., & Lefgren, L. (2005). *Principals as agents: Subjective performance measurement in education* (NBER Working Paper No. 11463). Cambridge, MA: National Bureau of Economic Research.

Kane, T. J., Rockoff, J. E., & Staiger, D. O. (2008). What does certification tell us about teacher effectiveness? Evidence from New York City. *Economics of Education Review*, *27*(6), 615–631.

Kane, T. J., & Staiger, D. O. (2002). *Volatility in school test scores: Implications for test-based accountability systems* (Brookings Papers on Education Policy). Washington, DC: Brookings Institution.

Koretz, D. M. (2002). Limitations in the use of achievement tests as measures of educators' productivity. *The Journal of Human Resources*, *37*(4), 752–777.

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

McCaffrey, D. F., Lockwood, J. R., Koretz, D. M., & Hamilton, L. S. (2003). *Evaluating value-added models for teacher accountability*. Santa Monica, CA: RAND.

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

McCaffrey, D. F., Sass, T. R., & Lockwood, J. R. (2008). *The intertemporal stability of teacher effect estimates*. National Center for Performance Incentives, Working Paper No. 2008-22. Retrieved from http://www.performanceincentives.org/data/files/news/PapersNews/McCaffrey_et_al_2008.pdf

McNeil, M. (2009). Starting gun sounds for "Race to the Top." *Education Week*, *29*(12), 1,18–19.

Murnane, R. J. (1984). Selection and survival in the teacher labor market. *The Review of Economics and Statistics*, *66*(3), 513–518.

Olson, L. (2007). Houston board OKs revamped performance-pay plan. *Education Week*, *27*(4), 11.

Raudenbush, S. W. (2004). What are value-added models estimating and what does this imply for statistical practice? *Journal of Educational and Behavioral Statistics*, *29*(1), 121–129.

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

Rockoff, J. E. (2004). The impact of individual teachers on student achievement: Evidence from panel data. *The American Economic Review*, *94*(2), 247–252.

Rogosa, D. R., & Willett, J. B. (1983). Demonstrating the reliability of the difference score in the measurement of change. *Journal of Educational Measurement*, *20*(4), 335–343.

Rothstein, J. (2007). *Do value-added models add value? Tracking, fixed effects, and causal inference*. Unpublished manuscript.

Rubin, D. B., Stuart, E. A., & Zanutto, E. L. (2004). A potential outcomes view of value-added assessment in education. *Journal of Educational and Behavioral Statistics*, *29*(1), 103–116.

Sanders, W. (2000). Value-added assessment from student achievement data: Opportunities and hurdles. *Journal of Personnel Evaluation in Education*, *14*(4), 329–339.

Sanders, W. L., & Horn, S. P. (1994). The Tennessee Value-Added Assessment System (TVAAS): Mixed-model methodology in educational assessment. *Journal of Personnel Evaluation in Education*, *8*(3), 299–311.

Sass, T. R. (2008). *The stability of value-added measures of teacher quality and implications for teacher compensation policy*. National Center for Analysis of Longitudinal Data in Education Research Brief No. 4. Retrieved from http://www.urban.org/UploadedPDF/1001266_stabilityofvalue.pdf

Tekwe, C. D., Carter, R. L., Ma, C., Algina, J., Lucas, M. E., Roth, J., et al. (2004). An empirical comparison of statistical models for value-added assessment of school performance. *Journal of Educational and Behavioral Statistics*, *29*(1), 11–35.

Todd, P. E., & Wolpin, K. I. (2003). On the specification and estimation of the production function for cognitive achievement. *Economic Journal*, *113*(485), F3–33.

Manuscript received January 31, 2009
Final revision received December 5, 2009
Accepted December 24, 2009