Using Subject Test Scores Efficiently to Predict Teacher Value-Added

Lars Lefgren and David Sims

EDUCATIONAL EVALUATION AND POLICY ANALYSIS 2012 34: 109
DOI: 10.3102/0162373711422377

Educational Evaluation and Policy Analysis
March 2012, Vol. 34, No. 1, pp. 109–121
DOI: 10.3102/0162373711422377

Using Subject Test Scores Efficiently to Predict Teacher Value-Added

Lars Lefgren
David Sims
Brigham Young University

This article develops a simple model of teacher value-added to show how efficient use of information across subjects can improve the predictive ability of value-added models. Using matched student–teacher data from North Carolina, we show that the optimal use of math and reading scores improves the fit of prediction models of overall future teacher value-added by up to a third for reading and a tenth for a composite measure (math and reading combined). Efficiency gains are greatest when value-added must be calculated on only 1 or 2 years of data. The methods employed are flexible and can be expanded to incorporate information from other subject or subitem test metrics.

Keywords: *teacher quality, teacher evaluation, value-added modeling*

THE current consensus in education research suggests that whereas true teacher quality is both highly variable and important to student achievement (e.g., Aaronson, Barrow, & Sander, 2007; Rivkin, Hanushek, & Kain, 2005; Rockoff, 2004) it is hard to discern through the current screening methods employed by school districts at the time of teacher hiring (Ballou 1996; Gordon, Kane, & Staiger, 2006). Because a more efficient system of teacher evaluation could provide gains to students on the order of hundreds of thousands of dollars (Staiger & Rockoff, 2010), academic researchers have recently gone to great lengths to seek out both better quality indicators that might be used in hiring (Rockoff, Jacob, Kane, & Staiger, 2011) and better measures of on-the-job performance either through administrative evaluation (Rockoff & Speroni, 2010) or through teacher value-added measures based on student test score data.

Although an array of statistical criticisms have arisen concerning the interpretation of these value-added models (e.g., Andrabi, Das, Khwaja, & Zajonc, 2011; Rothstein, 2010; Todd & Wolpin, 2003, 2006), there has also been recent scholarly support for the idea that lowering entry barriers to teaching along with a removal of the lowest performing teachers, even using imperfect value-added metrics, will lead to important gains in student learning over time (Goldhaber & Hansen, 2010; Hanushek, 2009). Given this movement to use value-added scores for teacher deselection, it seems likely that some variant of value-added methodology will play an increasingly important role in teacher and school accountability (Harris, 2009).

While an ideal measure of teacher effectiveness would measure the ability of a teacher to improve the full spectrum of student learning outcomes, current academic studies almost exclusively calculate teacher value-added to a specific test subject (e.g. mathematics or language arts). It is unlikely, however, that each subject value-added is equally informative about overall teacher effectiveness. Indeed, important studies such as

those by Gordon et al. (2006) and Goldhaber and Hansen (2010) implicitly reflect a general belief that math value-added will be a more important predictor of future teaching ability than reading value-added by using exclusively math value-added scores as outcomes.

Most contemporary value-added papers ignore potential benefits of using information on a teacher's performance in one domain to predict performance in another. The exception is mixed multilevel model methods, such as the type used in the Tennessee Value-Added Assessment System (TVAAS).[1] These models are frequently bypassed due to computational burdens and a lack of transparency. Additionally, existing studies offer no formal evaluation about the extent to which information on teacher performance across subjects improves the predictive power of value-added estimates. Also, the literature offers no guidance on the circumstances under which using information across subjects would be most useful. This paper develops an intuitive and computationally simple method for using information across teacher performance domains that addresses these gaps in the literature.

More specifically, we provide a statistical model of student learning and consider its implications for predicting teacher performance. In particular, we suggest a simple way of optimizing the predictive efficiency of traditionally derived teacher value-added metrics by optimally weighting subject-level measures of teacher effectiveness. We derive two methods to implement this weighting and test them using longitudinal data from North Carolina. Our results indicate that an optimal weighting of teacher performance in reading and math improves our ability to explain future teacher value-added by up to a third.

Because teacher ability to raise student test scores is positively correlated across subjects, our findings suggest that incorporating information from math value-added generates a more precise prediction of the ability to teach reading. Similarly, optimal use of the information in teacher reading and math value-added increases the precision of teacher value-added across a composite average of the subjects. This information is particularly valuable when contract rules require administrators to make permanent employment decisions based on only a few years of data and, hence, particularly imprecise estimates of teacher value-added. Also, because these improvements come from the use of test information already collected by school districts, this methodology represents a low-cost way to increase the relatively low reliability of teacher value-added measures (Koedel & Betts, 2007). Finally, our methodology can be further generalized to allow for the optimal weighting of teacher value-added scores across the particular competencies that make up a subject test score.

Model

Suppose that student learning in math can be represented as follows:

$$\Delta A^{M}_{i,j,t} = \alpha^{M} + \theta^{M}_{j} + \eta^{M}_{j,t} + \varepsilon^{M}_{i,j,t}, \quad (1)$$

where $\Delta A^{M}_{i,j,t}$ is the math gain of student i with teacher j in period t; α is the mean achievement gain; θ^{M}_{j} is teacher j's value-added in mathematics; $\eta^{M}_{j,t}$ is the classroom specific shock of teacher j in period t; $\varepsilon^{M}_{i,j,t}$ is the idiosyncratic innovation to student i's learning. Although our actual estimation will use a richer value-added specification that includes student-level covariates, we begin with this more parsimonious conceptual model in the interest of expositional clarity. Reading achievement evolves according to an analogous process,

$$\Delta A^{R}_{i,j,t} = \alpha^{R} + \theta^{R}_{j} + \eta^{R}_{j,t} + \varepsilon^{R}_{i,j,t}. \quad (2)$$

Suppose also that θ^{M}_{j}, $\eta^{M}_{j,t}$, $\varepsilon^{M}_{i,j,t}$, θ^{R}_{j}, $\eta^{R}_{j,t}$, and $\varepsilon^{R}_{i,j,t}$ are not observed but are known to be mean zero and normally distributed.[2] These component contributions to student learning are independent of each other both within and across periods with the following exceptions: $cov(\theta^{M}, \theta^{R}) \neq 0$, $cov(\eta^{M}, \eta^{R}) \neq 0$, and $cov(\varepsilon^{M}, \varepsilon^{R}) \neq 0$. That is, contemporaneous teacher, classroom, and student shocks may be correlated across reading and math achievement. We also assume that the expected math and reading gains, α^{M} and α^{R}, are known. In this framework, an unbiased estimate of a teacher's math value-added using data from a single period can be written as follows:

$$\begin{aligned}\hat{\theta}^{M}_{j,t} &= \frac{1}{N}\sum\nolimits_{i=1}^{N}(\Delta A^{M}_{i,j,t} - \alpha^{M}) \\ &= \theta^{M}_{j} + \eta^{M}_{j,t} + \frac{1}{N}\sum\nolimits_{i=1}^{N}\varepsilon^{M}_{i,j,t}. \end{aligned} \quad (3)$$

Rather than indexing N across classrooms, we assume that class sizes are homogenous across teachers and time periods and simplify our expression for single period math value-added to the following:

$$\hat{\theta}^{M}_{j,t} = \frac{1}{N}\sum_{i=1}^{N}(\Delta A^{M}_{i,j,t} - \alpha^{M}) = \theta^{M}_{j} + \nu^{M}_{j,t}, \quad (4)$$

where $\nu^{M}_{j,t} = \eta^{M}_{j,t} + \frac{1}{N}\sum_{i=1}^{N}\varepsilon^{M}_{i,j,t}$. We can write an analogous expression for a one-period estimate of teacher reading value-added. In this case, $var(\nu^{M}) = var(\eta^{M}) + \frac{1}{N}var(\varepsilon^{M})$ and $cov(\nu^{M}, \nu^{R}) = cov(\eta^{M}, \eta^{R}) + \frac{1}{N}cov(\varepsilon^{M}, \varepsilon^{R})$.

Incorporating data across multiple periods, an estimate of teacher math value-added can be written as

$$\hat{\theta}^{M}_{j} = \theta^{M}_{j} + \frac{1}{T}\sum_{t=1}^{T}\nu^{M}_{j,t}. \quad (5)$$

Without covariates, this corresponds to the estimate that would be produced by a fixed effects model of teacher quality. It is unbiased, but it is not efficient. In particular, because the distribution of true teacher value-added and $\nu^{M}_{j,t}$ is unknown, an empirical Bayes's technique that calculates the expectation of value-added given the observed noisy measure will produce a more efficient estimator. This shrinkage estimator is given by

$$\bar{\theta}^{M}_{j} = \hat{\theta}^{M}_{j}\frac{var(\theta^{M})}{var(\theta^{M}) + \frac{1}{T}var(\nu^{M})}. \quad (6)$$

For teachers with fewer student observations, the estimate of value-added is shrunk toward the mean (zero in this case) of the distribution. This occurs because the estimate is relatively noisy. For more experienced teachers, the degree of shrinkage is less because the empirical estimate is less noisy. This type of shrinkage estimator is very similar to what is done by hierarchical linear model (HLM) estimates of teacher effectiveness. Abstracting from the lack of covariates, this is the method used by Kane and Staiger (2008) to estimate teacher quality.

Whereas this type of estimation approach is common in value-added modeling, it does not efficiently use information across academic subjects. To the extent that teacher effectiveness is correlated across subjects, the best guess of a teacher's ability to increase mathematics test scores should also incorporate information regarding a teacher's ability to increase reading scores. Thus, we wish to generalize our estimator to calculate the expected value of true math value-added conditioned on both observed math value-added and observed reading value-added. Given normality, this expectation of math value-added given $\hat{\theta}^{M}_{j}$ and $\hat{\theta}^{R}_{j}$ is given by the following:

$$\theta^{M}_{j} = \hat{\theta}^{M}_{j}\frac{\left[var(\theta^{R}) + \frac{var(\nu^{R})}{T}\right]var(\theta^{M}) - \left[cov(\theta^{M},\theta^{R}) + \frac{cov(\nu^{M},\nu^{R})}{T}\right]cov(\theta^{M},\theta^{R})}{\left[var(\theta^{R}) + \frac{var(\nu^{R})}{T}\right]var(\theta^{M})\left[var(\theta^{M}) + \frac{var(\nu^{M})}{T}\right] - \left[cov(\theta^{M},\theta^{R}) + \frac{cov(\nu^{M},\nu^{R})}{T}\right]^{2}} + \hat{\theta}^{R}_{j}\frac{\left[var(\theta^{M}) + \frac{var(\nu^{M})}{T}\right]cov(\theta^{M},\theta^{R}) - \left[cov(\theta^{M},\theta^{R}) + \frac{cov(\nu^{M},\nu^{R})}{T}\right]var(\theta^{M})}{\left[var(\theta^{R}) + \frac{var(\nu^{R})}{T}\right]\left[var(\theta^{M}) + \frac{var(\nu^{M})}{T}\right] - \left[cov(\theta^{M},\theta^{R}) + \frac{cov(\nu^{M},\nu^{R})}{T}\right]^{2}} \quad (7)$$

A derivation of this result is given in the appendix (see the online appendix, available at http://eepa.sagepub.com/supplemental). Note that as the number of periods, T, approaches infinity, the weight placed on reading goes to zero. For any finite number of periods, however, an estimate of a teacher's reading value-added is informative regarding her effectiveness in math instruction. The corresponding expectation for reading value-added is symmetric. This highlights the conclusion that the use of multiple subject measures is likely to be more important when one is forced to evaluate teachers over a short time horizon.

Another possible approach to weighting multiple subjects is to use weights derived from ordinary least squares (OLS) regression coefficients. For example, we could run a regression of the form

$$\hat{\theta}^{M}_{jt} = \pi^{T}_{0} + \pi^{T}_{1}\hat{\theta}^{M}_{j,T,-t} + \pi^{T}_{2}\hat{\theta}^{R}_{j,T,-t} + \mu^{M}_{j,T,t}, \quad (8)$$

where the regression coefficients also represent prediction weights. This corresponds to a regression of the raw math value-added computed using data from period t on the raw reading and math value-added measures constructed from T prior periods, not including the reference period, $-t$, and is similar to the empirical Bayes's method outlined by Morris (1983). If our assumptions regarding the data generating process are correct, the OLS weights converge to those described in Equation 7. The OLS approach is also flexible. For example it is easily generalizable to using

TABLE 1
Summary Statistics for North Carolina Fourth-Grade Sample Students

Variable	North Carolina
Normalized reading score	0.000 (1.000)
Normalized math score	0.000 (1.000)
Student fraction male	0.507 (0.500)
Student fraction free lunch	0.460 (0.498)
Student fraction White	0.615 (0.487)
Student fraction Black	0.294 (0.455)
Student fraction Hispanic	0.045 (0.208)
Student fraction special education	0.128 (0.334)
Student fraction limited English	0.029 (0.168)
Student age	10.290 (0.499)
Parent education < high school	0.117 (0.322)
Parent education—high school	0.196 (0.397)
Parent education—college graduate	0.203 (0.402)
Parent education—graduate work	0.046 (0.210)

Note. Standard deviations are in parentheses. Test scores are normalized to be mean zero with unit standard deviation by state, year, grade, and subject. $n = 710{,}453$.

value-added measures across a larger number of subjects or assessment areas (e.g., science and social studies or math subitem scores). Additionally, whereas the former weighting scheme treats all prior years as equally important in predicting future teacher value-added, we can relax this assumption and allow value-added from recent years to have different weights than value-added from more distant periods. These approaches can be implemented by simply adding more regressors to the equation.

To this point our weighting procedures have abstracted from the possibility that teachers face different class sizes. A substantial variation in actual class size could represent an important source of heterogeneity in the precision of teacher value-added estimates. Modifying the weights found in Equation 7 to allow for class size heterogeneity is straightforward, and the resulting formula can be found in the appendix. In our results section we show that this modification does not improve the predictive power of our weights, mostly because there is relatively low variance for class size in our data (the standard deviation is about 3.6 students). The regression estimates of Equation 8 could also be modified, in principle, to account for differences in class size through the use of separate sets of terms for teachers of a particular class size, though this might quickly grow unwieldy in practice.

Empirical Implementation

To obtain raw estimates of teacher value-added, we estimate the following OLS regression separately for reading and math:

$$\Delta A^{M}_{i,j,t} = X_{i,j,s,t} B^{M}_{t} + e^{M}_{i,j,s,t} \quad . \tag{9}$$

Note that $\Delta A^{M}_{i,j,t}$ represents the change in math performance of student i, with teacher j, in school s, from period $t-1$ to period t. $X_{i,j,s,t}$ is a vector of student-level covariates shown in Table 1, and $e^{M}_{i,j,s,t}$ is the residual. The t subscript on the covariate coefficients indicates that we run a separate regression for each year. Although our initial specification lacks school fixed effects, we will also present results with them included. We calculate the raw teacher fixed effect for a given year by averaging the residuals for that teacher within that year. This corresponds to a random effects assumption about teacher quality, namely teacher quality is uncorrelated to the student-level controls.[3] For specifications in which we use multiple years of data to calculate the teacher value-added, we take a simple average of the yearly value-added measure across the years in our sample.

To calculate the appropriate weights for the procedures described in Equations 6 and 7, we require estimates of the relevant variances and

covariances. Our statistical model establishes the following relationships:

$$cov(\hat{\theta}^{M}_{j,t}, \hat{\theta}^{M}_{j,t-1}) = var(\theta^{M}), \quad (10)$$

$$var(\hat{\theta}^{M}_{j,t}) = var(\theta^{M}) + var(\nu^{M}), \quad (11)$$

$$cov(\hat{\theta}^{R}_{j,t}, \hat{\theta}^{R}_{j,t-1}) = var(\theta^{R}), \quad (12)$$

$$var(\hat{\theta}^{R}_{j,t}) = var(\theta^{R}) + var(\nu^{R}), \quad (13)$$

$$cov(\hat{\theta}^{R}_{j,t}, \hat{\theta}^{M}_{j,t-1}) = cov(\theta^{R}, \theta^{M}), \quad (14)$$

$$cov(\hat{\theta}^{R}_{j,t}, \hat{\theta}^{M}_{j,t}) = cov(\theta^{R}, \theta^{M}) + cov(\nu^{R}, \nu^{M}). \quad (15)$$

Note that all of the variances and covariances on the left hand side of these six equations can be directly estimated using our raw single-year measures of teacher value-added. Collectively, these equations identify the variances and covariances on the right hand side of the equations, which are necessary for the construction of our weights.

Data

The predictions of the above model are largely intuitive. Namely, the positive correlation in teacher ability across subjects suggests using multiple subject scores in combination can provide better information in judging teachers. However, it is important to determine how large the efficiency gain from multiple-subject evaluations may be in practice. To aid in this, we use a data set derived from North Carolina school administrative records maintained by the North Carolina Education Research Data Center. The primary data consists of student-year observations for all students in fourth grade in the state from 1998 to 2004. During this time period, North Carolina required end-of-course standardized exams for all these students in both reading and mathematics that were aligned to the state's learning standards.

We follow the data selection and standardization procedures of earlier researchers (Jacob, Lefgren, & Sims, 2010; Clotfelter, Ladd, & Vigdor, 2007). Namely, we transform the North Carolina test scores to reflect standard-deviation units relative to the state average for fourth graders in that year and then follow the algorithm described in detail in Clotfelter et al. (2007) to match teachers to students.[4] This allows an approximately 79% match success rate. Our selection of fourth graders for study is largely because of the comparatively high match rates for that grade. We have also run the analysis on the sample of fifth graders and found similar prediction improvements to those described below. In addition to the end-of-course examinations for the fourth graders, we have their third-grade test score data, which we use to calculate achievement gains for our model.

Table 1 reports summary statistics including the basic demographic controls for student race, ethnicity, free lunch, and special education status available in the data. Whereas the North Carolina sample is close to the national average in free lunch eligibility (46% compared to 42% nationally), it actually has smaller than average minority enrollments, comprised mainly of African American students and only a small percentage of non-native English speakers.

Results and Discussion

Our model suggests that, especially when limited years of teacher data are available, an optimal weighting of teacher value-added scores from reading and math can make better predictions about future teacher value-added than a single-subject model can. To test this we use multiple derivations of past teacher value-added from the years from 1998 to 2003 to predict the value-added of teachers in the 2003–2004 school year. We first compare the predictive power of the basic empirical Bayes's value-added measures (from Equation 6) based on a single subject, with those of the two-subject model of Equation 7. To do so we regress an unshrunken teacher value-added measure from a future year (the 2003–2004 school year) on the weighted value-added measures based on prior years.

Table 2 presents the results of this analysis, including the prediction coefficient (the regression beta) as well as the R^2 goodness-of-fit statistic. Furthermore, because of the potential for measurement error in value-added calculations, the maximum R^2 value for these exercises is unlikely to be 1. To provide a more intuitive measure of how

TABLE 2

Shrinkage Weights and Prediction Fit Measures for Alternative Value-Added Weighting Models

		Simple Bayes's Weighting (Equation 6)				Complex Bayes's Weighting (Equation 7)				
Years of Prior Data	Prediction Sample Size	Own-Subject Weight	Prediction Coefficient	Prediction R^2	R^2/Max R^2	Reading Weight	Math Weight	Prediction Coefficient	Prediction R^2	R^2/Max R^2
(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)
Predicting future reading value-added										
1	2,252	0.286	0.926	0.083	0.291	0.179	0.149	0.975	0.111*	0.388
2	1,603	0.445	0.917	0.120	0.419	0.299	0.166	0.960	0.151*	0.528
3	1,202	0.546	0.886	0.132	0.461	0.388	0.161	0.881	0.144	0.503
4	863	0.616	0.836*	0.125	0.437	0.457	0.152	0.818*	0.133	0.464
5	640	0.667	0.723*	0.118	0.412	0.512	0.142	0.804*	0.131	0.458
Predicting future math value-added										
1	2,252	0.537	1.044	0.322	0.599	0.028	0.526	1.047	0.324*	0.603
2	1,603	0.699	1.014	0.405	0.754	0.031	0.687	1.015	0.406	0.756
3	1,202	0.777	0.947	0.400	0.745	0.030	0.765	0.948	0.400	0.745
4	863	0.823	0.949	0.419	0.780	0.029	0.811	0.950	0.420	0.782
5	640	0.853	0.959	0.433	0.806	0.027	0.842	0.959	0.433	0.806
Predicting future composite value-added										
1	2,252	0.494	1.008	0.269	0.544	0.103	0.338	1.033	0.285*	0.577
2	1,603	0.661	0.995	0.352	0.712	0.165	0.426	1.010	0.367*	0.743
3	1,202	0.746	0.923*	0.344	0.696	0.209	0.463	0.924*	0.351	0.710
4	863	0.796	0.912*	0.352	0.712	0.242	0.482	0.913*	0.358	0.724
5	640	0.830	0.913	0.362	0.733	0.269	0.492	0.919	0.368	0.745

Note. Each row of the table represents a separate analysis using the indicated number of years of student data prior to 2003–2004 to calculate teacher value-added for teachers present in all included sample years. The weighted value-added data is then used to make predictions of teacher value-added in the 2003–2004 school year. Goodness of fit statistics are from these predictions. R^2/Max R^2 is the prediction R squared divided by the fraction of the variance in the 1-year teacher value-added that is attributable to teacher ability. Each panel predicts the outcomes of a different value-added subject measure. In panel C, the composite subject is a simple average of reading and math value-added scores. Reported sample size is that of teachers present in all required data years for whom predictions are made.

* in Columns 4 and 9 = the coefficient is significantly different from 1 at the 5% level. * in Column 10 indicates that the difference in R squared between Columns 5 and 10 is significantly different from 0 at the 5% level.

successful past value-added is at predicting the future, we also present a measure of the fraction of the variance attributable to teacher ability, rather than the total variance, that can be explained by the value-added measure. This measure, implemented by dividing the actual R^2 by the reliability of the 1-year teacher value-added estimate, can be found in Columns 6 and 11.

Each row of the table represents a separate analysis using the indicated number of years of student data prior to 2003–2004 to calculate teacher value-added for the right hand side of the regression. The prediction sample size refers to the number of teachers present in all relevant sample years. Because we require a teacher to be present in all of the years over which we calculate the predictive value-added, the sample size of teachers for which we are making predictions declines as the number of years of data increase. This not only is relevant to the paper but serves to highlight a real difficulty in teacher evaluation via value-added; namely, we have short information baselines for most teachers we might wish to evaluate. Column 3 shows the weight on past single subject value-added produced by a standard empirical Bayes's shrinkage estimator, where 1 would indicate full weight should be given to the raw value-added estimate. Similarly, Columns 7 and 8 give the weights accorded to raw math and reading value-added in the multisubject shrinkage procedure.

Because value-added estimation may be used to fulfill different objectives, each panel of the table predicts the outcomes of a different value-added subject measure. In Panel A, we use the past value-added measures to predict future teacher value-added in reading. As might be expected from the previous literature's focus on math scores, the traditional Bayes's estimator using past reading value-added is a poor predictor, explaining only 8% to 13% of the variation in future reading value-added. We also note that the R^2 measure is not monotonically increasing in the number of data years. The primary reason for the nonmonotonicity is that data from earlier years are less informative than are data from more recent years. Consequently, when we allow the weights to vary with time from the current period, as in Table 3, Column 3, this nonmonotonicity is largely eliminated, although there is still some due to the change in the composition of teachers as more years of data are added.

The results here also suggest a high degree of noise in the reading value-added estimates that is especially noteworthy when only 1 or 2 years of data are available. Even with 5 years of data, the uncertainty in value-added estimation leads the empirical Bayes's procedure to downweight the raw estimate by almost half. Finally, we reject the null hypothesis of no prediction bias (prediction coefficient equal to one) when more than 3 years of data are used.

We next compare these results with the predictive performance of the multisubject Bayes's weighting given in Equation 7. This differs from the earlier formulation in that it optimally incorporates information about a teacher's value-added in math to predict future value-added in reading. The weighting results in Columns 7 and 8 suggest that when only 1 year of data is available for teachers, their past math and reading performances convey roughly equivalent information about their future ability to teach reading. The weight placed on math remains at a roughly constant level as more years of data are added, thus diminishing in relative importance. However, comparing the prediction fit measures for traditional Bayes's weighting in Columns 5 and 6 with those for the multisubject weighting in Columns 10 and 11, we find that the use of multisubject weights improves the prediction in all cases. When only 1 or 2 years of data are available, it leads to a statistically significant improvement in prediction fit that exceeds 25%. Put another way, given 1 year of data on value-added, using optimal information on two subjects increases the ability to predict reading value-added by almost as much as gathering a 2nd year of data. Indeed, with 2 years of data, the use of multisubject weights improves prediction fit more than a 3rd year of data. Meanwhile, a comparison of prediction coefficients shows the same pattern, noted in the simple Bayes's results, of rejecting the null hypothesis of one when more than 3 years of teacher data are used, suggesting roughly similar performance in terms of prediction bias.

Panel B of the table considers the use of past value-added to predict future math performance and follows the same format as the first panel. Here, the results show that math is a much less noisy predictor than reading is. When math value-added alone is used as a predictor, the Bayes's weights are much larger and the resulting R^2 measures are three to four times as large as for reading,

TABLE 3
Prediction Fit Measures for Alternative Specifications

Years of Prior Data	OLS Complex Bayes's Equation 8	OLS Complex Bayes's: Time Varying Weights	Complex Bayes's: Weights Based on Number of Students	Simple Bayes's: Lagged Achievement Model	Complex Bayes's: Lagged Achievement Model	Simple Bayes's: School Fixed Effects	Complex Bayes's: School Fixed Effects
(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Predicting future reading value-added							
1	0.111	0.111	0.109	0.085	0.112	0.051	0.069
2	0.151	0.149	0.151	0.132	0.158	0.070	0.093
3	0.144	0.142	0.141	0.146	0.158	0.081	0.082
4	0.132	0.142	0.131	0.136	0.149	0.070	0.067
5	0.130	0.146	0.132	0.133	0.144	0.049	0.054
Predicting future math value-added							
1	0.325	0.325	0.326	0.310	0.307	0.244	0.243
2	0.407	0.407	0.406	0.345	0.360	0.288	0.288
3	0.400	0.412	0.401	0.405	0.406	0.243	0.243
4	0.420	0.439	0.419	0.433	0.433	0.243	0.242
5	0.433	0.461	0.432	0.435	0.435	0.225	0.225
Predicting future composite value-added							
1	0.285	0.285	0.284	0.254	0.269	0.193	0.205
2	0.367	0.367	0.367	0.335	0.360	0.241	0.255
3	0.351	0.358	0.351	0.343	0.354	0.212	0.212
4	0.358	0.376	0.357	0.364	0.375	0.204	0.201
5	0.366	0.393	0.369	0.378	0.385	0.185	0.184

Note. OLS = ordinary least squares. Each row of the table represents a separate analysis with its own data restrictions. Each row uses the indicated number of years of student data prior to 2003–2004 to calculate teacher value-added for all teachers present in all those sample years. The weighted value-added data is then used to make predictions of teacher value-added in the 2003–2004 school year. Goodness of fit statistics are from these predictions. Each panel predicts the outcomes of a different value-added subject measure. In panel C, the own subject is a simple average of reading and math value-added scores.

which suggests that past math results alone can explain up to 43% of the variation in future math value-added. Furthermore, incorporating reading into the predictions in either manner fails to meaningfully improve them, as it is assigned a relatively low weight.

The final panel of the table considers the prediction of a composite value-added—in this case the teacher's future ability to increase the simple average of math and reading achievement. This might be thought to approximate the sort of comprehensive measure districts would wish to use in evaluating teachers for retention, tenure, or performance rewards. Part C of the appendix briefly shows how we modify our conditional expectation formulas to reflect this composite score.

When we compare the performance of a simple average composite score with an optimal weighting of both subject scores, the latter improves predictive performance, increasing the R^2 by about 2% to 7%. As expected, the effect is statistically significant and largest when only 1 or 2 years of data are available to make value-added judgments. Whereas the composite results suggest that this is a modest improvement, it is also inexpensive to implement in most settings because states are commonly required to record and track classroom test data in multiple subjects.

Table 3 presents a variety of specification checks that examine the robustness of our central results. In each case we report the R^2 prediction fit measures that correspond to those in Table 2, Columns 5 and 10. We begin by estimating the generalized OLS model of Equation 8. Interestingly, it produces goodness-of-fit results that are almost identical to those given by the analytical weighting scheme in Equation 7.

These OLS results confirm an important asymmetry between the high signal value-added math estimates and their noisier reading counterparts. This relatively strong math signal drives the composite results. If both math and reading value-added were equally informative about future teacher quality there would be no benefit to differentially weighting them. This insight, in turn, suggests that this OLS approach could potentially be generalized further to include as many subjects as desirable. Alternatively, value-added measures based on test subitem category scores that target certain areas of competency may be more predictive than whole subject value-added and should be entered separately into the regression model. Thus, the model allows the relaxation of the restriction that components of a test score be combined into the same value-added. The weights can also change depending on the number of periods used for evaluation. The gains from our simple two-subject application of this methodology are modest, but the methodology itself has the flexibility to produce larger gains through further expansion. Unfortunately, our current data do not provide the scope to examine subitem scores.

Continuing with our specification checks, we test whether the restriction that all past teacher years are equally informative is driving our results by modifying the OLS model to consider past teacher value-added separately by year in making a prediction. The results are presented in Column 3. Whereas the two models are identical for 1 year of predictive data, we expect the time-varying model to gain more predictive power than the baseline as the number of years increases. Indeed, we see such a divergence in all cases, with the time varying model producing a 6% to 12% higher R^2 value with 5 years of data.

Column 4 incorporates class size information into the multisubject weighting in the manner discussed in the appendix. The results suggest that class size does not constitute an important weighting consideration in the present data because the results differ only from the baseline results produced by Equation 7 in the third decimal place.

To this point we have been using time differenced test scores, commonly referred to as the gains method, as the basis of our value-added calculation. A common alternative uses a past test score as an added regressor, instead of differencing test scores, to produce value-added estimates. The next two columns of our table show the R^2 results for both traditional Bayes's weighting and our proposed multisubject weighting using this lagged achievement model. Although individual results change somewhat, the broad pattern remains the same. The multisubject weights produce almost identical results in predicting math and somewhat better predictions of reading and a composite. The improvements are much larger when only 1 or 2 years of data are used.

The final columns of Table 3 modify the previous table's estimates by including school fixed effects in the model. Unlike the previous specification checks, we do not expect this

analysis to produce the same estimates as the baseline because there is a clear conceptual difference between comparing teacher value-added within a school as opposed to within the state. However, both ideas are likely to be useful in different policy contexts. Once school fixed effects are incorporated, the predictive power of both estimators decreases. This is to be expected as a correlation of teacher or student quality at a school level results in a higher preponderance of noise within school. Again, the pattern emerges; multisubject weighting improves predictive power, particularly for reading scores when using 1 or 2 years of data.

Although we have shown that our weighting procedures improve the predictive power for reading and composite value-added at low cost, changes in R^2 measures do not provide much of an intuitive feel for their practical importance. In Table 4, we attempt to translate the consequences of using our estimators into a policy relevant form. In particular, one of the most commonly suggested uses of value-added is to create a ranking of teachers to be used for rewards or sanctions. In particular, recent studies (Goldhaber & Hansen, 2010; Gordon et al., 2006; Hanushek, 2009) suggest that the best way to use value-added scores to improve student outcomes would be by letting go of the teachers ranked lowest by value-added metrics. In Table 4, we show how the use of multisubject weighting of value-added might inform such efforts.

Table 4 contains two sets of numbers. The first set gives the probability that a teacher's unshrunken 2003–2004 value-added performance will lie in a particular portion of the teacher quality distribution given that past value-added measures placed them in that portion of the distribution. We perform the exercise for four hypothetical distributional questions: top 10%, top 20%, bottom 10%, and bottom 20%. In each case we compare the simple Bayes's shrinkage estimator of Equation 6 with the complex Bayes's estimator of Equation 7. As in previous tables, we repeat the exercise for each possible combination of 1 to 5 data years and for predicting value-added in reading, math, and a simple composite average.

The results suggest that the multisubject weights increase the probability of correctly predicting the worst teachers in reading (the bottom 20%) by 2 to 3 percentage points when limited data are available. When more than 3 years of data are available there are no measurable gains. As before any gains in math prediction power are small, while the complex Bayes's estimator increases the correct composite predictions of the lowest 20% performing teachers by about 2 percentage points.

This exercise gives us some policy intuition but is lacking in one important respect. Because future value-added is itself a noisy measure, it is not actually a sound basis for judging the validity of our predictions. Instead we would like to ask what is the probability a teacher's future performance will lie in a certain portion of the true teacher quality distribution given a past value-added prediction that they are in that portion of the distribution. A methodology to provide such a measure is developed in Jacob and Lefgren (2008). Their method, formalized in the article's appendix, involves using the moments of the data to simulate the error rates inherent in the noisy future value-added distribution and then uses Bayes's rule to update the conditional probability of a correct prediction. Of course, this measure is not perfect in the current context. In particular, in our longer data samples with the number of teachers already limited, the precision of calculations that involve placing teachers in numerically small bins is troublesome. Still, we believe that a broad pattern of results produced by this method can be informative here.

Using this methodology, the bracketed set of numbers in Table 4 compares the simple and complex Bayes's estimators on their ability to predict a teacher's place in the tails of the true teacher quality distribution. These results show that the use of multisubject weights is much more valuable in practical prediction of a teacher's true quality in reading, using 1 or 2 years of data, than the goodness-of-fit measures might lead us to believe. Indeed, the results show that when only 1 year of data is available, the complex Bayes's weights improve the correct rate of predictions of teachers in the bottom tails of the distribution by around 10 percentage points. They also suggest that in limited data situations multisubject weights can be used to improve the correct prediction of the worst teachers in composite terms by 3 to 4 percentage points. Because labor agreements often force districts to make long-term teacher retention

TABLE 4
The Probability a Teacher is in Extreme (Top or Bottom) Deciles of Teacher Value-Added Given that Estimated Value-Added is in Corresponding Extreme Deciles

	Top 10%		Top 20%		Bottom 10%		Bottom 20%	
Years of Prior Data	Simple Bayes's	Complex Bayes's	Simple Bayes's	Complex Bayes's	Simple Bayes's	Complex Bayes's	Simple Bayes's	Complex Bayes's
(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
Predicting future reading value-added								
1	0.204 [0.494]	0.218 [0.545]	0.327 [0.602]	0.344 [0.658]	0.212 [0.525]	0.248 [0.659]	0.330 [0.613]	0.361 [0.711]
2	0.244 [0.644]	0.263 [0.714]	0.359 [0.705]	0.378 [0.765]	0.217 [0.544]	0.217 [0.544]	0.336 [0.633]	0.358 [0.702]
3	0.250 [0.667]	0.275 [0.762]	0.371 [0.742]	0.400 [0.834]	0.231 [0.567]	0.231 [0.567]	0.332 [0.618]	0.332 [0.618]
4	0.314 [0.909]	0.291 [0.821]	0.395 [0.819]	0.378 [0.764]	0.172 [0.373]	0.195 [0.461]	0.323 [0.592]	0.312 [0.555]
5	0.344 [1.022]	0.297 [0.844]	0.406 [0.854]	0.414 [0.879]	0.203 [0.490]	0.219 [0.549]	0.336 [0.631]	0.320 [0.581]
Predicting future math value-added								
1	0.422 [0.819]	0.427 [0.829]	0.482 [0.775]	0.480 [0.770]	0.381 [0.726]	0.389 [0.746]	0.481 [0.772]	0.486 [0.781]
2	0.450 [0.881]	0.463 [0.909]	0.563 [0.938]	0.559 [0.932]	0.335 [0.625]	0.329 [0.611]	0.486 [0.782]	0.495 [0.801]
3	0.500 [0.993]	0.500 [0.993]	0.567 [0.946]	0.567 [0.946]	0.355 [0.670]	0.347 [0.651]	0.498 [0.807]	0.494 [0.798]
4	0.512 [1.019]	0.512 [1.019]	0.564 [0.941]	0.558 [0.929]	0.322 [0.595]	0.322 [0.595]	0.492 [0.793]	0.497 [0.805]
5	0.500 [0.993]	0.500 [0.993]	0.594 [1.002]	0.609 [1.033]	0.328 [0.609]	0.344 [0.644]	0.484 [0.779]	0.484 [0.779]
Predicting future composite value-added								
1	0.329 [0.654]	0.342 [0.687]	0.46 [0.499]	0.46 [0.499]	0.354 [0.715]	0.376 [0.769]	0.475 [0.800]	0.497 [0.849]
2	0.444 [0.933]	0.441 [0.902]	0.541 [0.945]	0.547 [0.958]	0.366 [0.746]	0.360 [0.730]	0.477 [0.805]	0.498 [0.852]
3	0.442 [0.928]	0.458 [0.968]	0.521 [0.901]	0.525 [0.910]	0.331 [0.659]	0.331 [0.659]	0.456 [0.761]	0.473 [0.797]
4	0.465 [0.985]	0.465 [0.985]	0.535 [0.932]	0.529 [0.919]	0.321 [0.637]	0.321 [0.637]	0.468 [0.786]	0.468 [0.786]
5	0.453 [0.955]	0.438 [0.918]	0.547 [0.958]	0.547 [0.958]	0.359 [0.728]	0.344 [0.691]	0.453 [0.753]	0.477 [0.805]

Note. The table reports the probability that a teacher is in a particular part of the empirical 2003–2004 value-added distribution as a function of whether he or she was in the same range in the estimated simple or complex Bayes's value-added distribution in prior years. In the square brackets, we use the method in Jacob and Lefgren (2008) to estimate the probability that a teacher is in a particular quantile of the true teacher value-added distribution as a function of whether he or she was in the same quantile of the prior years' observed value-added distribution.

decisions on the basis of 1 or 2 years of data, the use of multisubject weights appears to offer an important low cost area for improving the quality of these essential staffing decisions.

Conclusion

The increased use of value-added measures to inform district policies about teacher retention, promotion, and rewards seems very likely in the near future. Thus, low-cost, easily implementable methods of improving the predictive power of value-added measures are important. Our model of teacher performance suggests two easily implementable estimators that produce optimal multisubject weighted value-added measures. Our results indicate these estimators produce large improvements in the ability of value-added to distinguish teacher quality in reading and smaller improvements in the ability to predict composite teacher quality. These improvements are largest when value-added must be calculated across only one or two years of data. They are also most directly applicable to elementary schools, where teachers are responsible for instruction across a variety of subjects.

Furthermore, the models provided in this paper can be generalized to allow the optimal weighting of additional subjects or even subsections of particular subjects. While data and space considerations place further investigation of these possibilities beyond the scope of the current paper, an important part of future research may well be determining if there are particular bellwether student competencies, whose measurement best predicts teacher ability. While there are undoubtedly numerous improvements that can be made to provide better data for teacher evaluation, our results suggest that better use of existing data would be a simple, low cost place to start.

Acknowledgments

We thank Jacob Vigdor and Brian Jacob for providing us with the data used in Clotfelter, Ladd, and Vigdor (2007) for our analysis and Mark Showalter and Eric Eide as well as anonymous reviewers for helpful comments and suggestions.

Declaration of Conflicting Interests

The authors declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The authors received no financial support for the research, authorship, and/or publication of this article.

Notes

1. Ballou, Sanders, and Wright (2004) provide a useful description of mixed multilevel models in a value-added context. The intuition that performance in one domain is informative regarding performance in another has been used in other contexts including the measure of school effectiveness (see Kane and Staiger, 2001) and hospital quality (see Staiger, Dimick, Baser, Fan, and Birkmeyer, 2009).

2. With arbitrary α^M or α^R, this assumption involves no loss of generality.

3. To check the importance of our random effects assumption, we have also estimated a model in which the teacher value-added measures enter as fixed effects. The correlation between the value-added estimates across the fixed and random effects strategies is 0.99.

4. The teachers identified in the student test file are those who proctored the exam, not necessarily those who taught the class. The authors describe the three-tiered system of matching students to actual teachers. The first assigns the proctor as the teacher if the proctor taught the correct grade and subject that year. They also look at the composition of the test-taking students and compare it with the composition of students in classes from the teacher file to find matches.

References

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

Andrabi, T., Das, J., Khwaja, A. I., & Zajonc, T. (2011). Do value-added estimates add value? Accounting for learning dynamics. *American Economic Journal: Applied Economics*, *3*(3), 29–54.

Ballou, D. (1996). Do public schools hire the best applicants? *Quarterly Journal of Economics*, *111*, 97–133.

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.

Clotfelter, C. T., Ladd, H. F., & Vigdor, J. L. (2007). *How and why do teacher credentials matter for student achievement?* [Working Paper #12828]. Cambridge, MA: National Bureau of Economic Research.

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

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

Hanushek, E. A. (2009). Teacher deselection. In D. Goldhaber & J. Hannaway (Eds.), *Creating a new teaching profession* (pp. 165–180). Washington, DC: Urban Institute Press.

Harris, D. N. (2009). Would accountability based on teacher value-added be smart policy? An examination of the statistical properties and policy alternatives. *Education Finance and Policy*, *4*, 319–350.

Jacob, B. A., & Lefgren, L. (2008). Can principals identify effective teachers? Evidence on subjective performance evaluation in education. *Journal of Labor Economics 25*(1), 101-136.

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

Kane, T. J., & Staiger, D. O. (2001). *Improving school accountability measures.* Working Paper #8156. Cambridge, MA: National Bureau of Economic Research.

Kane, T. J., & Staiger, D. O. (2008). *Estimating teacher impacts on student achievement: An experimental evaluation* [Working Paper #14607]. Cambridge, MA: National Bureau of Economic Research.

Koedel, C., & Betts, J. R. (2007). *Re-examining the role of teacher quality in the educational production function* [Working Paper #2007–03]. Nashville, TN: National Center on Performance Initiatives.

Morris, C. N. (1983). Parametric empirical Bayes' inference: Theory and applications. *Journal of the American Statistical Association*, *78*, 47–55.

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. *American Economic Review*, *94*, 247–252.

Rockoff, J. E., Jacob, B. A., Kane, T. J., & Staiger, D. O. (2011). Can you recognize an effective teacher when you recruit one? *Education Finance and Policy*, *6*, 43–47.

Rockoff, J. E., & Speroni, C. (2010). Subjective and objective evaluations of teacher effectiveness. *American Economic Review*, *100*, 261–266.

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

Staiger, D. O., Dimick, J. B., Baser, O., Fan, Z., & Birkmeyer, J. D. (2009). Empirically derived composite measures of surgical performance. *Medical Care 47*(2), 226–233.

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

Todd, P. E., & Wolpin, K. I. (2003). On the specification and estimation of the production function for cognitive achievement. *Economic Journal*, *113*, 3–33.

Todd, P. E., & Wolpin, K. I. (2006). *The production of cognitive achievement in children: Home, school and racial test score gaps.* Philadelphia, PA: University of Pennsylvania.

Authors

LARS LEFGREN is an associate professor of economics at Brigham Young University. His research examines questions relating to education, crime, and economic inequality.

DAVID SIMS is an associate professor of economics at Brigham Young University. His research interests include measuring the efficacy and unintended consequences of educational reforms.

Manuscript received September 9, 2010
Revision received July 18, 2011
Accepted August 1, 2011