Controlling for Student Background in Value-Added Assessment of Teachers

Dale Ballou, William Sanders and Paul Wright

JOURNAL OF EDUCATIONAL AND BEHAVIORAL STATISTICS 2004 29: 37
DOI: 10.3102/10769986029001037

Journal of Educational and Behavioral Statistics
Spring 2004, Vol. 29, No. 1, pp. 37–65

Controlling for Student Background in Value-Added Assessment of Teachers

Dale Ballou
Vanderbilt University

William Sanders
Paul Wright
SAS Institute

The Tennessee Value-Added Assessment System measures teacher effectiveness on the basis of student gains, implicitly controlling for socioeconomic status and other background factors that influence initial levels of achievement. The absence of explicit controls for student background has been criticized on the grounds that these factors influence gains as well. In this research we modify the TVAAS by introducing commonly used controls for student SES and demographics. The introduction of controls at the student level has a negligible impact on estimated teacher effects in the TVAAS, though not in a simple fixed effects estimator with which the TVAAS is compared. The explanation lies in the TVAAS's exploitation of the covariance of tests in different subjects and grades, whereby a student's history of test performance substitutes for omitted background variables.

Keywords: *accountability, teacher quality, value-added assessment*

1. Introduction

Throughout the United States, state and school districts are raising academic standards for students. At the same time, there are growing efforts to hold teachers and administrators accountable for the quality of the education they provide.

Although there is evidence that virtually all students, even those from disadvantaged backgrounds, can succeed in the right educational environment, it is much more difficult to raise the achievement of disadvantaged children to the new standards. Holding teachers and administrators accountable for student outcomes without regard for differences in students' backgrounds is manifestly unfair and, in the long run, counter-productive. Such policies will alienate educators, making it more difficult to staff schools serving the neediest population. The perception that educators are being held accountable for student achievement without due regard for factors beyond their control may ultimately discredit the standards movement itself.

The Tennessee Value-Added Assessment System

An alternative approach measures the efficacy of schools and teachers in value-added terms, based on student *progress* rather than the percentage of students able to meet an absolute standard. Measuring student progress requires controlling in some fashion for initial level of achievement. This is done most transparently if the pre- and post-tests are on the same achievement scale ("vertically equated"), in which case the analysis can be based on simple differences or gain scores. Alternatively, introducing a prior test score as a regressor controls for initial achievement, so that the contribution of schools and teachers to student progress is based on residual differences in the post-test scores. Because the value-added method measures gain from a student's own starting point, it implicitly controls for socioeconomic status and other background factors to the extent that their influence on the post-test is already reflected in the pre-test score.

Among the most prominent examples of value-added assessment in education is the Tennessee Value-Added Assessment System, developed by William Sanders and associates at the University of Tennessee (Sanders, Saxton & Horn, 1997). The statistical model underlying the TVAAS is highly parsimonious, employing no data on students other than test scores and the identities of schools and teachers. Within this parsimonious framework, however, the history of a student's test scores is brought to bear on the estimation of the effectiveness of the teachers that student has had.

Although the value-added approach to teacher assessment has widespread appeal, the TVAAS model has come under criticism for failing to do enough to control for SES and demographic factors (Linn, 2001; Kupermintz, 2002). The main concern is that such factors influence not only the starting point but also the rate at which a student learns. The Value-Added Research Consortium at the University of Florida College of Medicine explored the issue. Researchers at the Consortium estimated a variety of value-added models (though not the TVAAS), with and without contextual variables such as student income and race. They found that these variables were almost always statistically significant and that estimates of teacher and school effects were sensitive to the presence of such controls (University of Florida, 2000a; 2000b). Because the TVAAS was not among the models estimated, the implications of this research for the TVAAS are unclear. Nonetheless, the omission of such controls from the TVAAS has been controversial, with some members of the research and policy community skeptical that the contributions of schools and teachers can be measured accurately without controlling for contextual variables that also influence student achievement gains (Darling-Hammond, 1997; Popham, 1997).

The TVAAS can accommodate the introduction of covariates, suggesting that this controversy could be resolved by adding demographic and SES variables to the TVAAS model and using conventional statistical tests to ascertain whether they matter. This solution is not, however, as straightforward as it may first appear. Students are not assigned at random to teachers and schools. If better teachers are able to obtain jobs in schools serving an affluent student population, or if more affluent parents seek the best schools and teachers for their children (say, by residential location, or pressuring school administrators to place their children in desired classes), demographic and SES variables become proxies for teacher and school quality. Because they are correlated with otherwise unmeasured variation in school and teacher quality, the coefficients on these variables will capture part of what researchers are trying to measure with residuals. Predictors of school and teacher effectiveness will accordingly be biased toward zero.

Thus, when the data show that students from impoverished backgrounds do not gain as much from one year to the next as more affluent students, it is problematic whether to attribute that to the independent effect of their backgrounds on achievement or to the quality of their schooling. Simply introducing SES and demographic variables into a value-added model will not reveal which, if either, of these positions is correct, for both hypotheses imply an association between student background and achievement gain.

This issue has not always been appreciated in the literature. Aitkin and Longford (1986), in a discussion of various approaches to modeling school effects, point out that discrepant performance might be the consequence of SES, and could disappear on the inclusion of social class in the model. "Every effort needs to be made in school comparisons to avoid model mis-specification by the inclusion of all relevant variables in the initial model." (Aitkin & Longford, 1986, p. 22). But as we have just noted, if disadvantaged students are systematically assigned to less effective schools and teachers, inclusion of SES as a control can mask genuine differences in school and teacher quality.

One article that takes due notice of the problem is Raudenbush and Willms (1995). Although these authors are concerned with the estimation of school effects, not teacher effects, the statistical issues are the same. To control for context ("peer effects") they introduce the mean pre-test score among students at a given school (taken before enrollment at the school). The same model contains a school effect as a variance component (random effect). Individual student background is represented by the student's own prior test score. "Effects of school practice are conceived to be the unobservable part of the school effect that remains after removing the contributions of student background and context." As the authors expressly note, if the mean score on the prior test is not orthogonal to the random school effect (that is, if assignment of students to schools is not random), the coefficient on the context variable will pick up the correlation between the mean score and the true school effect. The model then underestimates the true variance of the school effects.

Fortunately it is possible to introduce these covariates in a way that avoids this difficulty. Although details appear below, we sketch our approach here. We first estimate a model predicting test score gains as a function of both student- and school-level SES and demographic variables and teacher fixed effects, using standard analysis of covariance methods. As a result, the coefficients on the SES and demographic covariates are independent of the teacher effects. We use these coefficients to compute "quasi-residuals" as $Y - \mathbf{X}b_{\text{cov}}$, where $\mathbf{X}$ is the vector of SES and demographic variables and b_{cov} the coefficient estimates. Because teacher fixed effects are not removed from the dependent variable, the contribution of teacher quality to achievement gains remains in the quasi-residual. Finally, we use these quasi-residuals as dependent variables in the TVAAS.[1]

2. An Overview of the TVAAS

Because the model and estimation methodology behind the TVAAS are not well understood, we begin with an overview.

Educational outcomes are represented by scores on the Tennessee Comprehensive Assessment Program (TCAP), a vertically linked series of achievement tests administered annually in one or more subjects. The sequence of test scores for a student who is first tested in 1994 in the third grade is assumed to satisfy the following equations:

$$Y_{94}^{3} = b_{94}^{3} + u_{94}^{3} + e_{94}^{3}, \tag{1a}$$

$$Y_{95}^{4} = b_{95}^{4} + u_{94}^{3} + u_{95}^{4} + e_{95}^{4}, \tag{1b}$$

$$Y_{96}^{5} = b_{96}^{5} + u_{94}^{3} + u_{95}^{4} + u_{96}^{5} + e_{96}^{5}, \tag{1c}$$

$$\text{etc} \ldots$$

where Y_t^k= test score in year *t,* grade *k,* b_t^k = district mean test score in year *t,* grade *k,* u_t^k= contribution of the grade *k* teacher to the year *t* test score, and e_t^k = student-level stochastic component in year *t,* grade *k.* Subscripts for students, teachers, and subjects have been suppressed in the interest of notational clarity. All variables in Equations 1a, 1b and 1c pertain to the same student and subject. The teacher subscript will normally change as a student advances through the grades, though it is possible for the third grade teacher in 1994, say, to be the same person as the fourth grade teacher in 1995.

Teacher effects are subscripted with years: the model does not constrain teacher effectiveness to be constant over time. Thus, “teacher effects” are actually teacher-within-year effects. Note, too, that the same teacher effect enters more than one equation, as achievement in later years is “layered” on top of earlier achievement. As a consequence, it is possible to estimate a teacher effect even if all data for the teacher during the year in question are missing. For example, if Y_{95}^3 were missing, u_{95}^3 would still appear in the equation for Y_{96}^4 provided the latter is not also missing.[2]

Teacher effects are deviations from the district average. If 4th-grade teachers in a district are particularly strong as a group, this appears as an increase in the district mean score, b_{95}^4, rather than the individual u_{95}^4s. No restrictions are placed on the within-student covariance matrix of the e_t, which includes covariances within the same subject over time and covariances across subjects and years. Covariances across students are set to zero.

More insight into the layered model is obtained if the equations are differenced. Solving Equation 1a for u_{94}^3 and substituting into Equation 1b, after rearranging terms we find that

$$Y_{95}^{4} - Y_{94}^{3} = b_{95}^{4} - b_{94}^{3} + u_{95}^{4} + e_{95}^{4} - e_{94}^{3} \text{ or,} \tag{2a}$$

$$u_{95}^{4} = (Y_{95}^{4} - Y_{94}^{3}) - (b_{95}^{4} - b_{94}^{3}) - (e_{95}^{4} - e_{94}^{3}). \tag{2b}$$

By repeatedly substituting and solving, we obtain analogous expressions for teacher effects in other grades and years. The teacher effect is therefore what remains of the year-to-year gain after removing the district mean gain and the contribution of factors idiosyncratic to the student. If we think of $(Y_{95}^{4} - Y_{94}^{3}) - (b_{95}^{4} - b_{94}^{3})$ as the residual gain at the student level, estimation of teacher effects is a matter of determining how much of the residual gain to attribute to student-specific factors and how much to the influence of the teacher. Although TVAAS constructs other measures of teacher effectiveness on the basis of Equation 2a, in this article our focus will be on the estimation of teacher within year effects such as u_{95}^{4}.[3]

Mixed-Model Methodology

TVAAS is estimated using a mixed-model methodology (Searle, 1971), applying to models of the form

$$\mathbf{y} = \mathbf{Xb} + \mathbf{Zu} + \mathbf{e}, \tag{3}$$

where **b** is a vector of fixed effects, **u** is a vector of random (teacher) effects, **X** and **Z** are incidence matrices (indicating which students have been assigned to which teachers in which subjects in which years), and **e** is a vector of random error terms. The use of boldface is intended to distinguish these vectors (which contain data for many students over multiple years) from the variables specific to students and years in Equations 1 and 2. **X** may include covariates. However, in the TVAAS, **Xb** consists solely of the district average score on a given test.

The teacher effects and random student components are assumed to be independent, normally distributed variates with variance-covariance matrices

$$\text{var}(\mathbf{u}) = \mathbf{D},\ \text{var}(\mathbf{e}) = \mathbf{R},\ \text{var}(\mathbf{Zu} + \mathbf{e}) = \mathbf{V} = \mathbf{ZDZ}' + \mathbf{R}.$$

D is assumed to be diagonal. Covariances are zero, even within-teacher covariances across years and subjects. **R,** as noted above, is block-diagonal, with unrestricted within-student covariance and zeros in the off-diagonal blocks. To the extent that test scores in different subjects and different years are correlated, TVAAS exploits that correlation to improve the precision of the estimates of **b** and **u.**

Estimates of **b** and **u** are obtained as solutions to the system

$$\begin{bmatrix} \mathbf{X}'\mathbf{R}^{-1}\mathbf{X} & \mathbf{X}'\mathbf{R}^{-1}\mathbf{Z} \\ \mathbf{Z}'\mathbf{R}^{-1}\mathbf{X} & \mathbf{Z}'\mathbf{R}^{-1}\mathbf{Z} + \mathbf{D}^{-1} \end{bmatrix} \begin{bmatrix} \mathbf{b} \\ \mathbf{u} \end{bmatrix} = \begin{bmatrix} \mathbf{X}'\mathbf{R}^{-1}\mathbf{y} \\ \mathbf{Z}'\mathbf{R}^{-1}\mathbf{y} \end{bmatrix}$$

The solution for the teacher effects is

$$\mathbf{u}^* = \mathbf{DZ'V}^{-1}(\mathbf{y} - \mathbf{Xb}^*) = \mathrm{E}(\mathbf{u} \mid \mathbf{y}). \tag{4}$$

u* is known as a shrinkage estimator. If u* contained a single element, rather than a vector, the formula for a single teacher effect would be $u^* = \rho[\Sigma(y - Xb^*)]/N$, where summation is over the *N* students in a teacher's class and ρ is the ratio $\mathrm{var}(u)/[\mathrm{var}(u) + \mathrm{var}(e)/N]$, i.e., the reliability of $y - Xb^*$ as a measure of the teacher effect. The expression $\mathbf{DZ'V}^{-1}(\mathbf{y} - \mathbf{Xb}^*)$ is the matrix generalization of $\rho[\Sigma(y - Xb^*)]/N$. The greater are the diagonal terms of **R,** the noisier is $\mathbf{y} - \mathbf{Xb}^*$ as an estimate of **u.** Accordingly **u*** is shrunk toward the grand mean of **u,** which is **0,** as var(**e**) increases. The shrinkage is optimal in the sense that **u*** is the minimum mean-squared error estimate of **u** when **u** is normally distributed. More generally, **u*** minimizes any symmetric loss function.

The TVAAS estimation problem is illustrated in simplified form in Figure 1. The district average test score for a given cohort of students in year t is given by the point labeled A. Between year *t* and year *t* + 1 these students advance one grade and are tested again in the same subject. The mean score in year *t* + 1 is labeled B.

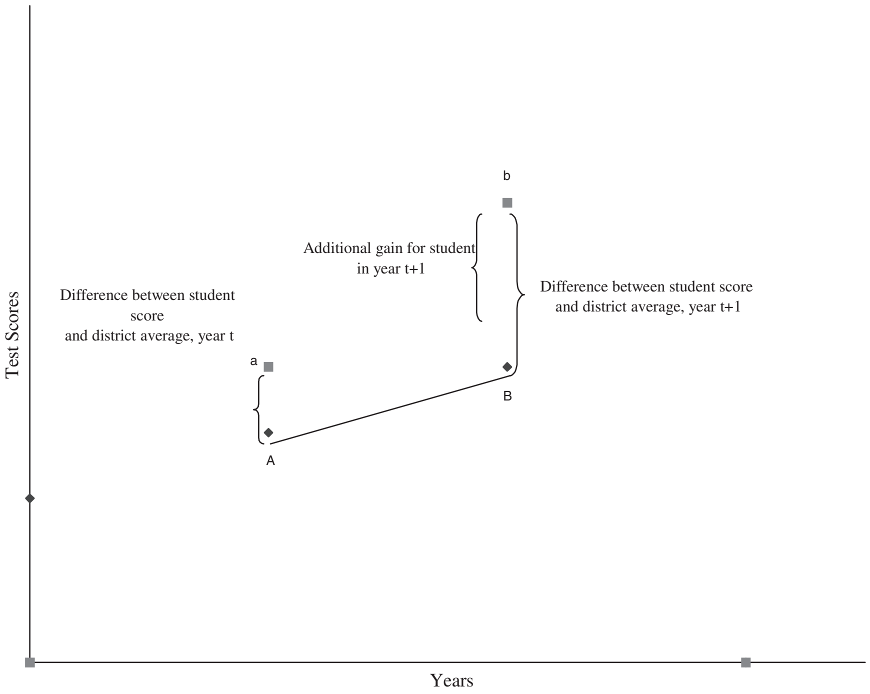
Test Scores
Years
b
Additional gain for student in year t+1
Difference between student score and district average, year t+1
Difference between student score and district average, year t
a
B
A

FIGURE 1. *TVAAS student and teacher effects.*

Scores are also depicted for a particular student ("Mitchell"), the points labeled with lower-case letters a and b. As shown, Mitchell exceeded the district mean in year t and in year $t + 1$. In fact, the gap between Mitchell and the district mean has grown: an extra learning gain occurred in year $t + 1$. The estimation problem is how much of this additional learning gain to attribute to the teacher and how much to idiosyncratic circumstances (Mitchell's genetic makeup, his home life, etc.).

If Mitchell's classmates also have gains that exceed the district average gain, his teacher in year $t + 1$ appears better than average. However, this is only a provisional conclusion: if it is normal for Mitchell and his classmates to outpace the rest of the district, the teacher in year $t + 1$ may no longer appear exceptionally effective.[4] Yet even this conclusion is provisional, for it might be that above-average gains in the past have been due to above-average teachers in those years as well, something we could ascertain only by casting our net wider still to examine the performance of students who are not Mitchell's classmates in year $t + 1$ but have shared teachers with him in the past. To determine whether the extra gain in Figure 1 should be attributed to the teacher or to the student, TVAAS estimates teacher and student effects for all students, teachers, grades, and subjects simultaneously. In so doing, TVAAS uses data from after year $t + 1$ as well as before: if tests in subsequent years show Mitchell doing no better than the district average, we will be more inclined to attribute the high score in year $t + 1$ to the teacher, other things being equal.

The notion that TVAAS apportions the extra gain between teacher and student effects is only a useful heuristic. While we can regard this as the typical situation, in the case of any one student either the teacher or student contribution can exceed the whole of the extra gain. Thus, if Mitchell's classmates made substantially less progress than the district average, the estimated teacher effect would likely be negative. This would imply that the whole of the extra gain shown in Figure 1 plus some should be attributed to student-specific idiosyncratic factors.[5]

The TVAAS requires tests that are vertically linked—scores for fourth graders, for example, must be expressed on the same developmental scale as scores for third graders, fifth graders, etc. In order to compare the progress of students over time, test forms must be equated across years. The Tennessee Comprehensive Assessment Program uses CTB/McGraw-Hill's Terra Nova tests. As a consequence of some well-publicized problems with the equating of these tests by the publisher, TVAAS makes adjustments to the scores provided by CTB/McGraw-Hill. These adjustments are described in Appendix A.

The Estimation Sample

The TVAAS is estimated using a five-year data window, pooling data across all cohorts that passed through grades 3–8 (or any portion thereof) during that period. Thus, to obtain a given set of teacher effects, data from ten different student cohorts are used, as illustrated in Table 1. Cohorts are identified by Roman numerals. Cohort I comprises students who were eighth graders in spring, 1996, cohort II those who were seventh graders in 1996, etc. The last cohort, X, are students who enter the estimation sample as third graders in 2000. Observe that only two cohorts,

TABLE 1
TVAAS Estimation Sample Using Five-Year Data Window

Grade	1995–96	1996–97	1997–98	1998–99	1999–00
3	VI	VII	VIII	IX	X
4	V	VI	VII	VIII	IX
5	IV	V	VI	VII	VIII
6	III	IV	V	VI	VII
7	II	III	IV	V	VI
8	I	II	III	IV	V

V and VI, are present for the maximum possible number of years, five. Two cohorts, I and X, are included in the estimation sample in only one year.[6] The mean number of years per cohort is three.

The TVAAS uses the data in this five-year window to estimate teacher effects for teachers in grades 4–8 for years 1998–2000. A year later, when the 2001 data become available, the data window will be advanced one year, with corresponding changes in the representation of cohorts: Cohort I will drop out of the data set, cohort XI will enter, and the grades in which the other cohorts are contributing data will change. The model will be re-estimated, obtaining teacher effects for years 1999–2001. Through the layering of teacher effects in the TVAAS model, the addition of data from 2001 will improve the precision of some teacher estimates—for example, the effect fourth grade teachers had on cohort IX in 2000—as this cohort will also appear as fifth graders in year 2001. There will be a loss of precision at the other end: for example, because the 5th-grade performance of Cohort IV will no longer enter the estimation sample, there will be less information about these students available to estimate the impact of their teachers in later grades (for example, the impact of 8th grade teachers in 2000).

The TVAAS is estimated county by county. (In Tennessee a county generally contains a single school district.) Students who move out of the school system(s) in the county are not followed to their new location, although their data for the years they spent in the county are retained in the estimation sample. Students who move into the county do not bring their history with them: neither their previous test scores (if they attended school elsewhere in Tennessee) nor information about their past teachers is used. A student who arrives in the county for the first time, say, as a sixth grader in 1997, will be represented by the equation $Y^6_{97} = b^6_{97} + u^6_{97} + e^6_{97}$, with no terms for the contribution of teachers in the previous years. This raises the possibility that some part of that contribution will be attributed to the 6th-grade teacher, although this is likely to be a significant problem only when a large proportion of a teacher's class consists of students new to the system.

A student counts as part of a teacher's class only if the student has spent at least 150 days in that class. As a result of student mobility, test scores are reported for many students who are not matched to any teachers. Students who meet this threshold are supposed to be claimed by the teachers who provided them instruction in tested subjects; fractions of a student can be claimed if teachers shared responsibility. Because explicit instruction in reading falls off in the higher grades, many 7th-and 8th-grade students go unclaimed on the reading test.

In Table 2 we compare claimed to unclaimed students. In all subjects and years, mean scores are higher among claimed students. Claimed students usually have higher gains as well. These differences should not be taken as proof that teachers are shunning responsibility for low achievers. Mobile students are more likely to be eligible for free and reduced-price lunch and to be non-white. In addition, performance is likely to be lower among mobile students, both because mobility may be a symptom of an unstable family life and because movement between classes itself disrupts a student's academic year. Nonetheless, the high incidence of unclaimed students combined with the mean difference in gains raises the possibility that some estimated teacher effects might be substantially different if unclaimed students were assigned to teachers on a prorated basis. This is a problem not with the TVAAS model but with the legislative restrictions placed on the assessment system: in Tennessee, teacher effects are not to be based on gains of students who have spent less than the requisite 150 days in the classroom. In this

TABLE 2
Comparison of Claimed to Unclaimed Students

Subject	Grade	Percentage of claimed students	Difference in mean scores	Difference in mean gains	Difference in FRL eligibility	Difference in proportion non-White
Reading	3	88.48	10.15	0.00	–0.04	–0.03
Reading	4	88.47	9.27	1.33	–0.04	–0.04
Reading	5	85.34	7.74	–0.56	–0.04	–0.03
Reading	6	82.97	7.43	1.98	–0.02	–0.05
Reading	7	56.67	7.31	0.55	–0.08	–0.07
Reading	8	56.87	10.21	0.62	–0.10	–0.09
Lang. Arts	3	88.39	8.90	0.00	–0.04	–0.04
Lang. Arts	4	88.43	9.87	2.79	–0.04	–0.04
Lang. Arts	5	85.50	8.22	1.42	–0.03	–0.02
Lang. Arts	6	82.49	7.32	0.42	–0.04	–0.04
Lang. Arts	7	77.05	12.28	2.54	–0.07	–0.03
Lang. Arts	8	78.18	10.77	1.40	–0.04	–0.01
Math	3	88.54	8.82	0.00	–0.03	–0.03
Math	4	88.80	8.35	0.84	–0.05	–0.05
Math	5	85.76	9.20	1.32	–0.03	–0.02
Math	6	83.71	6.29	–0.23	–0.03	–0.02
Math	7	78.46	13.85	2.73	–0.08	–0.04
Math	8	78.08	14.15	2.12	–0.07	–0.02

Note. FRL = free and reduced price lunch, Lang. Arts = language arts.

sense, the TVAAS is measuring (or attempting to measure) just what the legislature has asked it to.[7]

Although the data for unclaimed students do not enter the estimation of teacher effects directly, they are retained in the sample for the estimation of the within-student covariance matrix and district means. Thus, TVAAS uses all the data available on every student within the five-year window.

3. Modifying TVAAS To Control for SES and Demographics

Let X_{2t} represent the value of SES and demographic covariates in year t that may affect test score gains, and let b_{2t}^{k} be the corresponding grade-k coefficient. (As before, we omit subscripts for subject, teacher, and student.) Then the modified TVAAS model is of the form

$$Y_t^k = b_t^k + \Sigma_{\mathrm{s}=0,\mathrm{S}} X_{2t-\mathrm{s}} b_{2t-\mathrm{s}}^k + \Sigma_{\mathrm{s}=0,\mathrm{S}} u_{t-\mathrm{s}}^k + e_t^k, \tag{5}$$

where S is one less than the number of years the student has spent in her current school system. In vector notation, $\mathbf{u^* = DZ'V^{-1}(y - X^*b^*)}$**,** where $\mathbf{X^*}$ includes $\mathbf{X_2}$.

Non-random assignment of teachers to students may induce a correlation between X_2 and u. If Equation 5 is estimated by maximum-likelihood, a nonzero value of b_2 might mean that X_2 affects gains or it might mean that X_2 is correlated with teacher quality, picking up effects that should be attributed to u. Our solution is to replace $\mathbf{b^*}$ by an estimate of $\mathbf{b}$ that is unbiased even when the correlation between $\mathbf{X_2}$ and teacher quality is non-zero. Our choice is the analysis of covariance estimator, $\mathbf{b_{cov}}$, obtained by treating the teacher effects in Equation 5 as fixed rather than random. The resulting estimator of teacher effects is

$$\mathbf{\tilde{u} = DZ'V^{-1}(y - Xb_{cov})}, \tag{6}$$

which, we will refer to as the "modified TVAAS." In Equation 6 the covariance matrices V and D are assumed known. In reality, of course, they must be estimated. This means estimating the TVAAS using test score "quasi-residuals" $\mathbf{y^* = y - X_2b_{2,cov}}$ as the dependent variable. Apart from replacing $\mathbf{y}$ with $\mathbf{y^*}$**,** this is identical to the original TVAAS. When estimated as a mixed model by MLE, the modified TVAAS yields estimates of the random effects $\mathbf{u}$ that preserve all the attractive properties of the original TVAAS with respect to handling missing data, producing shrinkage estimates of the **u,** etc.[8]

To simplify the estimation of $b_{2,\mathrm{cov}}$, we use test score *gains* as the dependent variable for grades 4–8. (In the equations for grade 3, the first year in which students are tested, the dependent variable necessarily remains the achievement *level.*) For a given student in year *t,* grade *k,* the model is $\Delta Y_t^k = b_t^k - b_{t-1}^{k-1} + X_{2t}^k b_2^k + u_t^k + e_t^k - e_{t-1}^{k-1}$, which is estimated separately for each subject/grade combination. The estimates of b_{2i}^k are then summed as $\Sigma_{\mathrm{s}=0,\mathrm{S}}\, \mathrm{X}_{2t-\mathrm{s}}^k b_{2t-\mathrm{s}}^k$, which is subtracted from Y_t^k to obtain Y_t^{*k}.

Selection and Measurement of SES and Demographic Covariates

The matrix X_2 contains a dummy variable for the student's eligibility for free and reduced-price lunch (FRL), an indicator for race other than white (NW), gender (MALE), and the two-way interactions of these three variables. This is a very limited set of controls. However, it contains the variables most likely to be available to a school district without extra costs of data collection. The question we are addressing is therefore the practical one of whether a district engaged in value-added assessment is better off using the limited information it possesses on student SES and demographics than not using it.

We also sought to introduce aggregate versions of these variables to control for peer effects. Due to a high degree of collinearity, we reduced their number to one: the percentage of students eligible for free or reduced-price lunch (PCTFRL). In several respects this variable is problematic. For those students taking departmentalized courses, we do not know which of a teacher's classes a student attended. For those students who were not claimed by teachers, we do not know even that much, although we do know their grade and school. As a result, we use two alternative measures: the percentage of FRL-eligible students in a student's school, and the percentage of FRL-eligible students in a student's grade-within-school. We estimate the model both ways.[9]

Introducing PCTFRL into the model raises another problem. Because this variable takes the same value for all of a teacher's students in a given year, there is a perfect linear dependence between it and the u_t in Equation 6. Accordingly, in the front-end models that include PCTFRL, we constrain teacher effects to be constant over time. Thus we are identifying the effect of PCTFRL through intertemporal variation in the make-up of the school, or grade-within-school, in which a teacher is employed.

To summarize, we estimate three versions of Equation 7: an equation containing only student-level covariates and teacher-by-year fixed effects; and two equations containing student-level covariates and teacher effects constrained to be the same across years, with PCTFRL measured at the school level in one and at the grade-within-school level in the other.

4. Front-End Results

The first-stage fixed effects models (which we will call "front-end" models) and the modified TVAAS were estimated using data from a single large, diverse Tennessee district. Tests were given in grades 3–8 in five subjects: reading, language arts, mathematics, social studies, and science. Although all five subjects were included when the model was estimated, our discussion focuses on results for the first three only.[10] The five-year data window comprised 1997–2001. Estimates of teacher effects were obtained for academic years 1998–1999, 1999–2000, and 2000–2001.

Because the front-end models were estimated with gain scores as the dependent variable, the estimation sample comprised students with adjacent-year observations.

However, all records for a given student were used to calculate **y*** and estimate Equation 5, provided data were not missing on any of the covariates.[11] Table 3 contains descriptive statistics on level and gain scores by subject and grade. Mean gains tend to fluctuate. There is some convergence of gains as grade level increases until seventh grade, when the standard deviation of language arts and math gains increases.

Table 4 contains selected results from the front-end models for reading. The dependent variable in grades 4–8 is the gain score, in grade 3 the level score. Coefficients are shown for FRL, race (NW = 1 if non-white), and the FRL-race interaction (FRL × NW). Results for gender are much less interesting and are not shown.

Coefficients on student-level covariates lie within a band of ±3. Given that mean annual gains are usually 12–15 points, these effects do not appear to be large. Frequently they fail to attain statistical significance. Although poor, minority students start out behind, as seen in the third coefficients, overall there is little tendency for poor, minority students to fall farther back over time.

Like the student-level covariates, PCTFRL is often insignificant. However, the point estimates are often much larger (as are the standard errors). The coefficients

TABLE 3
Average Level and Gain Scores, 1996–2001

Subject	Grade	Number of students	Mean score	*SD*	Percentage of students with gain scores	Mean gain	*SD*
Reading							
	3	22400	625.6	38.0	0.0	—	—
	4	21907	638.9	36.1	85.9	14.1	23.1
	5	20047	646.9	34.7	86.5	9.4	22.6
	6	19476	657.9	35.5	87.1	13.0	21.6
	7	19526	667.5	35.5	81.9	9.9	21.5
	8	17505	682.3	35.3	86.0	12.7	20.9
Language Arts							
	3	22343	630.1	35.7	0.0	—	—
	4	21897	643.3	36.0	85.7	15.0	24.0
	5	20028	651.4	34.2	86.4	10.3	22.8
	6	19481	660.5	34.0	87.0	9.9	21.3
	7	19460	668.9	37.7	82.0	10.5	22.8
	8	17437	683.6	35.2	85.9	12.1	21.9
Math							
	3	22406	597.8	34.9	0.0	—	—
	4	21887	621.8	35.0	85.8	23.9	25.4
	5	20017	637.5	35.5	86.2	18.2	23.8
	6	19466	656.7	36.6	86.8	18.9	22.8
	7	19474	673.0	38.9	81.9	17.1	24.5
	8	17471	692.8	39.7	85.8	16.5	23.5

TABLE 4
Front-End Estimates, Reading

Grade	Variable	Student-level covariates		Plus PCTFRL, grade-within-school		Plus PCTFRL, school-level	
		Coeff.	*SE*	Coeff.	*SE*	Coeff.	*SE*
3	FRL	–14.35	0.92	–14.58	0.92	–14.68	0.92
	NW	–13.64	0.84	–13.75	0.84	–13.75	0.84
	FRL × NW	0.12	1.03	0.14	1.02	0.15	1.02
	PCTFRL/100	—	—	–12.90	4.00	–11.71	4.76
4	FRL	–1.67	0.72	–1.55	0.72	–1.62	0.72
	NW	0.57	0.65	0.55	0.65	0.55	0.65
	FRL × NW	0.53	0.81	0.52	0.80	0.54	0.80
	PCTFRL/100	—	—	0.44	3.37	6.28	4.36
5	FRL	0.14	0.72	0.16	0.71	0.23	0.71
	NW	1.79	0.64	2.00	0.63	2.01	0.63
	FRL × NW	–1.00	0.79	–1.14	0.79	–1.12	0.79
	PCTFRL/100	.	.	3.58	3.30	–3.24	4.55
6	FRL	–1.05	0.74	–1.30	0.74	–1.34	0.74
	NW	–1.22	0.67	–1.30	0.67	–1.30	0.67
	FRL × NW	–0.22	0.84	–0.03	0.84	–0.04	0.84
	PCTFRL/100	—	—	2.02	4.27	12.29	5.24
7	FRL	1.62	1.02	1.77	1.02	1.76	1.02
	NW	2.52	0.87	2.40	0.87	2.40	0.87
	FRL × NW	–1.31	1.19	–1.28	1.19	–1.27	1.19
	PCTFRL/100	.	.	3.81	8.52	7.11	10.90
8	FRL	–1.74	1.12	–1.57	1.12	–1.55	1.12
	NW	0.92	0.93	0.88	0.93	0.87	0.93
	FRL × NW	0.82	1.35	0.68	1.34	0.69	1.34
	PCTFRL/100	—	—	–8.03	6.06	–14.81	8.55

Note. Coeff. = Coefficient, FRL = free and reduced price lunch, PCTFRL = percentage of students eligible for free and reduced price lunch, NW = non-White.

are quite unstable from one grade to the next. For example, when PCTFRL is measured at the school level, a 100% FRL class gains an extra 12.3 points in sixth grade. The same class loses ground (14.8 points) in eighth grade. Even with large standard errors, both of these estimates are statistically significant. They are much larger than the corresponding estimates in the models that use grade-within-school measures of PCTFRL.

We summarize the cumulative effect of SES and demographic covariates on reading achievement levels in Figure 2, which depicts mean achievement for a black male student who is FRL-eligible. In the higher curve, PCTFRL is set equal

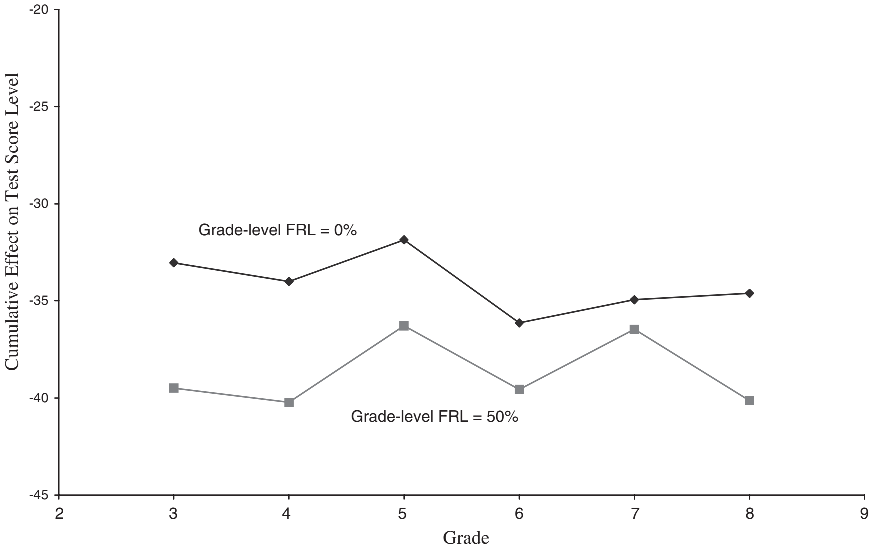

FIGURE 2. *Reading: Black, Male, FRL-eligible.*

to 0. In the lower curve, PCTFRL (at the grade-within-school level) equals 50. Because the effect of SES and demographic covariates on gains is inconsistent across grades, the curves fluctuate about a nearly flat trend.

Some of the same patterns are evident in math (Table 5). Coefficients on the student-level variables tend to be small, though they are nearly all negative, so that over time there is a larger cumulative effect. By and large, they are quite robust to how PCTFRL is measured. By contrast, coefficients on PCTFRL are very sensitive to whether this variable is measured at the school or grade-within-school level. There is a great deal of fluctuation across grades in the impact of PCTFRL, though most of the coefficients on PCTFRL are negative: high-FRL classes start behind and fall further behind. The drop is very great when PCTFRL is measured at the school level, though it should be kept in mind that the point estimates have large standard errors.

Table 6 contains front-end coefficients for language arts. Most of the earlier remarks apply here, too. Coefficients on student-level covariates are small and often insignificant. Coefficients on PCTFRL are generally larger, with large standard errors, and are unstable across grades. There is an enormous PCTFRL effect in 7th grade. The cumulative effect on achievement is depicted in Figure 4.

To sum up, the coefficients on the student-level covariates are generally small and robust towards changes in the model. By contrast, the coefficients on PCTFRL are larger and erratic. It is possible of course, that high values of PCTFRL promote learning gains in one grade and depress them in the next, if the curriculum builds in repetition across grades so that slow students can catch up. But changes in the

TABLE 5
Front-End Estimates, Math

		Student-level covariates		Plus PCTFRL, grade-within-school		Plus PCTFRL, school-level	
Grade	Variable	Coeff.	*SE*	Coeff.	*SE*	Coeff.	*SE*
3	FRL	−11.6	0.9	−11.6	0.9	−11.6	0.9
	NW	−9.6	0.8	−9.7	0.8	−9.7	0.8
	FRL × NW	0.2	1.0	0.4	1.0	0.4	1.0
	PCTFRL/100	—	—	−14.3	3.8	−20.7	4.5
4	FRL	−4.3	0.8	−4.1	0.8	−4.0	0.8
	NW	0.1	0.7	0.3	0.7	0.3	0.7
	FRL × NW	2.4	0.8	2.2	0.8	2.2	0.8
	PCTFRL/100	—	—	−3.3	3.3	−7.8	4.3
5	FRL	0.1	0.8	0.1	0.8	0.2	0.8
	NW	0.7	0.7	0.5	0.7	0.5	0.7
	FRL × NW	−1.2	0.9	−1.4	0.9	−1.4	0.9
	PCTFRL/100	—	—	−0.7	3.5	−7.2	4.6
6	FRL	−1.6	0.8	−1.5	0.8	−1.6	0.8
	NW	−0.5	0.7	−0.6	0.7	−0.7	0.7
	FRL × NW	0.8	0.9	1.0	0.9	1.0	0.9
	PCTFRL/100	—	—	−4.7	4.3	1.2	4.8
7	FRL	−0.4	1.0	−0.5	1.0	−0.5	1.0
	NW	−0.5	0.8	−0.5	0.8	−0.5	0.8
	FRL × NW	−1.4	1.1	−1.3	1.1	−1.3	1.1
	PCTFRL/100	—	—	−3.0	7.8	−18.0	8.7
8	FRL	−1.7	1.1	−2.0	1.1	−2.0	1.1
	NW	0.3	0.9	0.2	0.9	0.2	0.9
	FRL × NW	1.6	1.2	2.1	1.2	2.1	1.2
	PCTFRL/100	—	—	−12.7	6.2	−14.8	8.2

Note. Coeff. = Coefficient, FRL = free and reduced price lunch, PCTFRL = percentage of students eligible for free and reduced price lunch, NW = non-White.

sign and size of the PCTFRL coefficient often depend on whether PCTFRL is measured at the school or the grade-within-school level. It is difficult to explain this as the effect of curriculum design.

Coefficients on PCTFRL are also implausibly large in grades 7 (language arts and math) and 8 (math), though it is not implausible that there would be some increase in peer effects as students enter adolescence. The large impact of PCTFRL in seventh-, and eighth-grade math may also reflect decisions in the classroom about curriculum coverage.[12] In another paper we have explored possible sources of bias in the coefficients on PCTFRL (Ballou, Sanders, & Wright, 2003). The

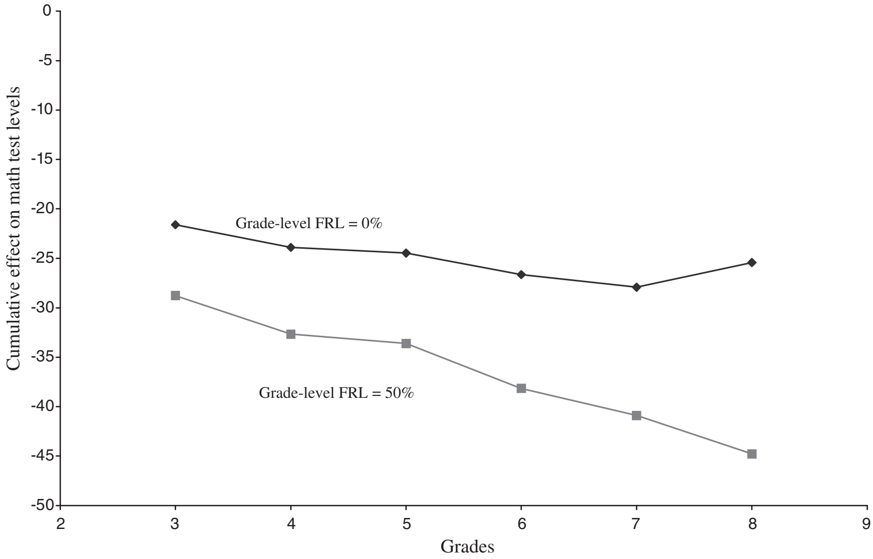

FIGURE 3. *Math: Black, Male, FRL-eligible.*

results of that investigation are inconclusive: although there are signs of bias, it does not appear to account for the pattern of coefficients in Tables 4–6. For now we acknowledge that the results for PCTFRL in some grades and subjects are suspect and move on to the central question: what effect does the introduction of covariates have on the TVAAS?

5. Teacher Effects in the Modified TVAAS

More than 5,000 teacher effects are estimated for the sample years. To illustrate the magnitude of these effects and their standard errors, we display the estimates for 7th- and 8th-grade mathematics teachers along with confidence intervals in Figure 5. Generally speaking teacher effects are not estimated with great precision. Nonetheless, approximately one fifth of these intervals lie entirely in the positive or the negative region of the graph. This proportion drops considerably in the other subjects, to eight percent of language arts teachers and only five percent of reading teachers in the same grades.

Coefficients from the three front-end models are used to produce three estimates of teacher effects, following Equation 6. Counting the estimates from the original, unmodified TVAAS, there are four sets of teacher effects in all. We begin by examining the pair-wise correlations among them. Four sets of estimates yield six pairwise correlations for every subject/grade combination. In Table 7 we display the *lowest* of these six correlation coefficients. Correlations across models are at least .9 in all but five subject/grades. The problematic subject/grades are language arts, grade 7, and math, grades 7 and 8. In the first, the correlation between models con-

TABLE 6
Front-End Estimates, Language Arts

		Adjustments:					
		Student-level covariates		Plus PCTFRL, grade-within-school		Plus PCTFRL, school-level	
Grade	Variable	Coeff.	*SE*	Coeff.	*SE*	Coeff.	*SE*
3	FRL	−14.1	0.87	−14.2	0.87	−14.2	0.86
	NW	−10.9	0.8	−10.9	0.79	−10.9	0.79
	FRL × NW	0.9	0.97	1	0.96	1	0.96
	PCTFRL/100			−12.9	3.8	−18.1	4.5
4	FRL	−1.94	0.75	−1.6	0.74	−1.5	0.74
	NW	0.2	0.67	0.1	0.67	0.1	0.67
	FRL × NW	0.6	0.83	0.23	0.82	0.24	0.82
	PCTFRL/100			3.5	3.5	2.7	4.5
5	FRL	1.39	0.75	1.21	0.75	1.14	0.75
	NW	1.78	0.67	1.73	0.66	1.74	0.66
	FRL × NW	−0.8	0.83	−0.7	0.83	−0.78	0.82
	PCTFRL/100			2.98	3.66	8.52	4.94
6	FRL	−0.69	0.75	−0.94	0.74	−0.95	0.74
	NW	−0.62	0.67	−0.86	0.66	−0.86	0.66
	FRL × NW	1.96	0.84	2.19	0.84	2.19	0.84
	PCTFRL/100			−2.19	4.32	−0.35	5.33
7	FRL	−1.91	0.91	−1.95	0.9	−1.98	0.9
	NW	0.17	0.78	0.32	0.78	0.35	0.78
	FRL × NW	0.58	1.05	0.34	1.04	0.31	1.04
	PCTFRL/100			−33.9	0.52	−39.1	6.12
8	FRL	1.12	1.01	0.94	1.01	0.95	1.01
	NW	−0.84	0.84	−0.9	0.84	−0.9	0.84
	FRL × NW	0.43	1.18	0.68	1.18	0.7	1.18
	PCTFRL/100			8.35	4.38	8.72	5.01

Note. Coeff. = Coefficient, FRL = free and reduced price lunch, PCTFRL = percentage of students eligible for free and reduced price lunch, NW = non-White.

taining PCTFRL and the others is less than .6. This is the largest discrepancy in our results. In 7th- and 8th-grade math, the model containing PCTFRL at the school level is out of line with the other results. These are, of course, the subject/grades where we obtained very large coefficients on PCTFRL in the front-end models.

Estimates that are highly correlated can differ in ways important to policy makers. Suppose (for purposes of argument) that our front-end models simply halve the teacher effects in the original, unadjusted TVAAS, with standard errors unchanged. The four sets of estimates will be perfectly correlated but the original

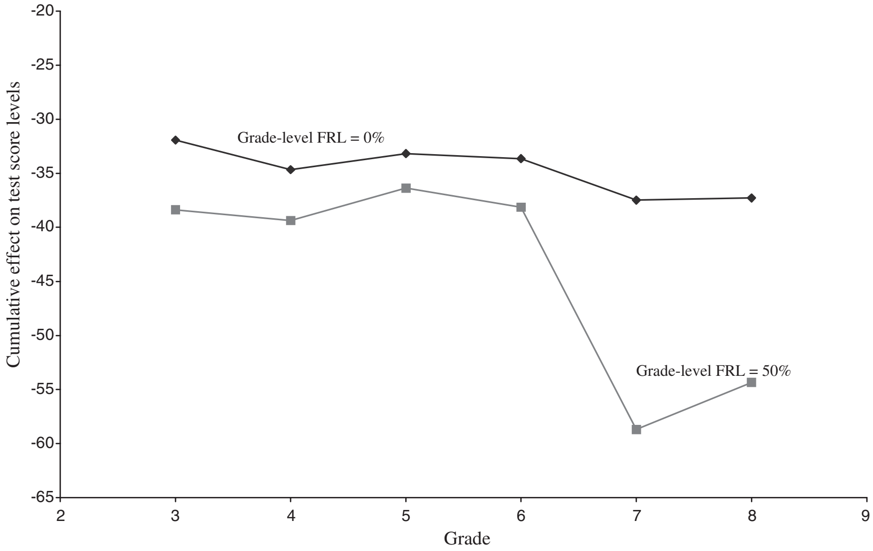

FIGURE 4. *Language Arts: Black, Male, FRL-eligible.*

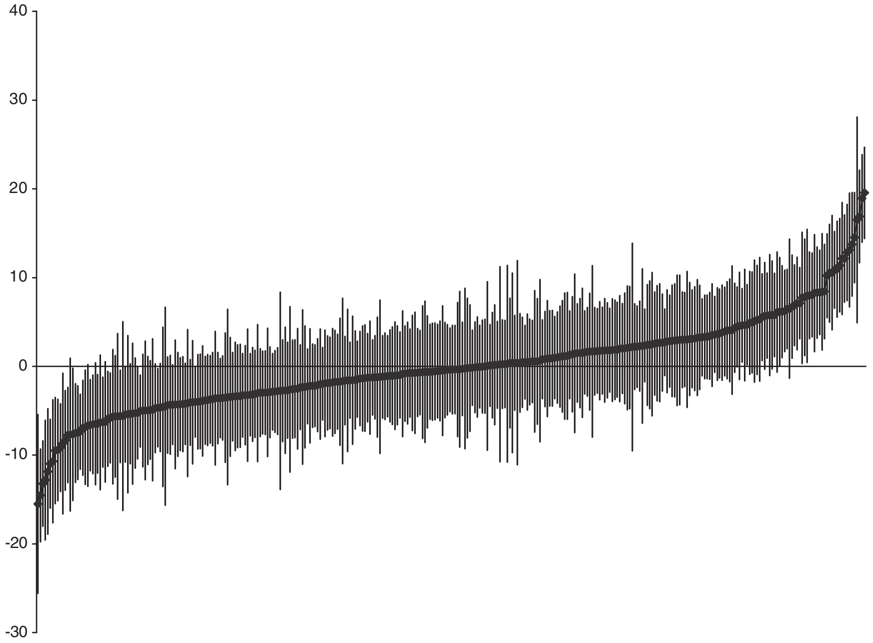

FIGURE 5. *Teacher effects with 95% confidence intervals: 7th-and 8th-grade mathematics.*

TABLE 7
Correlation Across Models

Grade	Reading	Language Arts	Math
4	.96	.98	.92
5	.95	.97	.89
6	.90	.97	.93
7	.91	.53	.76
8	.96	.86	.79

Note. Entries are the smallest of the six pair-wise correlation coefficients for teacher effects estimated by the four models, the original TVAAS and the three models with front-end adjustments.

model will identify many more teachers who are better or worse than average. Thus we ask how many exceptional teachers the four models find, and how often they agree on the identity of these teachers.

We classify a teacher as exceptional if the teacher effect divided by its standard error exceeds 1.5 in absolute value. Table 8 compares the number of teachers so classified by the unadjusted model to the number in each of the other three models. Teachers in the "yes" ("no") row did (not) meet the cutoff when the TVAAS was estimated without adjustments for SES and demographics. The columns headed "no" and "yes" indicate how many of these teachers met this criterion in the other models. There are relatively few observations off the main diagonal when only student-level covariates are added. Substantially more discrepancies arise when the front-end model controls for PCTFRL, though as already noted, these are the estimates in which we have less confidence.

Discrepancies between the models are probably of little concern if the disagreements are minor. Teachers classified "exceptional" by the unadjusted model

TABLE 8
Teachers Significantly Above or Below Average

		Front-End Models Control For:					
	Unadjusted	Student covariates		Student plus grade-level PCTFRL		Student plus school-level PCTFRL	
Subject	Significant	No	Yes	No	Yes	No	Yes
Reading	No	1580	9	1572	17	1530	59
	Yes	13	59	8	64	4	68
Lang. Arts	No	1561	27	1518	70	1513	75
	Yes	15	145	37	123	38	122
Math	No	1237	25	1181	81	1099	163
	Yes	18	365	58	325	77	306

Note. Lang. Arts = Language Arts.

may nearly be “exceptional” in the adjusted models. In Table 9 we display the absolute mean standardized teacher effects in the cells depicted in Table 8. The relevant cells have been highlighted in boldface. There is high agreement between the unadjusted model and the model with only student-level covariates. Teachers that make the cutoff in the former but miss the cutoff in the latter are clearly superior to those that miss the cutoff in both models. In math, for example, their mean standardized effect is 1.39, just .11 below the threshold. There is less agreement when the first stage adjusts for PCTFRL at the grade level, especially in language arts, and when it adjusts for PCTFRL at the school level, especially in math.

Finally, we ask how a teacher of disadvantaged students fares in the modified TVAAS compared to the original TVAAS. The variable of interest is the standardized teacher effect in the modified TVAAS minus the standardized teacher effect in the original TVAAS. We regress this variable on PCTFRL (at the grade or school level), the percentage of the teacher’s students who are non-White, and dummy variables for years. Table 10 reports the difference that the introduction of SES and demographic controls makes to a teacher whose class is 100% poor or 100% minority. Estimation samples pool data from grades 4 to 8. On average it makes little difference to teachers of disadvantaged students whether such controls are introduced at the student level. For example, a math teacher whose students are all FRL-eligible gets a boost of .26 in his standardized effect when controls for FRL-eligibility are entered at the student level. An increase of this size moves the average teacher only one sixth of the way to the threshold for “exceptional” teachers established above. However, the same is not true of front-end models that include PCTFRL. Particularly in math, front-end adjustments for PCTFRL have a substantively significant impact on the standardized teacher effect.

TABLE 9
Mean Absolute Standardized Teacher Effects

		Significant with student-level covariates?		Significant with student covariates plus grade-level PCTFRL?		Significant with student covariates plus school-level PCTFRL?	
Subject	Significant in unadjusted model?	No	Yes	No	Yes	No	Yes
Reading	No	0.42	1.73	0.44	1.79	0.53	2.19
	Yes	**1.23**	2.41	**1.22**	2.55	—	2.85
		no	yes	no	yes	no	yes
Lang Arts	No	0.5	1.62	0.53	2.09	0.52	2.2
	Yes	**1.3**	2.28	**1**	2.31	**1.07**	2.3
		no	yes	no	yes	no	yes
Math	No	0.62	1.6	0.6	1.79	0.71	2.06
	Yes	**1.39**	2.44	**1.2**	2.55	**0.98**	2.54

Note. PCTFRL = percentage of students eligible for free and reduced price lunch.

TABLE 10
Impact of Front-End Adjustments on Standardized Teacher Effects for Teachers of Disadvantaged Students

	Front-End Models Control For:		
	Student-level covariates	Plus PCTFRL at grade level	Plus PCTFRL at school level
Effect of front-end adjustments when PCTFRL = 100%			
Reading	+0.17	−0.11	−0.98
Language Arts	+0.15	+0.55	+0.47
Math	+0.26	+1.4	+2.5
Effect of front-end adjustments when PCTNW = 100%			
Reading	+0.36	+0.39	+0.31
Language Arts	−0.08	+0.01	+0.1
Math	+0.09	+0.28	+0.47

Note. PCTFRL = percentage of students eligible for free and reduced price lunch, PCTNW = percentage non-White.

6. Discussion

By multiple measures, inclusion of SES and demographic covariates at the student level has little effect on the TVAAS. For all subjects and grades, the correlation between initial teacher effects and those obtained after the introduction of these controls exceeds .9. The adjusted and unadjusted models agree far more often than they disagree on the identity of the teachers who are significantly above or below average: agreement is 2.7 times more likely than disagreement in reading, 3.5 times more likely in language arts, and 8.5 times more likely in mathematics.[13] As we have just seen, controlling for student-level covariates has only a moderate impact on estimated teacher effects, even for teachers whose classes are entirely poor or entirely minority.

Controlling for SES aggregated at the grade or school level has a substantial impact on TVAAS estimates in some subjects and grades, but these models are suspect: the coefficients on PCTFRL used in these adjustments have large standard errors, swing erratically from positive to negative, are not robust to changes in the measurement of PCTFRL, and are implausibly large in some grades and subjects.

To conclude, the results in which we have the greatest confidence—where we have controlled for SES and demographics at the student-level—exhibit very little change from the original, unmodified TVAAS. This will come as a surprise to some researchers. We consider four possible explanations for this finding.

(1) *There is insufficient variation in the make-up of teachers' classes.* If the great majority of teachers have roughly the same mix of poor and non-poor students, White and non-White, then adjusting for demographics and SES will not change estimated teacher effects.

There is no support in the data for this hypothesis. The mean proportion of FRL-eligible students over teachers is .47, with a standard deviation of .23. The corresponding figures for the proportion of non-white students is .54 and .21. If SES and race matter, clearly different teachers will be affected to varying extents.

(2) *The impact of SES and demographics on achievement growth is not large enough to make an appreciable difference to estimated teacher effects.* In Figures 2, 3, and 4, we depicted the trajectory of mean test scores for a student who was poor, Black, and male. The upper curve in each figure corresponded to the model containing only these student-level covariates. The combined effect of these variables on mean scores was at most five points from third to eighth grades. It may be that this is simply not enough to make a difference to the estimation of teacher effects.

If this explanation is correct, it should hold not just for the TVAAS but for the fixed effects estimates as well. Accordingly, we test it by comparing the teacher effects estimated in our front-end models to teacher fixed effects obtained from a model without any SES or demographic covariates. The latter is, of course, simply the mean gain for a teacher's students over the year. There are considerable discrepancies between the two sets of estimates. In reading the correlation between the two is only .31, in language arts .35, and in math .39. Clearly controlling for demographics and SES makes a substantial difference when teacher effects are estimated in this simple manner. By contrast, TVAAS teacher effects obtained after first controlling for student-level demographics and SES were highly correlated ($\rho > .9$) with those from the original, unadjusted TVAAS model. This suggests the explanation lies in those features of the TVAAS that the fixed-effects estimator lacks.

(3) *The high correlation between the adjusted and unadjusted TVAAS estimates is caused by shrinkage.* Estimated teacher effects in the original TVAAS are given by $\mathbf{u^*} = \mathbf{DZ'(ZDZ' + R)^{-1}(y - Xb)}$, where $\mathbf{Xb}$ is the matrix of district mean scores, $\mathbf{D}$ is the diagonal covariance matrix of $\mathbf{u}$, $\sigma^2_u\mathbf{I}$, and $\mathbf{R}$ is the block-diagonal variance-covariance matrix of student-level error terms. Even if $\mathbf{R}$ were a diagonal matrix $\sigma^2_e\mathbf{I}$, $\mathbf{u^*}$ would differ from the fixed effects estimator by the presence of the additive term $\sigma^2_e\mathbf{I}$ in the matrix $(\mathbf{ZDZ'} + \sigma^2_e\mathbf{I})$, which has the effect of shrinking $\mathbf{u^*}$ toward the overall mean of zero, with the degree of shrinkage dependent on the ratio of noise to signal in the within-teacher mean gains.[14] Thus it is useful to think of the difference between $\mathbf{u^*}$ and the fixed effects estimator as stemming from two sources: one the exploitation of the covariance structure $\mathbf{R}$ and the other shrinkage based on the ratio of noise to signal in the within-teacher mean scores. The two can be separated practically as well as conceptually: if the model were estimated using only one observation per student (say, by estimating it separately for each subject/grade combination), there would be no within-student covariances to exploit and $\mathbf{R}$ would reduce to a diagonal matrix. However, there would still be shrinkage. Is it shrinkage in this sense, absent any of the information contained in other test results, that accounts for our finding that estimated teacher effects are not very sensitive to the omission of SES and demographic variables from the TVAAS?

The question is worth asking, for shrinkage can have a dramatic effect on the correlation of two random variables that are both noisy predictors of a third variable. For example, consider $\chi_i = \omega_i + \varepsilon_i$ and $\xi_i = \omega_i + \upsilon_i$, where ε_i and υ_i are random variables independent of each other and of ω_i. χ_i and ξ_i are alternative predictors of the unobserved ω_i. The variances of ε_i and υ_i are not constant but depend on *i*. We shrink the predictors by multiplying each χ_i and ξ_i by their reliability coefficients $\lambda_i = \text{var}(\omega_i)/[\text{var}(\omega_i) + \text{var}(\varepsilon_i)]$ and $\gamma_i = \text{var}(\omega_i)/[\text{var}(\omega_i) + \text{var}(\upsilon_i)]$. The shrunk predictors are therefore weighted variables, with the weights an increasing function of the correlation between the original, unshrunk variables. Obviously the correlation between variables so weighted will be higher than the correlation between the original χ and ξ.

Like the two predictors in this example, the **u*** obtained in the original TVAAS and the $\tilde{\mathbf{u}}$ in the modified TVAAS are both noisy predictors of the true **u.** Thus it might seem that weighting both by their reliability coefficients would increase their correlation. If so, the high correlation between the two might be largely an artifact of shrinkage. However, shrinkage in the TVAAS does not follow this pattern, because the error in the two predictors is not independent. Given that we are assuming **R** diagonal, it suffices to examine the effect of shrinkage on the two fixed effects estimators. Assume that the true model is $y_{ij} = \mu_{ij} + u_j + e_{ij}$, where y_{ij} is the gain of the ith student of teacher j, μ_{ij} is the contribution of student-level SES and demographic factors, u_j is the teacher effect, and e_{ij} the student-level error. When this model is fit using a fixed effects estimator, the estimated teacher effect is $u_{1j} = y_j - \mu_j$, where the latter are the within-teacher means of y and μ, respectively. The error in this estimate is e_j, which has a variance of σ^2/n_j. The reliability of u_{1j} is $\lambda_{1j} = 1 - [\sigma^2/n_j]/\text{var}(u_{1j})$. Alternatively we can estimate the teacher effect using the mean gain score, $u_{2j} = y_j$, which absorbs μ_j into the teacher effect. The reliability $\lambda_{2j} = 1 - [\sigma^2/n_j]/\text{var}(u_{2j})$. The covariance of u_{1j} and u_{2j} is $\text{var}(u_j + e_j)$. λ_{1j} and λ_{2j} are therefore inversely related to the correlation between u_{1j} and u_{2j}, as e_j is a component of both the latter. If the same value of σ^2/n_j is used to estimate both λ_{1j} and λ_{2j}, the correlation between the shrinkage estimates will be lower than the correlation between the original u_{j1} and u_{j2}, as we are giving greater weight to observations with a smaller covariance.[15]

To summarize: the degree of shrinkage is an increasing function of the contribution of the student errors to the within-teacher means. Giving less weight to teacher effects that contain a large common noise component reduces the correlation between the original, unadjusted TVAAS estimates and estimates obtained after the front-end adjustments. Accordingly, shrinkage is not responsible for the high correlation between the unadjusted and the modified TVAAS teacher effects.

(4) *SES and demographic covariates add little information beyond that contained in the covariance of test scores.* **R** is not a diagonal matrix but a block diagonal matrix, in which the off-diagonal elements in each block represent the covariances of student error terms across subjects and years. Premultiplication of $(\mathbf{y} - \mathbf{Xb})$ by $\mathbf{DZ}'(\mathbf{ZDZ}' + \mathbf{R})^{-1}$ does not merely shrink the mean within-teacher gain by a factor that depends on the ratio of signal to noise; it reweights each of the student scores entering that mean by incorporating information from other subjects

and years through the covariance matrix **R.** Thus, above average gains in a given subject and year are attributed to the teacher only to the extent that we would not have expected such gains on the basis of students' complete history of test results.

If the off-diagonal elements of **R** were small, this would not matter very much. This is not the case, however. In the original, unmodified TVAAS, correlations between scores in different subjects typically range from .6 to .7. Correlations between scores in the same subject at different grades average about .8. Thus, scores on other tests contain a considerable amount of information about how well a student is likely to do on any one examination. In this sense, the longitudinal history of a student's performance serves as a substitute for other data on the student, such as SES and demographic controls.

This is evident when we compare the estimate of **R** in the original, unmodified TVAAS to the TVAAS with front-end adjustments for student-level covariates. There is a substantial decline in the size of the elements of **R** in the latter model, with the typical element falling about 18%. Clearly, in the absence of explicit controls for SES and demographics, other test scores contain much of the same information. By contrast, when no use is made of test covariances, as in the two fixed effects estimators u_1 and u_2, inclusion of SES and demographic controls makes a considerable difference.

7. Conclusion

The TVAAS is one of the most prominent examples of value-added assessment of teachers. Compared to other assessment systems, the TVAAS uses a highly parsimonious model that omits controls for SES, demographic, or other factors that influence achievement. As a result researchers who believe that even in a value-added model, controls are needed to account for the possible influence of such factors on the rate at which students learn, have criticized the TVAAS.

Controlling for SES and demographic factors is not as straightforward as might first appear, due to the possible correlation between these variables and teacher quality. We first remove the influence of the latter by estimating the effect of SES and demographics in a model with teacher fixed effects. We then remove the influence of SES and demographic factors from student test scores, using the residual scores as dependent variables in the TVAAS.

We find that controlling for SES and demographic factors at the student level makes very little difference to teacher effects estimated by the TVAAS. There is a larger impact when our controls include the percentage of students in the school or grade who are eligible for free or reduced-price lunch, but these models are not estimated with as much precision and we do not have much confidence in the final results.

The concerns of TVAAS critics are well placed when less sophisticated models are used to estimate teacher effects. It makes a considerable difference to a simple fixed-effects estimator whether the model includes student-level controls for SES and demographics. Unlike the TVAAS, this estimator does not exploit the covariance of tests in different subjects and grades.

The importance of the covariance matrix explains some of the choices that have been made in the design of the TVAAS. To exploit this information fully, a student's entire test score history must be used. Student data exhibit frequent lacunae: students miss tests, their eligibility for free or reduced-price lunch is not recorded, etc. The TVAAS is estimated in test-score levels so that observations need not be discarded whenever a missing score from a prior year prevents the calculation of a simple gain score. Likewise, one reason for the highly parsimonious specification of the TVAAS is the frequency of missing data on SES and demographics. Thus, in some respects, the analysis in this article has actually been unfair to the TVAAS, as our TVAAS estimation sample has been limited to the observations for which we could estimate the front-end gain score models. In practice, the TVAAS uses a larger estimation sample with corresponding improvements in efficiency.

This research does not resolve all questions about the TVAAS. In a Monte Carlo analysis of mixed models of the TVAAS type, McCaffrey et al. (2004) have shown that the extent to which the within-student covariance matrix can substitute for information on student demographics and SES depends on the extent to which the school system "mixes" students and teachers. In the district we have studied, there has been sufficient mixing, but we do not know whether this will be true of other school systems with greater social and racial stratification. It is also plausible that the make-up of the school influences achievement through peer effects. Because the covariance structure of the TVAAS does not capture the effects of student clustering (covariances across students, even those in the same class, are zero), we cannot be confident that the TVAAS controls for contextual variables in the same way that it controls for the influence of student-level SES and demographics. Several unresolved issues remain regarding the best way to control for peer influences, however, and our work in this area continues.

Notes

[1] This approach is anticipated in Raudenbush and Willms (1995), who also propose using a pooled, within-school regression coefficient on the student-level background variable (the prior test score) in their model. However, this estimator is not offered as a way of dealing with bias. Because the mean school-level score on the pre-test is also in the model as a context variable, the coefficient on the individual score is necessarily a pooled, within-school estimate (Aitkin & Longford, 1986). In this model the authors (rightly) are concerned with the bias arising when the school effect is not orthogonal to the context variable. Their solution is to use estimates from models with and without the context variable to bracket the true variance of the school effect. By contrast, we propose to exploit the panel character of the TVAAS data to obtain unbiased estimates of the effect of both student-level and school-level covariates, such as school mean SES.

[2] This must be qualified in the following limited sense. If the students of the teacher whose data are missing all go on to the same teacher the next year, and, moreover, if there are no other students in the class of the second teacher, then it would not be possible to identify u_{95}^{3} separately from u_{96}^{4}; however, this is a very unlikely scenario.

[3] As the formulation in Equations 2a and 2b shows, teacher effects can be regarded as estimates of value-added only if there is a prior-grade score to subtract from the current-year score. Because testing starts in the third grade, no valued-added estimates are obtained for third grade teachers.

[4] Evidence on student performance in other years and subjects bears on the decision through the covariance matrix R. In principle one could allow teacher performance in other subjects and years to have a symmetric influence, but this is precluded by the assumption that the covariance matrix of teacher effects, D, is diagonal.

[5] This description of the TVAAS has been necessarily brief. Readers desiring further detail will find a numerical illustration in Sanders, Saxton and Horn (1997).

[6] Although every cohort spends six years in grades 3–8, in no case are all six years of data used.

[7] From a research standpoint, one might want to know how sensitive the teacher effects are to the omission of students who are presently unclaimed. Random imputation might be used for this purpose, but it would be extremely computer intensive: simply estimating the TVAAS once for a large district can take nearly a week.

[8] Our estimate of the standard errors of the $\tilde{u}$ is $DZV^{-1} = (Z'R^{-1}Z + D^{-1})^{-1}$, the two-level MLF formula given in Bryk and Raudenbush (1992). This formula ignores estimation error in the coefficients in the first-stage model.

[9] PCTFRL and FRL are also measured with error inasmuch as some students eligible for free and reduced-price lunch decline to apply for it. Our estimator adjusts for measurement error, as described in Ballou and Wright (2003).

[10] The tests in social studies and science are shorter and appear to lack the validity of the other examinations. All tests are used in estimating the model, however, to enhance the information contained in the covariance matrix **R.**

[11] Students whose value of FRL (or, much less commonly, race or sex) was missing were therefore dropped from the estimation sample. However, the expected value of FRL for these students was used to compute the aggregate PCTFRL measures at the school and grade-within-school levels. The incidence of missing FRL values ranged from 8.5% in 1997 to 14.2% in 1995. The number of missing values for race or sex was negligible.

[12] These are transitional grades in which students prepare for the study of algebra. (Eighth grade students who actually enroll in Algebra I are given a different test in the TCAP and do not enter the estimation sample.) How much pre-algebra material to introduce is left to the discretion of the classroom instructor, resulting in substantial variation in the mathematics curriculum for these grades. Some teachers prepare their students for the pre-algebra items that begin to appear in the Terra Nova in these grades; those who review arithmetic do not. If these decisions are correlated with SES, students in low SES classes who are ready for more advanced material may not be getting it.

[13] These numbers are the ratio of the entries in the lower right-hand cells of Table 8 to the number of observations off the main diagonal.

[14] The former depends on $\sigma^2_{\mathbf{e}}$ and the number of students in the class, the latter on $\sigma^2_{\mathbf{u}}$.

[15] Matters are not quite this straightforward, as the same value of σ^2/n_j is not used to compute both λ_{1j} and λ_{2j}. When the teacher effect is estimated using the mean gain, $u_{2j} = y_j$, the value of σ^2/n_j that is used in λ_{2j} contains the within-teacher variance in μ_{ij}, which tends to mitigate the effects just described.

Appendix: Adjustments to Test Scales

Since its inception, the Tennessee Comprehensive Assessment Program has used tests prepared by CTB/McGraw-Hill, initially the CTBS Series IV, then, starting in 1998, the Terra Nova series. Disparities between the 1999 results and those of earlier years raised serious doubts about the procedures used to equate the forms of the tests used in different years. Tennessee's experience in this regard was not unique. Erratic test results convinced school officials elsewhere (e.g., New York, Indiana) that there were serious flaws in the scoring of these tests (Steinberg & Henriques, 2001).

After considerable discussion between researchers responsible for TVAAS and representatives of CTB/McGraw-Hill, the company agreed to make changes in the scoring of the 1999 tests. After this revision the 1999 tests appeared reasonably well equated with the 1998 tests. However, there was no indication that the fundamental deficiencies in the estimation of item difficulty parameters had been addressed. In order to avoid future recurrences of the problem, TVAAS has made a statistical correction to the test data received from CTB/McGraw-Hill in subsequent years. The correction is illustrated in Figure A1 (using hypothetical data). Mean school-level scores in a particular subject (e.g., reading) are plotted for 1999 and 2000 against the 1998 scores and regression lines are fit to the plot. As is evident, the 2000 scores show a steeper relationship to those of 1998 than do the 1999 scores: between 1998 and 2000 there appears to have been greater growth at the high end of the scale than the low end. The correction to the 2000 data consists of rotating the 2000 line about the point where it intersects the 1999 line until the two lines coincide. For these illustrative data, the correction raises scores below the mean and reduces scores above the mean. These correction factors are then applied to the student-level data. These transformed data are then used in all subsequent analyses. This procedure has been followed in post-1999 years in which a substantial disparity exists between the slopes of the current-year line and the 1999 line.

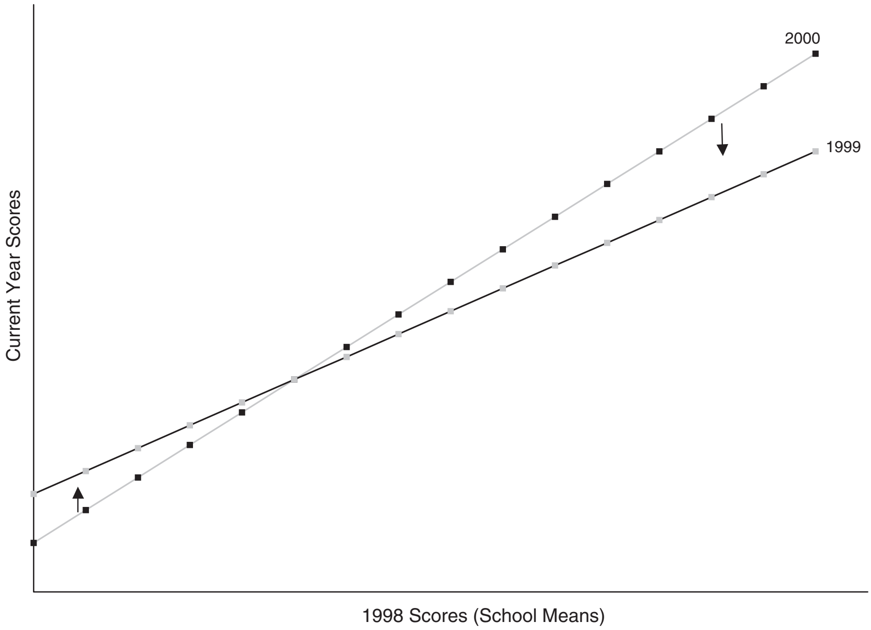

FIGURE A1. *Test scale correction.*

References

Aitkin, M., & Longford, N. (1986). Statistical modeling issues in school effectiveness studies. *Journal of the Royal Statistical Society, A, 149,* 1–43.

Ballou, D., & Wright, P. (2003). *Measurement error in free and reduced-price lunch.* Unpublished manuscript.

Ballou, D., Sanders, W., & Wright, P. (2003). *Student background and value-added assessment: Is there a middle way?* Unpublished manuscript.

Bryk, A. S., & Raudenbush, S. W. (1992). *Hierarchical linear models.* Newbury Park, CA: Sage.

Darling-Hammond, L. (1997). Toward what end? The evaluation of student learning for the improvement of teaching. In J. Millman, (Ed.), *Grading teachers, grading schools. Is student achievement a valid evaluation measure?* (pp. 248–263). Thousand Oaks, CA: Corwin.

Kupermintz, H. (2002). *Teacher effects as a measure of teacher effectiveness: Construct validity considerations in TVAAS (Tennessee Value-Added Assessment System).* CSE Technical Report 563. University of California, Los Angeles.

Linn, R. L. (2001). *The design and evaluation of educational assessment and accountability systems.* CSE Technical Report 539. University of California, Los Angeles.

McCaffrey, D. F., Lockwood, J. R., Koretz, D., Louis, T. A., Hamilton, L., Kirby, S. (2004). Models for value-added modeling of teacher effects. *Journal of Educational and Behavioral Statistics, 29*(1), 67–102.

Popham, J. (1997) The moth and the flame: Student learning as a criterion of instructional competence. In J. Millman, (Ed.), *Grading teachers, grading schools. Is student achievement a valid evaluation measure?* (pp. 264–274). Thousand Oaks, CA: Corwin.

Raudenbush, S. W., & Willms, J. D. (1995). The estimation of school effects. *Journal of Educational and Behavioral Statistics, 20*(4), 307–335.

Sanders, W. L., Saxton, A. M., Horn, S. P. (1997). The Tennessee Value-Added Assessment System: A quantitative, outcomes-based approach to educational assessment. In J. Millman, (Ed.), *Grading teachers, grading schools. Is student achievement a valid evaluation measure?* (pp. 137–162). Thousand Oaks, CA: Corwin.

Sanders, W. L., & Rivers, J. C. (1996). *Cumulative and residual effects of teachers on future student academic achievement.* University of Tennessee Value-Added Research and Assessment Center.

Searle, S. R. (1971). *Linear models.* New York: John Wiley & Sons.

Steinberg, J., & Henriques, D. (2001). When a test fails the schools, careers and reputations suffer. *New York Times,* May 21, 2001. Retrieved October 17, 2002, from http://query.nytimes.com/search/restricted/.

University of Florida. (2000a). *Prototype analysis of school effects.* Value-Added Research Consortium.

University of Florida. (2000b). *Measuring gains in student achievement: A feasibility study.* Value-Added Research Consortium.

Webster, W. J., & Mendro, R. L. (1997). The Dallas Value-Added Accountability System. In J. Millman, (Ed.), *Grading teachers, grading schools. Is student achievement a valid evaluation measure?* (pp. 81–99). Thousand Oaks, CA: Corwin.

Wright, S. P., Horn, S. P., & Sanders, W. L. (1997). Teacher and classroom context effects on student achievement: Implications for teacher evaluation. *Journal of Personnel Evaluation in Education, 11,* 57–67.

Authors

DALE BALLOU is Associate Professor of Public Policy and Education, Department of Leadership, Policy and Organizations, Vanderbilt University, 230 Appleton Place Nashville, TN 37203-5701; dale.ballou@vanderbilt.edu. His areas of specialization are economics of education, teacher labor markets, and assessments of schools and teachers.

WILLIAM SANDERS is Senior Research Fellow, Value-Added Research and Assessment, SAS Institute, Inc., 100 SAS Campus Drive, Cary, NC 27513-8617; Bill.Sanders@sas.com. His areas of specialization are statistical mixed linear models.

PAUL WRIGHT is Statistician, Value-Added Research and Assessment, SAS Institute Inc., 100 SAS Campus Drive, Cary, NC 27513-8617; paul.wright@sas.com. His area of specialization is linear mixed models.

Manuscript received May 2003
Revision received September 2003
Accepted December 2003