Educational Evaluation and Policy Analysis
http://eepa.aera.net

Teacher Value-Added at the High-School Level: Different Models, Different Answers?

Dan D. Goldhaber, Pete Goldschmidt and Fannie Tseng

EDUCATIONAL EVALUATION AND POLICY ANALYSIS 2013 35: 220 originally published online 16 January 2013
DOI: 10.3102/0162373712466938

The online version of this article can be found at:
http://epa.sagepub.com/content/35/2/220

Published on behalf of

American Educational Research Association

and

SAGE

http://www.sagepublications.com

Educational Evaluation and Policy Analysis
June 2013, Vol. 35, No. 2, pp. 220–236
DOI: 10.3102/0162373712466938

Teacher Value-Added at the High-School Level: Different Models, Different Answers?

Dan D. Goldhaber
University of Washington-Bothell

Pete Goldschmidt
California State University-Northridge and New Mexico Public Education

Fannie Tseng
Berkeley Policy Associates

This article reports on findings based on analyses of a unique dataset collected by ACT that includes information on student achievement in a variety of subjects at the high-school level. The authors examine the relationship between teacher effect estimates derived from value-added model (VAM) specifications employing different student learning assumptions. Specifically, they compare teacher effectiveness estimates derived from a traditional lagged score VAM using pretests and posttests in a single subject area to those derived from VAM specifications employing a cross-subject student fixed-effects approach. The latter approach offers advantages for teacher evaluation systems due to significantly reduced data requirements; however, there is evidence that both the estimated effect size of teacher quality as well as estimates of individual teacher performance vary depending on the VAM methodology. In particular, teacher effects identified based on within-student cross-subject variation results in significantly smaller effect size estimates than do those generated from a more traditional lagged score model. The correlation across model specification ranges from .25 to .96 depending on the subject area.

Keywords: *value-added, teacher effectiveness, teacher evaluation*

A CONSIDERABLE amount of research since, and including, the *Coleman Report* (Coleman, 1966) shows that teacher quality is the most important of all the school-related factors that affect student achievement (e.g., Aaronson, Barrow, & Sander, 2007; Rivkin, Hanushek, & Kain, 2005) and that teachers vary considerably from one another in effectiveness.[1] It is not terribly surprising, therefore, that policymakers are considering options that tie consequential labor market decisions (e.g., compensation, professional development, and tenure) to measures of individual teacher performance, which may be judged in part on student test performance.

The ability to use student assessment results to draw inferences about teacher performance is limited to only a portion of the teacher workforce because most of teachers do not teach in subjects or grades that are assessed with state standardized assessments. Still when possible, value-added models (VAMs) of one sort or another are a leading candidate for estimating

the contribution of individual teachers toward student achievement growth on standardized tests. The U.S. Department of Education's recent Race to the Top initiative, for instance, offered large monetary incentives for states to adopt policies focused on measuring and acting on teacher quality, including measuring teachers based on student test achievement results.[2]

Most methods of deriving teacher effectiveness or performance (we use the terms interchangeably) based on student achievement test results entail estimating gains (or controlling for a prior level of achievement in a subject area).[3] But estimating these types of models at the high school level is complicated level by the fact that high school students are often not tested annually—No Child Left Behind (NCLB) requires state tests only once over the course of high school—and usually not in a subject area that might be considered contiguous.[4] Thus, in many high-school level classes there is no clear test that measures prior achievement; a biology course, for instance, is not necessarily a good proxy for previous achievement in a chemistry course even if it is the science course that was taken in the prior school year.[5] As a consequence, new research at the secondary level (e.g., Aslam & Kingdon, 2011; Clotfelter, Ladd, & Vigdor, 2010; Dee & Cohodes, 2005; Xu, Hannaway, & Taylor, 2007) relies on an alternative methodology where across-subject (rather than across-time) variation in test performance is used to identify teacher effects.[6]

This article reports findings based on analyses of a unique dataset collected by ACT[7] that includes information on student achievement in a variety of subjects at the high-school level and, importantly, includes both explicit links between teachers and students as well as information on student achievement in each subject at the beginning and end of each school year. These data allow us to assess the impact of model specifications that rely on different theoretical assumptions about student learning on estimates of teacher effectiveness at the high-school level. Specifically, we compare teacher effect estimates derived from a traditional VAM specification that includes a pretest and posttest in a subject area to specifications that use a cross-subject student fixed-effect approach to draw conclusions about the assumptions underlying different VAMs, and we answer two specific questions:

1. To what extent do the estimated magnitudes of teacher effectiveness on student learning vary according to the VAM specification employed?
2. How correlated are teacher performance rankings generated based on different VAM specifications?

We find evidence that the estimated effect size of teacher quality and estimates of individual teacher performance are model specification dependent. For example teacher effect estimates based on a student fixed-effects specification results in significantly small smaller effect size estimates than do those generated from a more traditional lagged score model. We cannot definitively say whether one model specification is preferred over another. And whereas the policy import of the sensitivity of teacher effect estimates to model specification is a normative question, our findings strongly suggest that assumptions about what drives student achievement, and consequently VAM specification, at the high-school level warrants further research.

The remainder of the article is arranged as follows. Section II describes our data and analytic approach. Our findings are presented in Section III, and in Section IV we offer some concluding thoughts on the policy implications of the findings.

II. Data and Analytic Approach

Data

The data we use for this study were collected by ACT as part of a pilot of their QualityCore end-of-course assessments. The target population consisted of high school students in Grades 10 through 12 in public and private schools. A stratified random-sampling design was used, with explicit strata consisting of size, type (public or private), and geographical location (ACT, n.d.).[8]

Each student included in this pilot was tested in one, two, or three subject areas. The data also

include student, teacher, and school characteristics: student gender and ethnicity; the average ACT college entrance score, and class size for each school; teachers' college GPA, college major, highest degree, certification status, and years of experience.

The analysis we perform benefits from specific properties of the outcome that further enhance generalizability. The assessment outcomes are end-of-course (EOC) exams that include problem-based items embedded in contexts that are designed to be accessible and relevant to high school students. The assessments are designed to measure learning outcomes students need to attain to succeed in college; within subject area, the EOC exams were scaled to have constant meaning across forms (ACT, n.d.). That is, pretest and posttest forms by subject were considered parallel forms and, hence, inferences related to what students know and can do, based on student scores, are the same from one form to another. The reliabilities of the EOCs range from a low of .78 (geometry) to .94 (English 11; ACT, n.d.), which are moderate to high but acceptable (DeVellis, 2003) and imply correlations with true performance of .88 to .97 (Yen, 1986).

Given the unique situation that parallel forms were given to students at the start and end of the course, scale scores retained constant meaning, meeting an important requirement for analyzing changes in student performance (Raudenbush, 2001). The scale scores, however, do not have constant meaning across subject. We standardize these scaled scores for all students within each subject area to be ~N(0,1).[9]

An important aspect of the outcome is that it is a low-stakes assessment. As such, it is more likely that estimated student gains may be attributed to desired teacher behaviors as opposed to, for instance, teaching specifically to a high-stakes test (Koretz, 2002, 2005). On the other hand, one might worry that students do not really try to do well on these low-stakes tests, so they do not provide a very good measure of their true learning gains (Koretz, 2008).[10] Another potential problem is that pretest scores in what may be unfamiliar subjects to students, such as biology and chemistry, may by uninformative predictors of posttest scores. If students begin the year with little baseline knowledge in those subjects, then including the baseline test does little to control for initial student ability. However, our findings suggest this is not the case.[11]

We restrict our analysis dataset to schools and classrooms in which teachers and students can be uniquely linked, which includes a total of 23 schools (9 of which are private schools), 205 teachers, and 8,002 students (in Grades 9 through 12).[12] In Table 1, we report selected school and teacher characteristics and include a comparison of the poverty level of these with information on high schools in the Midwest (to be comparable with our sample, which his also from the Midwest) from the 2007–2008 Common Core of Data and teacher characteristics from the 2007–2008 wave of the Schools and Staffing Survey. In terms of poverty level (percentage receiving free or reduced price lunch), the public schools in our sample appear to be very similar to those in the Midwest as a whole, though slightly less disadvantaged.[13] The teachers in our sample are less likely to hold an EdD or PhD and are slightly more experienced than other teachers in the U.S., but these differences are relatively small.[14]

We might expect to see a high correlation of test scores across subjects within students; that is, a student who performs well in one subject area is likely to also perform well in others. This is in fact what one would hope to see if it turns out that one subject area serves as a good control for achievement in other subjects (as we describe in the next subsection, this is an assumption of the VAM specification that might be used at the high-school level where pretests are typically not available). The evidence presented in Panel A of Table 2 shows some evidence of this in that there are generally positive correlations of posttest scores across subjects. It is interesting, for instance, to see that English test achievement is often highly correlated with both science and math achievement. The correlations, however, vary widely, ranging from .01 to .92. Some of the low correlations likely result from the fact that sample sizes are quite small, as students are quite unlikely to be tested across some combinations of subjects in the same year; in the case of algebra and geometry, for instance, there are only 14 students who are tested in those two subject areas.

TABLE 1
Selected School and Teacher Sample Characteristics

School Characteristics	Public High School Sample	All Public High Schools in Midwest[a]
Percentage of students eligible for free/reduced price lunch		
0–25	36%	32%
26–50	43%	35%
51–75	14%	11%
76–100	0%	4%
Unknown	7%	17%
Teacher Characteristics	**Teacher Sample (Includes Public and Private Secondary School Teachers)**	**Public and Private Secondary School Teachers in United States (2007–2008)[b]**
Less than a BA	0%	2%
BA/BS	51%	44%
MA/MS	46%	46%
EdD/PhD	2%	8%
Fewer than 4 years of teaching experience	18%	19%
4–9 years of teaching experience	29%	28%
10 or more years of teaching experience	53%	53%

a. Common Core of Data (2007–2008).
b. Schools and Staffing Survey (2007–2008a, 2007–2008b).

TABLE 2
Panel A. Subject Test Scores Correlations Among Students With Multiple Tests

Subject Test Score Correlation	Algebra I	Algebra II	Biology	Chemistry	10th-Grade English	11th-Grade English	Geometry
Algebra I	1						
Algebra II	n/a *p* value = n/a *N* = 2	1 *N* = 2					
Biology	0.540 *p* value = .000 *N* = 406	0.560 *p* value = .000 *N* = 96	1 *N* = 502				
Chemistry	0.344 *p* value = .092 *N* = 25	0.619 *p* value = .000 *N* = 512	0.924 *p* value = .008 *N* = 6	1 *N* = 1,626			
10th-grade English	0.254 *p* value = .001 *N* = 160	0.609 *p* value = .000 *N* = 247	0.649 *p* value = .000 *N* = 738	0.503 *p* value = .000 *N* = 296	1 *N* = 1,441		
11th-grade English	0.543 *p* value = .007 *N* = 23	0.504 *p* value = .000 *N* = 613	0.600 *p* value = .000 *N* = 624	0.610 *p* value = .000 *N* = 624	0.301 *p* value = .297 *N* = 14	1 *N* = 1,711	
Geometry	0.008 *p* value = .977 *N* = 14	0.359 *p* value = .051 *N* = 30	0.587 *p* value = .000 *N* = 643	0.575 *p* value = .000 *N* = 233	0.460 *p* value = .000 *N* = 833	0.225 *p* value = .002 *N* = 183	1 *N* = 1,963

Note. n/a refers to subject tests that were taken by fewer than three students.

TABLE 2 (CONTINUED)

Panel B. Subject Pretest Scores Means and Standard Deviations by the Number of Tests Taken

	Students With 1 Subject Test	Students With 2 Subject Tests	Students With 3 Subject Tests	*F* Test	Prob < *F*	*N*
Algebra I	143.41 (3.11)	143.28 (3.29)	143.43 (2.66)	0.28	0.76	1,426
Algebra II	142.92 (3.43)	143.65 (3.32)	143.96 (3.47)	10.11	0.00	1,453
Biology	145.22 (4.56)	145.89 (4.44)	147.34 (4.82)	32.80	0.00	1,840
Chemistry	142.31 (3.10)	142.81 (3.22)	142.02 (2.94)	8.73	0.00	1,626
10th-grade English	152.83 (7.12)	153.18 (6.30)	153.86 (5.76)	3.88	0.02	1,804
11th-grade English	153.88 (6.91)	152.44 (6.69)	153.06 (6.27)	7.30	0.00	1,711
Geometry	143.07 (3.08)	142.55 (2.99)	143.03 (3.03)	4.86	0.01	1,798

Students were not uniformly tested across the same subjects, which raises the prospect that students in the sample who have more subject area assessment results may be systematically different from those with fewer assessments.[15] We assess this in Panel B of Table 2, which shows the average test scores in seven subjects (Algebra I, Algebra II, Geometry, Biology, Chemistry, and 10th- and 11th-grade English) for students who sat for one, two, or three or more subject tests.[16] There are some slight, and often statistically significant, differences between the average scores of students in the data with only one subject area test and those with more than one; however, there is not a generalized pattern in the means in that the mean scores for students with multiple tests are higher than the means for students with only one test in some subjects and lower in others.

Analytic Approach

Since the 1966 *Coleman Report* (Coleman, 1966), there have been literally hundreds of studies that analyze the relationship between educational inputs and student achievement on standardized tests, and, in general, the findings show only a weak relationship between teacher credentials or characteristics (or other schooling resources) and achievement (Goldhaber, 2002; Hanushek, 1986, 1997).[17] Estimation of teacher effects at the high-school level is complicated because of the lack of standardized student assessment results. NCLB testing requirements have considerably broadened the potential for using VAMs in elementary- and middle-school levels. However, NCLB only requires one high-school grade be tested so one cannot use what has become a standard value-added framework (Hanushek, 1979) that entails regressing student achievement on a test in one grade against a set of school or teacher variables and individual student covariates including achievement in a prior grade.

More recently, Clotfelter et al. (2010) and Xu et al. (2007) have addressed this limitation by employing a student fixed-effects model on an administrative dataset from North Carolina to examine the relationship between teacher characteristics and credentials and student achievement at the high-school level. Because the North Carolina state administrative data lack a clear-cut baseline test for prior subject-specific student achievement, both sets of authors model within-student variation across subjects based on EOC exams. Clotfelter et al. found that teachers' credentials (in-subject test scores) have positive effects on students, particularly in math. Xu et al. assessed the differential impacts of Teach for America (TFA) and traditional classroom teachers and find TFA teachers to be more effective.[18]

The data described in the prior subsection allow for an exploration of the assumptions underlying the type of value-added model employed by Clotfelter et al. (2010) and Xu et al. (2007). In particular, when same-subject pretest scores are not available, as is typical at the high-school level, one can instead exploit within-student cross-subject variation to derive individual teacher value-added effectiveness estimates.

Consider the following generalized version of a typical value-added model:

$$A^S_{ijt} - A^S_{ij(t-1)} = \mu^G_i + \mu^S_i + X_{it}\gamma + C_t\beta + T^S_{jt}\tau + \varepsilon^S_{ijt}, \quad (1)$$

where *i* represents students, *j* represents teachers, *S* represents subject area, and *t* represents the

school year. Student learning gains, $A^S_{ijt} - A^S_{ij(t-1)}$, in a particular subject and in a particular teacher's class are a function of generalized student ability, μ^G_i, which is common across all subject areas, subject-specific student knowledge, μ^S_i, which only affects achievement in subject *S*, time-varying student and family background characteristics, X_{it}, classroom or school covariates, C_t, the student's teacher in subject *S* in year *t*, T^S_{jt}, as well as a disturbance which is orthogonal to the previous terms.

Data availability and the structure of the school setting play an important role in the VAM specification that is ultimately used by researchers. When estimating value-added at the elementary level, where student test histories in a subject are available, researchers typically use some variant of the following model:

$$A^S_{ijt} = \alpha A^S_{ij(t-1)} + X_{it}\gamma + C_t\beta + T^S_{jt}\tau + \varepsilon^S_{ijt}, \quad (2)$$

where $\alpha A^S_{ij(t-1)}$ serves as a proxy for learning prior to time *t* and unobserved ability (i.e., the assumption here is that $\alpha A^S_{ij(t-1)}$ serves as a proxy for both μ^G_i, and μ^S_i,).[19]

When repeated measures of student performance in a single subject are unavailable (so there is no measure of prior year student achievement), but there are tests available across multiple subjects, one can instead estimate a cross-subject student fixed-effect variant of Equation 1:

$$A^S_{ijt} = \mu^G_i + X_{it}\gamma + C_t\beta + T^S_{jt}\tau + \varepsilon^S_{ijt}. \quad (3)$$

The important distinction between Equations 2 and 3 is that the measure of prior achievement has been replaced by a measure of general student ability, μ^G_i, which is assumed to be constant across all subject areas, and $\mu^S_i = 0$ is assumed for all subjects, referred to here as the "blank slate assumption."[20] In Equation 3, teacher effects are identified based on within-student variation in achievement across subjects as opposed to within-student variation in achievement over time. Demeaning both sides of Equation 3 within students yields

$$A^S_{ijt} - \bar{A}_{ijt} = \left(T^S_{jt} - \bar{T}_{jt}\right)\tau + \left(\varepsilon^S_{ijt} - \bar{\varepsilon}_{ijt}\right). \quad (4)$$

There are two key assumptions implicit in the typical high school specification represented in Equations 3 and 4. The first, arising from the fact that there is no pretest score on the right hand side of the model, is that students start a subject with a blank slate, controlling for general ability.[21] In other words, the blank slate assumption inherent in Equation 4 assumes that there are no differences between students in knowledge of subject *S* at time *t-1* that influence their achievement in that subject at time *t*. The second, which is also crucial to the derivation of Equation 4, is that student achievement at the high-school level is unidimensional in the sense that the student-related factors (e.g., general ability) influencing achievement in one subject area have the same effect in other subjects as well. This implies that in Equation 1, $\mu^S_i = 0$ for all subjects and thus μ^G_i, drops out after demeaning the variables in Equation 3. Were this not the case, then the within-student relative subject-specific ability term, $\mu^S_i - \bar{\mu}^S_i$, would be subsumed into the error term and may bias the teacher-effect estimates because the adjusted gain across subject areas could not be attributed strictly to the educational resources at the school.[22]

Research on the role of general student ability across content is not conclusive, but evidence suggests that people who do well in one area tend to do well in others (Carroll, 1993). This, however, does not guarantee that achievement is unidimensional. In fact, Heckman (1995) argues that both general ability and other factors play a role in affecting achievement.[23]

Clotfelter et al. (2010) examined the issue of unidimensionality of student ability in two ways. First they found that the probability of students' enrollment in advanced algebra and English courses is strongly predicted by average absolute ability in math and reading (as measured by a test in a prior year) but is unrelated to relative ability in those subjects—evidenced by the fact that absolute scores in both math and reading are positively correlated with enrollment in both advanced algebra and English and that neither the relative math nor reading score appears to be a better predictor of enrollment in an advanced placement course. Second, they note the result from an ordinary least squares regression that relative student scores (from the eighth grade) are not a significant predictor of relative teacher licensure scores (for future high

school teachers), a finding that they argue is consistent with the assumption that students are not assigned to teachers based on their relative math and reading abilities. They conclude that schools consider student ability to be single dimensional. Clotfelter et al.'s conclusion is reached in the context of student achievement models based on observable teacher characteristics. If most of a teacher's effectiveness derives from unobservable characteristics that are relatively uncorrelated with observable characteristics and if students are systematically assigned to teachers based on relative ability, then a violation of unidimensionality implies that student fixed-effects value-added models may, in fact, be biased.

Where our study departs from previous research is that we are able to more rigorously test the blank slate and unidimensionality assumptions and to judge the effects of possible violations of these assumptions on estimated teacher effects. We can do this because the data described above include student test results across a variety of subjects at two points in time (at the beginning and end of the school year). Specifically, we estimate models that expand on the Clotfelter et al. (2010) and Xu et al. (2007) frameworks by adding within-student demeaned pretest scores to the right-hand side of Equation 4:

$$A_{ijt}^{S} - \bar{A}_{ijt} = \left(A_{ij(t-1)}^{S} - \bar{A}_{ij(t-1)}\right)\alpha + \left(T_{jt}^{S} - \bar{T}_{jt}\right)\tau + \left(\varepsilon_{ijt}^{S} - \bar{\varepsilon}_{ijt}\right). \quad (5)$$

We term this model the "comprehensive model" in discussion of the results.[24]

We can empirically test the effect of different assumptions about student learning (cumulative subject-specific achievement, as in Equation 2, or the unidimensionality and blank-slate assumptions represented in Equation 3) on teacher effectiveness estimates from VAMs. In particular, in Equation 2, the specification that can generally be used at the high-school level assumes that $\alpha = 0$. Consider the implication of a finding that $\alpha \neq 0$ in estimating the comprehensive model. One explanation for the finding is that unidimensionality holds (i.e., $\mu_i^S = 0$) but that the blank-slate assumption does not (i.e., students come into a high-school subject with knowledge that is not accounted for by μ_i^G). On the other hand, suppose that the blank-slate assumption does hold but that unidimensionality does not (i.e., $\mu_i^S \neq 0$). In this case, the coefficient on the lagged relative score picks up the effect of subject-specific ability in the same way that a coefficient on the lagged level score picks up the effect of overall student ability in the traditional VAM formulation. Of course, rejecting $\alpha \neq 0$ can result from the both blank-slate and the unidimensionality assumptions' not being tenable.

To the degree that the above assumptions do not hold, one might expect differences in the estimates of value-added, which is potentially relevant to policy. Personnel policies that use a measure of value-added, for instance, are more likely to rely on a teacher's rank in the distribution rather than the magnitude of the estimated value-added effects. Thus, differences in estimated effects between models may not be important if they do not have much impact on the value-added ranking of teachers. We can assess this by examining the rank correlations between the γ estimates from these different models. And if the estimates from these different models are highly correlated, we would conclude that lagged achievement is a good proxy for general student ability and that subject-specific ability is negligible or at least is uncorrelated with teacher assignment.

We might also expect the different models to generate varying estimates of the teacher effect size. In particular, the estimates from the student fixed-effects models are identified based on within- versus between-student variation in achievement. If there is positive (unobserved) selection to student ability to teacher ability, then the within-student models would tend to show less variation in teacher effectiveness because we only observe students with a narrow range of teacher effectiveness. The reverse, of course, is also true.

In calculating effect sizes, we are concerned about sampling error inflating the estimates. There are several ways to deal with this issue.[25] For instance, let $\hat{\tau}_j$ be the raw teacher effect estimate for teacher j, and let $\mathrm{SE}(\hat{\tau}_j)$ be the standard error for this estimate. Furthermore, let V_{TRUE} be the squared effect size corrected for sampling error (what we want to estimate), and let V_{ERROR} be the portion of the total variance of the teacher effects that is due to sampling

error. If we assume that the teacher effects estimates are uncorrelated with the sampling error, we can express the total variance of the raw teacher effects, V_{TOTAL}, as

$$V_{TOTAL} = V_{ERROR} + V_{TRUE}. \quad (6)$$

A natural estimator for the first term in Equation 6 is the sample variance of the raw teacher effect estimates, $\hat{\tau}_1,...,\hat{\tau}_n$, while an estimate of the second term in Equation 6 is the average of the squared standard errors of these estimates, weighted by the number of observations, k_j, contributing to each estimate:

$$\hat{V}_{TOTAL} = \mathrm{Var}(\hat{\tau}_1,...,\hat{\tau}_n) \quad (7)$$

$$\hat{V}_{ERROR} = \frac{1}{\sum k_j}\sum [\mathrm{SE}(\hat{\tau}_j)]^2 k_j. \quad (8)$$

From Equation 6, one approach to estimate V_{TRUE} is to take subtract $\hat{V}_{TOTAL}$ and $\hat{V}_{ERROR}$:

$$\hat{V}_{TRUE(1)} = \hat{V}_{TOTAL} - \hat{V}_{ERROR}. \quad (9)$$

We can also shrink the individual teacher estimates before calculating the adjusted effect size. We calculate two types of individual Empirical Bayes (EB)–adjusted teacher effect estimates: $\hat{\tau}_j^{EB(1)}$, which uses the same shrinkage factor for all teachers; and $\hat{\tau}_j^{EB(2)}$, which uses a shrinkage factor that is informed by the reliability of the individual estimate (i.e. inversely proportional to the sampling error for that estimate):

$$\hat{\tau}_j^{EB(1)} = \frac{\hat{V}_{TRUE}}{\hat{V}_{TOTAL}}\hat{\tau}_j \quad (10)$$

$$\hat{\tau}_j^{EB(2)} = \bar{\tau} + \frac{\hat{V}_{TRUE}}{\hat{V}_{TRUE} + [\mathrm{SE}(\hat{\tau}_j)]^2}(\hat{\tau}_j - \bar{\tau}). \quad (11)$$

The final two estimates of V_{TRUE}, then, are simply the sample variances of each type of EB-adjusted estimates:

$$\hat{V}_{TRUE(2)} = \mathrm{Var}(\hat{\tau}_1^{EB(1)},...,\hat{\tau}_n^{EB(1)}) \quad (12)$$

$$\hat{V}_{TRUE(3)} = \mathrm{Var}(\hat{\tau}_1^{EB(2)},...,\hat{\tau}_n^{EB(2)}). \quad (13)$$

Below we report the EB estimates that shrink teacher effects in proportion to their reliability since the individual teacher class sizes in the sample vary; however, as it turns out in practice, the EB shrinkage methodology makes little differences to our findings.[26]

III. Results

There are significantly more studies at the elementary level reporting the impact of teacher quality on student achievement than at the middle- or high-school levels. For example, of the 18 studies on the effect size of teacher quality in a recent review by Nye, Konstantopoulos, and Hedges (2004), only 4 include exclusively sixth grade or higher, and only 1 includes grades higher than ninth grade. However, the estimates of teacher effects at the middle- and high-school levels are consistent with the elementary literature in showing teacher quality can have a large impact on student achievement relative to other schooling inputs.[27]

We use the model specifications outlined in Equations 2, 4, and 5 to estimate the impact of a 1 standard deviation impact in teacher effectiveness on student achievement.[28] Equation 2 is consistent with the basic value-added formulation that includes a same-subject student pretest as a control variable. There is no pretest score on the right hand side of Equation 4; rather, this specification attempts to control for heterogeneity of individual ability by including a student fixed effect in the model, and it assumes that: (a) there are no differences between students in initial knowledge of subjects; and (b) student achievement at the high-school level is unidimensional in the sense that the student-related factors that influence achievement in one subject area have the same effect in other subjects as well (and there is no selection into classrooms based on students' subject-specific ability). Finally, Equation 5 augments the control variables with within-student demeaned pretest scores. A finding of significance on the coefficient of the within-student demeaned pretest indicates that one or both of the key assumptions (blank slate or unidimensionality) does not hold in the case of the type of VAM used when a pretest scores are unavailable.

In Table 3, we show the estimated teacher effect size. In each cell of the table, we report unadjusted effect size estimates as well as effect size estimates that are corrected for sampling

TABLE 3

Standard Deviation of Teacher Effect Estimates Under Different Model Specifications (unadjusted standard deviation/Empirical Bayes–adjusted standard deviation)

	Number of Teachers	(1) Traditional Lagged Score Model	(2) Student Fixed-Effects Model	(3) Student Fixed-Effects With Lagged Score Model
Algebra I teachers	22	0.405/0.334	0.214/0.156	0.182/0.125
Algebra II teachers	36	0.387/0.317	0.222/0.169	0.185/0.126
Biology teachers	31	0.403/0.264	0.190/0.062	0.164/0.034
Chemistry teachers	26	0.442/0.281	0.197/0.035	0.163/NA*
10th-grade English teachers	25	0.258/0.025	0.186/0.034	0.156/NA*
11th-grade English teachers	34	0.426/0.294	0.206/0.086	0.180/0.055
Geometry teachers	38	0.420/0.350	0.191/0.135	0.115/0.133

*We cannot calculate adjusted effect sizes for the student fixed effects VAMs because the estimate of the error variance (the weighted average of the standard errors of the teacher effects) is greater than the total variance of the estimated teacher effects.

error using EB methods. In estimating the effect size (and throughout the analysis), we keep the sample of students informing each teacher's performance estimate constant so differences that arise in the results are driven by model specification, not the sample itself.[29] The effect sizes vary by subject area, but our findings also suggest that they differ by methodological approach.[30] Specifically, Column 1 of the table shows the results when we use the traditional lagged score approach, Column 2 shows the results from a specification with student fixed effects (consistent with Clotfelter et al., 2010, and Xu et al., 2007), while Column 3 reports results from our comprehensive model (Equation 5).

The unadjusted estimates range from about 0.16 to 0.44 across all subjects. The EB-adjusted estimates are often significantly smaller (in cases where there was considerable measurement error), ranging from 0.03 to 0.35. These estimates are not out of line with typical estimates at the elementary level. For example, in a recent review of published students on teacher effects, Hanushek and Rivken (2010) reported effect sizes, adjusted for measurement error, that range from 0.08 to 0.26 using reading tests and 0.11 to 0.36 in math.

What is more notable is that there is a substantial difference in the estimate impact of teachers across model specifications. The changes in estimated teacher quality from the traditional model (Column 1) is consistently far larger than either of the two student fixed-effects models (Columns 2 and 3); it is about 1.5 to 2 times as large as the point estimates derived from the student fixed-effects specification (Column 2), and even larger relative to the comprehensive (Column 3) model. The same pattern tends to be present for the EB estimates as well. These smaller effect sizes in the student fixed effects model specification are consistent with the notion that students are positively matched to teachers across subjects so the effect sizes are identified based on a relatively narrower range of teacher performance than is the case for the cross-sectional traditional value-added specification.[31]

The estimate of α, the coefficient on the within-student demeaned pretest score variable, is positive ($\alpha = 0.31$) and highly significant. Recall that the inclusion of this variable in the comprehensive model is designed to test the degree to which the blank slate and/or unidimensionality assumptions implicit in the student fixed-effects models without demeaned lagged achievement (see Equation 4) appear to hold. That α is significant indicates that one of the assumptions underlying this specification is violated, but, unfortunately, we cannot tell whether it is one or both assumptions that do not hold.

The results for the comprehensive model in Column 3 of Table 3 show that accounting for nonrandom student teacher matching based on subject specific ability by including the demeaned pretest score in the regression may reduce upward bias in estimates of teacher effectiveness in the student fixed-effect model that does not include demeaned lag scores. All the unadjusted (and most of the EB-adjusted) standard deviations in Column 3 are all less than their counterparts in Column 2. However, as noted above, both fixed effects specifications may capture only a narrower range of teacher

TABLE 4
Spearman Correlations of Teacher Effect Estimates Across Model Specifications

	(1) Traditional and Student Fixed-Effects Models	(2) Student Fixed-Effects and Fixed-Effects With Lagged Score Models	(3) Traditional and Student Fixed-Effects With Lagged Score Models
Algebra I teachers	0.802**	0.941**	0.851**
Algebra II teachers	0.429**	0.957**	0.453**
Biology teachers	0.249	0.917**	0.346
Chemistry teachers	0.408*	0.952**	0.538**
10th-grade English teachers	0.293	0.921**	0.245
11th-grade English teachers	0.674**	0.925**	0.529**
Geometry teachers	0.464**	0.901**	0.677**

*p = .05. **p = .01.

performance. It is probably appropriate to think of the comprehensive model (student fixed effects specification with demeaned lags) as a lower bound estimate of the overall variation of teacher effectiveness since it likely is estimated based on restricted range of teacher effects and includes controls for the possibility that student assignments to teachers could be based on relative ability in a subject (i.e., it allows for violations of the blank slate and unidimensionality assumptions).

In Table 4, we present correlations between individual teacher effect estimates from the three models.[32] Column 2 of Table 4 shows that individual teacher effect estimates from the student fixed-effects models and comprehensive models are highly correlated (all over 0.9), and Columns 1 and 3 show that estimates from the traditional model are far less correlated with those from either of the other two models; the correlations range from a low of 0.25 for English 10 to about 0.85 for Algebra I.[33]

That the correlations between the two student fixed effects specifications are far higher than the correlation between the models that include lag scores suggests that the within-student identification strategy has a much larger impact on estimated teacher effects than does the inclusion of controls for prior-year achievement. This suggests greater sensitivity to model specification than other similar analysis at the elementary- (Papay, 2011) and middle-school (Ballou, Sanders, & Wright, 2004; Lockwood et al., 2007) levels, though these studies do not estimate analogous models to the within-student cross-subject fixed-effects models.[34] Interestingly, however, the magnitudes at the low end (e.g., 0.25 for biology and 0.29 for 10th-grade English teachers) are not out of line with what has been reported for teacher effectiveness correlations in other contexts: Goldhaber and Hansen (in press), for instance, reported adjacent year correlations in teacher reading effectiveness (at the elementary level) of about 0.30; Papay (2011) found correlations across different reading tests that are in the range of 0.16 to 0.58 depending on the test and model specification.

In short, our results show that model specification affects both the estimated teacher effect size as well as individual estimates of teacher effectiveness. Clearly, the choice of how to control for student ability is a nontrivial one. We explore the extent to which the differences in teacher-effect estimates are relevant to policy in greater detail in Section V.

IV. Public Policy Implications and Conclusions

In the previous sections, we showed that model specification has a large influence on both the estimated impact of teacher quality on student achievement and estimates of individual teacher effectiveness. To help illustrate the policy relevance of the correlations we report in Table 4, in Table 5 we present a transition matrix that shows the percentage of teachers who fall into a quintile given one specification of the model and the same or a different quintile

TABLE 5
Transition Tables

	Student Fixed-Effects Model				
Panel A	1	2	3	4	5
Traditional model					
1	38.0%	22.0%	24.0%	16.0%	0.0%
2	26.1%	28.3%	15.2%	19.6%	10.9%
3	20.0%	20.0%	20.0%	24.4%	15.6%
4	13.0%	23.9%	26.1%	13.0%	23.9%
5	9.3%	4.7%	11.6%	27.9%	46.5%
Panel B	Student Fixed-Effects With Lagged Score Model				
Student fixed-effects model					
1	82.0%	18.0%	0.0%	0.0%	0.0%
2	17.4%	47.8%	32.6%	2.2%	0.0%
3	2.2%	26.7%	48.9%	20.0%	2.2%
4	0.0%	6.5%	17.4%	63.0%	13.0%
5	0.0%	0.0%	0.0%	16.3%	83.7%

using an alternative specification. Quintile 1 represents the highest ranked teachers and Quintile 5 represents the lowest ranked teachers.

Were it the case that model specification was irrelevant to a teacher's rank in the distribution (and measurement error was not an issue), we would expect the diagonal elements of the matrix to each be 100% and the off-diagonal elements to be zero. This, clearly, is not the case. There is a considerable number of teachers who end up switching quintiles based solely on model specification; this is especially pronounced in Panel A, when we move from the traditional value-added model to the student fixed-effects model. For instance, we observe (Table 5, Panel A) that about 9% of teachers who fall into the lowest quintile in the traditional model are found to be in the highest quintile in the student fixed-effects model (extreme bottom-left cell). The correlation between estimates from the student fixed-effects and comprehensive models is higher, as is shown in the corresponding transition table in Panel B. In this case, a significantly larger proportion of teachers fall in cells close to the diagonal.

Some suggest that value-added measures are not reliable enough to use; Hill (2009), for instance, argued,

> It seems irresponsible, given what we know about value-added scores, to use them in high-stakes situations absent other information about teacher quality. Even lower-stakes situations, such as singling out teachers for extra professional development or peer mentoring, can induce substantial psychological costs. (p. 706)

But, whether value-added estimates ought to be used is debateable. There is little evidence that non-VAM teacher evaluation methods are rigorous and there is evidence that value-added estimates of teacher effectiveness are better predictors of future student achievement than are other teacher credentials typically used for employment determination and compensation in the teacher labor market (Glazerman et al., 2010; Goldhaber & Hansen, 2010).

Should the sensitivity of estimated teacher performance to model specification make policymakers wary about using value-added methods at the high-school level (a cross-subject student fixed effects approach in particular) in the evaluation of individual teachers? There is no empirically derived right answer to this question. The bottom line for teacher evaluation of any sort is that there will be some level of teacher misclassification and misclassification associated with VAM estimates must be juxtaposed against teacher evaluation systems that, today at least, look to be less than rigorous (Toch & Rothman, 2008). Clearly, there is some risk associated with using value-added models that provide poor information about true teacher effectiveness, but it is also true that one cannot use teacher performance to inform personnel decisions if the performance evaluation systems

in place fail to distinguish teachers from one another (Weisburg, Sexton, Mulhern, & Keeling, 2009). Of course we probably care more about whether the use of teacher performance estimates has causal impacts on the quality of the teacher workforce than the accuracy of the performance measure.[35] And this is not something that can be convincingly determined outside of actual policy variation because the behavioral response of teachers to the use of value-added estimates may affect their performance.

Our study shows that value-added measures can differ according to the model and connects these differences to assumptions about student learning. The limitation is that it does not provide information about what model specification is likely to be the most accurate reflection of true teacher effectiveness. Ultimately it is necessary to have an external means of assessing the validity of different VAM specifications. For instance, the recent study by Chetty, Friedman, and Rockoff (2011) assesses whether value-added measures at the elementary- and middle-school levels are meaningful predictors of later life outcomes, such as college-going behavior and earnings. This type of research is also promising for estimates of value-added at the high-school level. Another promising direction for future research is to study VAM estimates in an experimental setting where students are randomly matched to teachers, such as has been done at the elementary level by Kane and Staiger (2008).

Acknowledgments

We wish to acknowledge ACT for supplying the data used in the analyses described here; Jim Sconing, Zeyu Xu, and James Cowen for helpful comments; and Philip Sylling for research assistance. All errors in the article are solely the authors' responsibility and the views expressed do not necessarily represent ACT or the authors' institutions.

Notes

1. Estimates of the effect size impact of *teacher quality* (a term used interchangeably here with *teacher effectiveness* and *teacher job performance*) range roughly from a low-end estimate of .10 (Rivkin, Hanushek, & Kain, 2005; Rockoff, 2004) to a high-end estimate of about .50 (Nye, Konstantopoulos, & Hedges, 2004). But even the lower bound estimates of the effect size suggest teacher quality is quite important relative to other schooling interventions. For example, the low-end teacher effect estimate of .10 is on the same order of magnitude as lowering class size by 10 to 13 students (Rivkin et al., 2005).

2. For more information on Race to the Top, see http://www2.ed.gov/programs/racetothetop/index.html.

3. See, for instance, Rivkin (2009), Rothstein (2010), and Todd and Wolpin (2003), for a discussion of the assumptions underlying traditional value-added models (VAMs).

4. Additionally, many of the datasets used to assess the contribution of individual teachers, and thus the effect size of changes in teacher quality, use methods to infer linkages of individual teachers and students, and these methods often do not apply at the secondary level; for example, the proctor of a standardized test is a good proxy for an elementary student's teacher, but is more likely the homeroom teacher for a high school student (Xu, Hannaway, & Taylor, 2007).

5. At the elementary level, one might attribute the gain in student achievement from the end of the fourth grade to the end of the fifth grade as due in part to the fifth-grade teacher because elementary coursework follows a linear progression and the annual testing effectively provides pretests and posttests.

6. In other words, this work measures gains in the distribution of achievement in one subject, controlling for a student's position in the distribution in one or more alternate subjects.

7. ACT was formerly known as the American College Testing Program.

8. Schools with fewer than 100 students were excluded (ACT, n.d.).

9. Using a normalized metric in the model does not affect inferences regarding relative performance (Goldschmidt, Choi, Martinez, & Novak, 2010).

10. About 23% of student gains in our sample are negative. However, this may not result in biased estimates of the importance of teacher quality if student scores remain roughly normally distributed, which we find to be the case. We further eliminate the extreme 1% and then the extreme 2% of student gains and find that the estimated variance of teacher quality is indeed smaller, but the relative differences across subjects and model specifications are qualitatively similar.

11. Specifically, the correlations between pretests and posttests ranged from .56 to .77, which is similar to state standards based assessments that range from about .70 to .85 (in adjacent years).

12. Beginning- and end-of-course assessment scores were originally collected from classrooms in

62 schools located in the Midwest, but in many of these schools teachers and students are not linked. There were 52 teacher-class observations with fewer than 8 students; these were dropped from the sample.

13. Students in private schools are eligible to receive free or reduced price lunches; however, our dataset did not include information about free or reduced-price lunch eligibility for the private schools.

14. We could not find comparable data for the Midwest teachers only.

15. For example, one might hypothesize that more motivated students would be more likely to complete assessments in more subject areas.

16. The data also include a small number of students who are tested in 12th-grade English, but the student test sample is too small to include in the analyses.

17. See, for instance, Clotfelter, Ladd, and Vigdor (2007); Goldhaber and Anthony (2007); Goldhaber and Brewer (1997); Monk and King (1994); Rockoff (2004); and Rowan, Chiang, and Miller (1997) as examples.

18. A limitation of this approach is that end-of-course (EOC) tests may not be well aligned with assessments taken in a prior year because many states have moved toward specific EOC assessments that focus narrowly on material covered by a particular course (Yen, 1986). As a consequence, models using prior EOC assessments as controls might not meet a key assumption that the outcome has constant meaning over time (Raudenbush, 2001), and it is not based on a vertically equated item response theory–based scale score that is on an interval scale and is comparable across grades. Theoretically, this is the optimal metric to use when examining change in student performance (Hambleton & Swaminathan, 1987). We describe the specific characteristics and unique implementation of the EOCs we use in this analysis.

19. If three or more observations of student achievement are available, then researchers often replace time-invariant variables in X_{it} with a student fixed effect, μ_i (e.g., see Harris & Sass, 2011), which account for both time-invariant observed and unobservable student factors influencing achievement. Note, however, that this would not account for dynamic factors that are unobserved (Rothstein, 2010). An alternative specification (based on different assumptions) treats μ_i as a random effect, which allows for the estimation of time invariant X_{it} (McCaffrey, Lockwood, Koretz, & Hamilton, 2004).

20. This is akin to assuming that there is no persistence in teacher effects from prior years. Previous research generally focusing on elementary grades indicates prior teacher effects fade out over time, with estimates of the teacher persistence parameter ranging from about .2 to .3 (Jacob, Lefgren, & Sims, 2008; Konstantopoulos & Chung, 2011; Lockwood & McCaffrey, 2007; McCaffrey et al., 2004).

21. Models that include a measure of prior student learning are thought to account for schooling inputs in prior periods because these would have been incorporated into the prior year test achievement score in the model. See Todd and Wolpin (2003) for a more in-depth discussion of the issue of decay in student achievement.

22. Imagine, for example, that students who excel in math tend to be assigned to math teachers who hold master's degrees and, further, that these students tend to perform poorly in English. In this case, the use of average student achievement across all subjects as a control variable would tend to overstate the effect of having a math teacher with a master's degree because the model would underpredict a student's true ability in math and, therefore, attribute the relatively good performance to having a teacher with a master's degree.

23. Recent evidence suggests that general ability directly affects specific broad cognitive abilities, thus indirectly playing a substantial role in (mathematics) achievement (Taub, Floyd, Keith, & McGrew, 2008), and cross-subject assessment results have been included in achievement models specifically to account for general ability (Goldschmidt, Martinez-Fernandez, Niemi, & Baker, 2007). Teacher effect estimates have also been found to be sensitive to the characteristics of the outcome measures used to estimate them (Lockwood et al., 2007). Lockwood et al. (2007) found that teacher effect estimates had correlations of about .01 to .46 when based on the same students' scores on two subsets of the same mathematics test. It is not clear whether these differences are due to nonunidimensionality of student ability, content coverage, or nonunidimensionality in teacher quality.

24. Note that typical value-added models that include student fixed effects and a lagged score would be biased because the demeaned lagged achievement variable includes contemporaneous achievement (i.e., the dependent variable). The solution for this is typically to instrument for the lagged variable. This is not a problem in estimating Model 5 because the lagged demeaned variable does not included contemporaneous achievement (given that the demeaning is based on a cross-subject mean rather than a mean over time). Also, this model does still assume that subject-specific ability or knowledge must decay at the same rate as do other inputs (see Todd & Wolpin, 2003).

25. See, for instance, Aaronson, Barrow, and Sander (2007), Jacob and Lefgren (2008), and Kane and Staiger (2008). For a more extensive discussion of the assumptions underlying different types of

Empirical Bayes (EB) corrections, see Goldhaber and Hansen (in press).

26. The estimated effects sizes were generally found to be slightly smaller when shirking based on the individual reliability of the estimate than shrinking by the same factor for all teachers, but the differential in estimated effect sizes between the two EB estimates is generally in the thousands decimal place. We also estimated models that excluded teachers in classes with less than 12 students; this sample also produced very similar findings for teacher effect sizes.

27. For elementary teachers, quality effect estimates range from .02 to .46 (Koedel & Betts, 2007; Leigh, 2009; Nye et al., 2004) based on quasi-experimental data. Teacher quality effects for elementary school teachers based on experimental data range from .12 to .38 (Kane & Staiger, 2008; Nye et al., 2004) and results for middle- and high-school teachers suggest effect sizes in the range of .10 to .35 (Koedel, 2009; Nye et al., 2004).

28. These estimates were derived using the FESE command in Stata.

29. The pattern of findings, however, is consistent if we instead allow the maximum sample size possible for each model specification.

30. For example, the EB effect sizes from the traditional model range from about 0.2 to 0.4, whereas effect sizes from the student fixed effects model are much larger.

31. This is somewhat analogous to a finding of smaller teacher effects in models that include school fixed effects because the school effects absorb some of the between school variation in teacher performance that arises given the nonrandom sorting of teachers into schools.

32. We estimated several additional variants of the models discussed in Section III. In particular, we estimated a less restrictive version of Model 2 that uses multiple beginning-of-year test scores by including each test score available as a separate regressor and a set of dummy variables for students missing test scores for particular subjects. We do not discuss this variant in any detail because the teacher effect estimates were very highly correlated with those generated from the traditional model; the lowest Pearson correlation was 0.67 (for Algebra I), and for the remaining subjects the correlation was 0.83 or higher. We also estimated model variants that included school fixed effects, but these produced quite noisy estimates because our sample includes a number of schools with only a handful of teachers in the sample. Finally, we estimated models predicting the relationship between observed teacher variables (specifically we included in the model teachers' grade point average [GPA], indicators for undergraduate major and if this is the same as the subject they taught, highest degree, years of experience, whether they were deemed highly qualified under No Child Left Behind [NCLB], and whether they were certified by the National Board for Professional Teaching Standards) and student achievement (similar to Clotfelter et al., 2010). *F* tests show these teacher variables to be jointly significant, but few are individually significant. For example, teachers holding an advanced degree are not found to be more effective. Nor is there much consistent evidence that a teacher's GPA is predictive of effectiveness. Indicators for teachers being fully certified or considered to be a highly effective teacher under NCLB are generally not statistically significant, but consistent with empirical evidence (Clotfelter et al., 2010; Rivkin et al., 2005; Rockoff, 2004), teachers are found to become more effective with additional experience early on in their careers.

33. To examine if these results are sensitive to subject area, we also estimate out models separately for the three English courses and then for the five math and science courses. The pattern of correlations of teacher effects across models is qualitatively similar to that shown in Table 4.

34. Ballou et al. (2004) found correlations between various VAM specifications to be as low as 0.53 in seventh grade and 0.79 in eighth grade. The lower bound for the correlations between models compared by Harris and Sass (2006) are 0.39, whereas Lockwood et al. (2007) find correlations as low as 0.49.

35. Recent research (e.g., Boyd, Lankford, Loeb, & Wyckoff, 2011; Goldhaber & Theobald, 2011; Hanushek, 2009) suggests the potential of value-added measures to affect the quality of the workforce through the composition of who is in it. This work, however, ignores the possibility that the use of performance measures might also have behavior effects on teachers (i.e., influence who opts to become a teacher or who opts to remain in the profession).

References

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

ACT. (n.d.). Retrieved November 1, 2009, from http://www.act.org/qualitycore/index.html

Aslam, M., & Kingdon, G. (2011). What can teachers do to raise pupil achievement? *Economics of Education Review*, *30*, 559–574.

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

Boyd, D., Lankford, H., Loeb, S., & Wyckoff, J. (2011). Teacher layoffs: An empirical illustration

of seniority versus measures of effectiveness. *Education Finance and Policy*, *6*, 439–454.

Carroll, J. B. (1993). *Human cognitive abilities: A survey of factor-analytic studies*. New York, NY: Cambridge University Press.

Chetty, R., Friedman, J. N., & Rockoff, J. E. (2011). *The long-term impacts of teachers: Teacher value-added and student outcomes in adulthood*. Cambridge, MA: National Bureau of Economic Research. Retrieved May 31, 2011, from http://obs.rc.fas.harvard.edu/chetty/value_added.pdf

Clotfelter, C. T., Ladd, J., & Vigdor, J. (2007). *How and why do teacher credentials matter for student achievement?* [Calder Working Paper 2 and NBER Working Paper 12828]. Cambridge, MA: National Bureau of Economic Research.

Clotfelter, C. T., Ladd, H., & Vigdor, J. (2010). Teacher credentials and student achievement in high school: A cross-subject analysis with student fixed effects. *Journal of Human Resources*, *45*, 666–681.

Coleman, J. S. (1966). *Equality of educational opportunity* [NBER Working Paper 12828]. Washington, DC: U.S. Dept. of Health, Education, and Welfare, Office of Education.

Common Core of Data. (2007–2008). *Public elementary/secondary school universe survey* (v. 1b). Alexandria, VA: National Center for Education Statistics. Available from http://nces.ed.gov/ccd/pubschuniv.asp

Dee, T. S., & Cohodes, S. R. (2005). *Out-of-field teachers and student achievement: Evidence from "matched-pairs" comparisons* [NBER Working Paper]. Cambridge, MA: National Bureau of Economic Research.

DeVellis, R. F. (2003). *Scale development: Theory and applications* (2nd ed.). Thousand Oaks, CA: Sage.

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

Goldhaber, D. (2002). Teacher quality and teacher pay structure: What do we know, and what are the options? *Georgetown Public Policy Review*, *7*(2), 81–94.

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

Goldhaber, D., & Brewer, D. (1997). Why don't schools and teachers seem to matter? Assessing the impact of unobservables on educational productivity. *Journal of Human Resources*, *32*, 505–523.

Goldhaber, D., & Hansen, M. (2010). Using performance on the job to inform teacher tenure decisions. *American Economic Review*. *100*, 388–392.

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

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

Goldschmidt, P., Choi, K. C., Martinez, F., & Novak, J. (2010). Using growth models to monitor school performance: Comparing the effect of the metric and the assessment. *School Effectiveness and School Improvement*, *21*, 337–357.

Goldschmidt, P., Martinez-Fernandez, J. F., Niemi, D., & Baker, E. L. (2007). The relationship among measures as empirical evidence of validity: Incorporating multiple indicators of achievement and school context. *Educational Assessment*, *12*, 239–266.

Hambleton, R. K., & Swaminathan, H. (1987). *Item response theory: Principles and applications*. Boston, MA: Kluwer.

Hanushek, E. A. (1979). Conceptual and empirical issues in the estimation of educational production functions. *Journal of Human Resources*, *14*, 351–388.

Hanushek, E. A. (1986). The economics of schooling—Production and efficiency in public schools. *Journal of Economic Literature*, *24*, 1141–1178.

Hanushek, E. A. (1997). Assessing the effects of school resources on student performance: An update. *Educational Evaluation and Policy Analysis*, *19*, 141–164.

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

Hanushek, E. A., & Rivkin, S. (2010). Generalizations about using value-added measures of teacher quality. *American Economic Review*, *100*, 267–271.

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

Heckman, J. J. (1995). Lessons from the bell curve. *Journal of Political Economy*, *103*, 1091–1120.

Hill, H. C. (2009). Evaluating value-added models: A validity argument approach. *Journal of Policy Analysis and Management*, *28*, 700–709. doi: 10.1002/pam.20463

Jacob, B., & Lefgren, L. (2008). Can principals identify effective teachers? Evidence on subjective performance evaluation in education. *Journal of Labor Economics*, *26*, 101–136.

Jacob, B., Lefgren, L., & Sims, D. (2008). *The persistence of teacher-induced learning gains* [Working Paper Series 14065]. Cambridge, MA: National Bureau of Economic Research.

Kane, T., & Staiger, D. O. (2008). *Are teacher value added estimates biased? An experimental validation of non-experimental estimates* [Working Paper Series 14607]. Cambridge, MA: National Bureau of Economic Research.

Koedel, C. (2009) An empirical analysis of teacher spillover effects in secondary in secondary school. *Economics of Education Review*, *28*, 682–692.

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

Konstantopoulos, S., & Chung, V. (2011). The persistence of teacher effects in elementary grades. *American Educational Research Journal*, *48*, 361–386.

Koretz, D. (2002). Limitations in the use of achievement tests as measures of educators' productivity. *Journal of Human Resources*, *374*, 752–777.

Koretz, D. (2005). *Alignment, high stakes, and the inflation of test scores* [Report 655]. Los Angeles, CA: National Center for Research on Evaluation, Standards, and Student Testing (CRESST) Center for the Study of Evaluation (CSE) Graduate School of Education & Information Studies, University of California.

Koretz, D. (2008). *Measuring up: What educational testing really tells us*. Cambridge, MA: Harvard University Press.

Leigh, A. (2009). *Estimating teacher effects from two-year changes in students' test scores* [Discussion Paper]. Acton, ACT: Australian National University, Center for Economic Policy Research.

Lockwood, J. R., & McCaffrey, D. (2007). Controlling for individual heterogeneity in longitudinal models, with applications to student achievement. *Electronic Journal of Statistics*, *1*, 223–252.

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

McCaffrey, D. F., Lockwood, J. R., Koretz, D. M., & Hamilton, L. S. (2004). Models for value-added modeling of teacher effects. *Journal of Educational and Behavioral Statistics*, *29*(1), 67–101.

Monk, D. H., & King, J. (1994). Multilevel teacher resource effects on pupil performance in secondary mathematics and science: The case of teacher subject-matter preparation. In R. Ehrenberg (Ed.), *Contemporary policy issues: Choices and consequences in education* (pp. 29–58). Ithaca, NY: Industrial and Labor Relations, Cornell University.

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

Papay, J. P. (2011). Different tests, different answers: The stability of teacher value-added estimates across outcome measures. *American Education Research Journal*, *48*, 163–193.

Raudenbush, S. (2001). Comparing personal trajectories and drawing causal inferences from longitudinal data. *Annual Review of Psychology*, *52*, 501–525.

Rivkin, S. (2009). The estimation of teacher value added as a determinant of performance pay. In D. Goldhaber & J. Hannaway (Eds.), *Creating a new teaching profession* (pp. 181–194). Washington, DC: Urban Institute Press.

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

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

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

Rowan, B., Chiang, F.-S., & Miller, R. J. (1997). Using research on employees' performance to study the effects of teachers on students' achievement. *Sociology of Education*, *70*, 256–284.

Schools and Staffing Survey. (2007–2008a). Percentage distribution of school teachers, by highest degree earned, school type, and selected school characteristics: 2007–08 [Table]. Alexandria, VA: National Center for Education Statistics.

Schools and Staffing Survey. (2007–2008b). Percentage distribution of school teachers, by total years of full-time teaching experience, years teaching at current school, school type, and selected school characteristics: 2007–08 [Table]. Alexandria, VA: National Center for Education Statistics.

Taub, G. E., Floyd, R. G., Keith, T. Z., & McGrew, K. S. (2008). Effects of general and broad cognitive abilities on mathematics achievement. *School Psychology Quarterly*, *23*, 187–198.

Toch, T., & Rothman, R. (2008) *Rush to judgment: Teacher evaluation in public education*. Washington, DC: Education Sector.

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

Weisberg, D., Sexton, S., Mulhern, J., & Keeling, D. (2009). *The widget effect: Our national failure to*

acknowledge and act on differences in teacher effectiveness. New York, NY: New Teacher Project.

Xu, Z., Hannaway, J., & Taylor, C. (2007). Making a difference? The effects of Teach for America in high school. *Journal of Policy Analysis and Management*, *30*, 447–469.

Yen, W. M. (1986). The choice of scale for educational measurement: An IRT perspective. *Journal of Educational Measurement*, *23*, 299–325.

Authors

DAN D. GOLDHABER is the Director of the Center for Education Data & Research and a professor in Interdisciplinary Arts and Sciences at the University of Washington-Bothell. His work focuses on issues of educational productivity and reform at the K–12 level, the broad array of human capital policies that influence the composition, distribution, and quality of teachers in the workforce, and connections between students' K–12 experiences and postsecondary outcomes.

PETE GOLDSCHMIDT, PhD, is currently on leave from his associate professor position in the College of Education at California State University-Northridge and serving as the assistant secretary for assessment and accountability in the New Mexico Public Education Department. His research interests include furthering methods in longitudinal modeling, educator evaluation, and school accountability.

FANNIE TSENG, PhD, is a Principal Research Analyst at Berkeley Policy Associates. Dr. Tseng's work focuses on the use of statistical models to measure the impacts of social programs on individual outcomes. She has conducted evaluation research in the areas of K-12 and post-secondary education, labor training and youth development. Dr. Tseng holds a BA in Economics and East Asian Studies from Brown University and a PhD in Economics from the University of Pennsylvania.

Article received June 21, 2010
First revision received October 19, 2011
Second revision received July 27, 2012
Accepted October 2, 2012