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

**CEDR Working Paper 2011-4.0**

Dan Goldhaber
Center for Education Data and Research
University of Washington

Pete Goldschmidt
California State University Northridge and CRESST/UCLA

Philip Sylling
University of Washington

Fannie Tseng
Berkeley Policy Associates

*The views expressed here are those of the author(s) and do not necessarily represent those of their affiliated institution(s), or funder(s). Any errors are attributable to the author(s). CEDR working papers have not gone through final review and should be cited as working papers. They are intended to encourage discussion and suggestions for revision before final publication.*

*The authors wish to acknowledge ACT for supplying the data used in the analyses described here and Jim Sconing and Zeyu Xu for helpful comments.*

*The suggested citation for this working paper is:*

Goldhaber, Dan, Peter Goldschmidt, Philip Sylling, and Fannie Tseng. *Teacher Value-Added at the High School Level: Different Models, Different Answers?* Working paper no. 2011-4.0. Center for Education Data & Research (CEDR), University of Washington, 2010. Web. <http://www.cedr.us/publications.html>.

## ABSTRACT

This paper 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, which allow us to examine the relationship between teacher effect estimates derived from VAM specifications employing different student learning assumptions. Specifically we explore the impact of assuming student learning to be unidimensional, prior knowledge ignorable, and achievement not influenced by concurrent teachers (spillover). We find evidence that both the estimated effect size of teacher quality changes and the estimates of individual teacher performance vary depending on learning assumptions. In particular, teacher effects estimated assuming unidimensionality and ignorable prior knowledge results in significantly smaller effect size estimates than those generated from a more traditional lagged score model, as well as changes in inferences about teachers. Similarly, teacher spillovers across subjects are found to have some impact on primary teacher estimates. Results imply that estimating high school teacher effects without explicitly considering the underlying assumptions results in biased estimates of performance; the substantive relevance is a normative question, but, our findings strongly suggest that VAM specification at the high school level warrants further research.

## I. Introduction

A considerable amount of research since, and including, the *Coleman Report* (Coleman, 1966) shows that of all the *school-related* factors that affect student achievement, teacher quality is the most important (e.g. Aaronson et al., 2007; Rivkin, et al., 2005).[1] A newer body of research (e.g. Kane et al., 2006) suggests teachers vary considerably from one another in effectiveness. It is not terribly surprising therefore that policymakers are considering options that tie consequential labor market decisions (e.g. compensation and tenure) to measures of individual teacher performance, often judged in part, on the student test performance.

The ability to use student assessment results to draw inferences about teacher performance, however, is potentially limited to only a portion of the teacher workforce, due to the fact that students are not tested in a content area in consecutive years in all grades and subjects. Specifically, value-added models (VAMs) of one sort or another are a leading method of 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

[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 et al., 2005;
[2] For more information on Race to the Top, see: http://www2.ed.gov/programs/racetothetop/index.html. See also

for a prior level of achievement in a subject area).[3] But these types of models are complicated at the high school level by the fact that high school students are often not tested annually—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 and Kingdon, 2010; Clotfelter et al., 2007a; Dee and Cohodes, 2005; Xu et al., 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 paper reports on 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 and, *information on student achievement in each subject at the beginning and end of each school year*. These data allow us to estimate the impact of changes in teacher effectiveness on student achievement in different subject areas at the high school level, and compare teacher effect estimates derived

---

[3] See, for instance, Rivkin (2009), Rothstein (2009), Todd and Wolpin (2003), for a discussion of the assumptions underlying traditional 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 et al., 2007).
[5] Whereas at the elementary level, one might attribute the gain in student achievement from the end of the 4th grade to the end of the 5th grade as due in part to the 5th grade teacher because elementary coursework follows a linear progression and the annual testing effectively provides pre- and post-tests.
[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."

from VAM specifications that rely on very different theoretical assumptions about student learning.

We find evidence that both the estimated effect size of teacher quality changes and the estimates of individual teacher performance vary depending on model specification. In particular, teacher effects estimates based on a student fixed effects specification results in significantly small smaller effect size estimates than those generated from a more traditional lagged score model. We cannot definitively say whether one model specification is preferred over another. And while the policy import of the importance of model specification for determining teacher effects is a normative question, our findings do strongly suggest that assumptions about what drives student achievement, and consequently VAM specification, at the high school level warrants further research.

The paper is arranged as follows. Section II provides more background on value-added models at the high school level. In Section III we describe our data and analytic approach. Our findings are presented in Section IV. Finally, in Section V we offer some concluding thoughts on the policy implications of the findings.

## II. Value-added Research at the High School Level

Since the 1966 *Coleman Report* (Coleman et al., 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).[8] Estimation of teacher effects at the high school level are complicated due to 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 utilize 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. (2007a) 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 lacks a clear-cut baseline test for prior subject-specific student achievement, both sets of authors model within-student variation *across subjects* in a year. Clotfelter et al. find that teachers' credentials (in-subject test scores) have positive effects on students, particularly in math. Xu et al. assess the differential impacts of Teach for America (TFA) and traditional classroom teachers and find TFA teachers to be more effective.

Another option is the use of End of Course (EOC) exams that constitute the standardized assessments that students take at the end of a course in high school. A limitation is that EOCs may not be well-aligned with assessments taken in a prior year because many states have moved towards 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

[8] See, for instance, Clotfelter et al. (2006); Goldhaber and Brewer (1997a,b); Goldhaber and Anthony (2007); Monk (1994); Rockoff (2004); and Rowen et al. (1997) as examples.

a key assumption that the outcome has constant meaning over time (Raudenbush, 2001)[9], and is not based on a vertically equated IRT-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 and Swaminathan, 1987).

Estimation of teacher effects at the high school level are also complicated by the typically more complex structure of these schools. Students usually receive instruction from multiple teachers, each specializing in a subject area, and we know little about whether estimates of teacher effectiveness at the secondary level are influenced by complementarities across subjects like mathematics and science or the baseline achievement controls that are used in VAMs. As we touched on above, this is a severe limitation for any policy that relies on VAM estimates of individual teacher effects.[10] A finding, for instance, that effectiveness estimates in one subject are influenced by teacher quality in a second, call into question policies that target individual teachers and utilize VAMs for high-stakes purposes. To our knowledge, only one published paper comes close to addressing this issue. Koedel (2009) estimates value-added reading achievement models and tests whether students' reading achievement is influenced not only by the quality of their English teachers but also by their mathematics, social studies, and science teachers. He finds evidence that math teachers have significant spillover effects on student reading performance but concludes that this finding may be due to the type of nonrandom

[9] Below we describe the specific characteristics and unique implementation of the EOCs we use in this analysis.
[10] A recent paper at the elementary level, (Jackson and Bruegmann, 2009), suggests that a non-trivial (about 12 percent) proportion of estimates of individual teacher effectiveness is explained by the quality of a teacher's peers. This line of work, however, investigates whether a teacher's effectiveness at teaching her own students impacts her peers' effectiveness at teaching their own students. We focus on the direct impact of a teacher on her students' performance in other courses.

student-teacher matching described in Rothstein (2009) as future math teachers also have predictive power for current reading performance.

In the next section we describe the data and analytic approach we use to explore the relationship between teacher quality and student achievement at the secondary level and the extent to which the estimates of individual teacher impacts are influenced by VAM specification—specifically whether a cross-subject model provides estimates that are consistent with more traditional estimates

## III. Analytic Approach and Data

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

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

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, $\mu^G_i$, which is common across all subject areas, subject-specific student ability, $\mu^S_i$, which only impacts subject *S* (which may be zero if unobserved student ability does not differ across subjects), 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.[11]

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

[11] Note that in some model specifications described in the results section we replace the vector of teacher indicators with observable teacher characteristics.

elementary level, where student test histories may be minimal at the earliest grades, researchers typically utilize some variant of the following model:

$$A_{ijt}^{S} = \alpha A_{ij(t-1)}^{S} + X_{it}\gamma + C_{t}\beta + T_{jt}^{S}\tau + \varepsilon_{ijt}^{S} \qquad (2)$$

where $\alpha A_{ij(t-1)}^{S}$ serves as a proxy for learning prior to time t and unobserved ability. The assumption here is that $\alpha A_{ij(t-1)}^{S}$ serves as a proxy for both $\mu_i^G$ and $\mu_i^S$.[12],[13]

The strategy used to estimate VAMs at the high school level, where repeated measures of student performance in a subject are typically unavailable (so there is no measure of prior year student achievement), is to utilize a variant of (1) where achievement across a variety of subject areas is used to obtain an estimate of $\mu_i^G$, and, importantly, $\mu_i^S = 0$ is assumed for all subjects (referred to here as the "blank slate assumption")[14], This results in the following student fixed effects model:

$$A_{ijt}^{S} = \mu_i^G + X_{it}\gamma + C_{t}\beta + T_{jt}^{S}\tau + \varepsilon_{ijt}^{S} \qquad (3)$$

The significant difference between (2) and (3) is that the measure of prior achievement has been replaced by a measure of general student ability, $\mu_i^G$, which is constant across all subject areas. Here teacher effects are identified based on within-student variation in achievement across

---

[12] 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, $\mu_i$ which should 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, 2009).

[13] An alternative specification (based on different assumptions) treats $\mu_i$ as a random effect, which allows for the estimation of time invariant $X_{it}$ (McCaffrey, et. al., 2004).

[14] 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 can last up to three years Konstantopoulos, S. & V. Chung (2011) and the persistence parameter ranges from about .2 to .3 (Jacobs et. al.,, 2008, McCaffrey, et. al., 2004), Lockwood & McCaffrey, 2007). However, as others have pointed out, Jacobs (2008), persistence may be less of an issue in less common subjects (e.g. biology) where there has been little prior instruction.

subjects as opposed to within-student variation in achievement over time. Demeaning both sides of equation (3) within students yields:

$$A_{ijt}^{S} - \overline{A}_{ijt} = \left(T_{jt}^{S} - \overline{T}_{jt}\right)\tau + \left(\varepsilon_{ijt}^{S} - \overline{\varepsilon}_{ijt}\right) \qquad (4)$$

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 pre-test score on the right hand side of the model, is that students start a subject with a blank slate, controlling for general ability.[15] In other words, this specification assumes that there are no differences within-students in initial knowledge across subjects. 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 that influence achievement in one subject area have the same effect in other subjects as well. This implies that in equation (1) above, $\mu_i^S = 0$ for all subjects and thus $\mu_i^G$ drops out after demeaning the variables in equation (3). Were this not the case then the within-student relative subject-specific ability term, $\mu_i^S - \overline{\mu}_i^S$, would be subsumed into the error term and may bias the teacher effect estimates since the adjusted gain across subject areas could not be attributed strictly to the educational resources at the school.[16]

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 (Dickens, 2008). In fact,

---

[15] Models that include a measure of prior student learning are thought to account for schooling inputs in prior periods since 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.

[16] 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 since the model would underpredict a student's true ability in math and therefore attribute the relatively good performance to teachers with master's degrees.

Heckman (1995) argues that both general ability and other factors play a role in affecting achievement.[17]

Clotfelter et al. (2007a) examine the issue of unidimensionality of student ability in two ways. First, they find 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, for example, advanced algebra. Second, they note the result from an OLS regression that relative student scores (from the eighth grade) are not a significant predictor of relative teacher licensure scores (for future high school teachers). This finding is consistent with the assumption that students are not assigned to teachers based on their relative math and reading ability. 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

[17] Recent evidence suggests that general ability directly impacts specific broad cognitive abilities, thus indirectly playing a substantial role in (mathematics) achievement (Taub, Floyd, Keith, and McGrew, 2008) and cross-subject assessment results have been included in achievement models specifically to account for general ability (Goldschmidt, Martinez-Fernandez, and 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 non-unidimensionality of student ability, content coverage, or non-unidimensionality in teacher quality.

assigned to teachers based on relative ability, then a violation of unidimensionality implies that fixed-effects estimates of teacher quality may still 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. This is possible because the unique data we employ for this study includes both within-student, pre-post, and multi-subject test results for a subsample of the students. Specifically, we estimate models that expand on the Clotfelter et al. (2007a) and Xu et al. (2007) frameworks by adding *within*-student demeaned pre-test scores to the right-hand side of equation (4):

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

We term this model the "comprehensive model" in discussion of the results. Consider the implication of a finding that $\alpha \neq 0$. 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 $\mu_i^G$). 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.

We can also 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 through the correlation between the γ estimates from these different models. If the estimates

from these 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.

The fact that we have students in the sample who have multiple teachers in a year also allows us to assess the role of joint production of achievement or "spillover" effects as in Koedel (2009). Controlling for joint production at the high school level is potentially important for two reasons. First, students may be non-randomly assigned to teachers as they reach and progress through high school. For example, if students with the best English teachers are also assigned the best biology teachers and above average instruction in biology positively influences English achievement, then the overall variance of teacher effectiveness of English teachers is overestimated and their individual teacher effect estimates are biased (downward in the stylized example). This story is typical if students are stratified by ability in high school and tracked into different clusters of classes. On the other hand, if students were compensated for "bad" teacher assignments in one subject with "good" teacher assignments in another subject, the variance of teacher effectiveness may be understated in a given subject. In either case, failing to account for heterogeneity in students' teacher assignments across classes may bias individual estimates of teacher effectiveness. Second, a richer understanding of joint production or complementarities across subjects is important for its potential for increasing *total* education production.

Specifically, we accomplish this by augmenting equation (2) with teacher fixed effects from cross-subject teachers, $T_{kt}^{S_{\neq w}}$, where k indexes teachers in a subject other than the target subject w. We test whether these cross-subject teachers are jointly significant in the estimated models:

$$A_{ijkt}^{S_w} = \alpha A_{ijk(t-1)}^{S_w} + X_{it}\gamma + C_t\beta + T_{jt}^{S_w}\tau_1 + T_{kt}^{S_{\neq w}}\tau_2 + \varepsilon_{ijkt}^{S_w} \quad (6)^{18}$$

The ACT data we utilize allows us to explore the issue of cross-subject controls because it includes beginning-of-year and end-of-year test scores in multiple subject areas as well as matched teacher identifiers for all subjects in which a student is tested. Each student in our sample is 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 for each school, class size; teachers' college GPA, college major, highest degree, certification status, and years of experience.

The analysis benefits from specific properties of the outcome that further enhances generalizability. The assessment outcomes are 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 in order to succeed in college, and, within subject area, the EOC exams were scaled to have constant meaning across forms (ACT, 2010). The reliabilities of the EOCs range from a low of .78 (geometry) to .94 (English 11) (ACT, 2010), 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

[18] We also try to include cross-subject teachers in the other two models but this significantly increases the complexity because the own-subject effect and the cross-subject effect must enter the equation when student test scores are demeaned within students. Identification is difficult if there is insufficient mixing across student-teacher combinations.

not have constant meaning across subject. We standardize these scaled scores for all students within each subject area to be ~N(0,1).[19]

An important aspect of the assessment outcome is the fact that it is a low-stakes assessment, which is less likely to suffer from coaching, or score inflation, rendering student gains generally more attributable to (desired) teacher behavior (Koretz, 2002), which assessment alignment to standards is insufficient to guarantee (Koretz, 2005). On the other hand, one might worry that students do not really try to do well on these low-stakes tests so that we do not get a very good measure of their true learning gains (Koretz, 2008).[20] Another potential problem is that pretest scores in relatively unfamiliar subjects, such as biology and chemistry, may by uninformative predictors of posttest scores, if, for instance, most students begin the year with little baseline knowledge in those subjects.[21]

The data used in the analyses were collected by ACT as part of a pilot of their QualityCore end-of-course assessments. We restrict our analysis dataset to schools and classrooms where teachers and students can be uniquely linked. This includes 23 schools (9 of which are private schools), 205 teachers, and 8,002 students (in grades 9 through 12).[22] In **Table 1** below we report selected school (Panel A) and teacher (Panel B) characteristics, and include a

---

[19] Using a normalized metric in the model does not affect inferences regarding relative performance (Goldschmidt, Choi, Martinez, and Novak, 2010).

[20] About 23 percent 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 percent and then the extreme 2 percent 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.

[21] We test this here and find this not to be the case. Specifically, pretest scores in biology and chemistry explain much more of the variation in their respective posttest scores than do pretest scores in algebra I, which might have been thought to be a better predictor of ability coming into a science course. Moreover, the magnitude of the coefficient estimates on the biology and chemistry pretests, 0.72 and 0.54, respectively, are within the range of what is typically found in models where the pre- and post-tests are thought to be well-enough aligned to be used for value-added analyses.

[22] 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 our not linked.

comparison of the poverty level of these with information on high schools in the Midwest from the 2007-08 *Common Core of Data,* and teacher characteristics from the 2007-08 wave of the *Schools and Staffing Survey*. In terms of poverty level (percent 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.[23] The teachers in our sample are slightly less credentialed and slightly more experienced than teachers in the U.S., but these differences are quite small.[24]

If student achievement is unidimensional, we would expect to see a high correlation of test scores across subjects within students, i.e. a student who performs well in one subject area is likely to also perform well in others. We find that there is equivocal evidence of unidimensionality in Panel A of **Table 2**, which shows the correlations between selected subject post-test scores for students who took multiple subject tests. The standardized student post-test scores are positively correlated across subjects, but the correlations vary widely, ranging from .01 to .92. While the positive correlation is generally consistent with the argument that student ability is unidimensional, we cannot tell from this simple analysis the degree to which these post scores are influenced by teacher effects and/or whether measurement error may explain the low correlation that we observe for some subject combinations.

Panel B of **Table 2** shows the average test scores in seven subjects (Algebra I, Algebra II, Geometry, Biology, Chemistry, and $10^{th}$ and $11^{th}$ grade English) for students who sat for one,

[23] 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.
[24] We could not find comparable data for the Midwest teachers only.

two, or three or more subject tests.[25] 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.

## IV. Results

We begin by briefly describing our findings on the relationship between student achievement and observable teacher characteristics. To obtain these results, we estimate models consistent with equations 2, 4, and 5 above, but including observable teacher characteristics in place of teacher indicator variables.[26]

As we described above, most value-added studies do not show a consistently strong relationship between observed teacher credentials and student achievement, at least in the expected directions. Our findings are similar. F-tests show the joint significance of the 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 grade point average 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., 2007a; Rivkin et al., 2005; Rockoff, 2004), teachers are found to become more effective with additional experience early on in their careers.

---

[25] 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.

[26] These results are available from authors upon request.

### *A. Influence of Model Specification on Estimated Teacher Effects*

As we mention above, 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 eighteen studies on the effect size of teacher quality in a recent review by Nye et al. (2004), only four include exclusively 6th grade or higher, and only one includes grades higher than 9th 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] 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 error using an Empirical Bayes (EB) approach (consistent with Aaronson et al., 2007).[29] Note that 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.[30] The effect sizes vary

---

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

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

[29] The average sampling variance of the individual teacher effect estimates is subtracted from the total variance of the sample of observed estimates based on $\sigma^2_{\hat{\gamma}} = \sigma^2_{\gamma} + \sigma^2_{\eta}$ where $\sigma^2_{\gamma}$ represents the true variance of the teacher effects and $\sigma^2_{\eta}$ represents the portion of the variance of the observed teacher effects that is due to sampling error and is estimated by calculating the average sampling variance of the individual teacher effect estimates.

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

by subject area, but our findings also suggest that they differ by methodological approach.[31] Specifically, column 1 of the table shows the results when we utilize the traditional lagged score approach, column 2 shows the results from a specification with student fixed effects (consistent with Clotfelter et al., 2007a, and Xu et al., 2007), while column 3 reports results from our comprehensive model (equation 5 above).

Looking across all subjects, the unadjusted estimates range from about 0.16 to 0.44. 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) report 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 the fact that there is a substantial difference in the estimate impact of teachers across model specifications. The estimated contribution of teachers based on the traditional model (column 1) is consistently far larger than the student fixed effects model (column 2). Specifically, the effect size derived from the traditional model is about one and one-half to two times as large as the point estimates derived from the student fixed effects specifications (the same pattern tends to be present for the EB estimates as well).

The coefficient estimate on $\alpha$ is statistically significant (and positive) indicating that one of the assumptions underlying equation 4 is violated. Unfortunately, we cannot definitely tell whether it is the blank slate or unidimensionality assumption, or both. If subject-specific ability is an important component of overall ability and, in fact, students are negatively matched to

[31] For example, the EB effect sizes from the traditional model range from about 0.2 to 0.4 while effect sizes from the student fixed effects model are much larger.

teacher quality based on this component rather than overall ability (e.g. students who are especially poor at math receive the best math teachers) then the student fixed effects model may understate the importance of teacher effectiveness. While the traditional model imperfectly controls for unobserved ability, it does proxy for overall and subject-specific components and so the model may be more resistant to bias caused by nonrandom student teacher matching.

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 pre-test score in the regression may reduce upward bias in estimates of teacher effectiveness from the student fixed effect model. All the unadjusted (and most of the EB adjusted) standard deviations in column 3 are all less than their counterparts in column 2. Within the context of the comprehensive model, the demeaned pre-test score effect results are consistent with prior research focusing on learning decay and persistence of teacher effects (Jacobs et. al., 2008; Lockwood et. al, 2007: McCaffrey et. al., 2004)

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) while columns 1 and 3 show that estimates from the traditional model are far less correlated with those

[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 utilizes multiple beginning of year test scores by including each of 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 since our sample includes a number of schools with only a handful of teachers in the sample.

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]

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 non-trivial one. We explore the extent to which the differences in teacher effect estimates are policy-relevant in greater detail in Section V below.

### *B. Assessing Spillover Effects: The Influence of Cross-Subject Teachers*

We begin an assessment of assessing the potential for spillover effects by conducting F-tests for the hypothesis that all cross-subject teacher effects are zero in our traditional model, as shown in column 1 of **Table 5**.[34] With the exception of Geometry, the hypothesis of no cross-subject teacher effect is consistently rejected at the 95 percent confidence level. This finding of teacher spillovers from one subject to another is consistent with Koedel (2009) who finds statistically significant effects of math and science teachers on student reading performance. This is prima facie evidence that spillovers ought to be considered in estimating teacher effects. However, it is not inconceivable that cross-subject teachers are important, but accounting for their contributions to student learning does little to change the estimated impact of own-subject teachers.

Column 2 of the table shows the Spearman correlation between own-subject effect estimates (from the traditional model) that omits cross-subject teachers with those from the

---

[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] We cannot estimate the effect of cross-subject teachers in the fixed effects or comprehensive models due to the collinearity between the subject tests and the cross-subject teacher dummies.

traditional model that includes them is greater than 0.7 for teachers in all subjects. Comparing these correlations to the much lower ones in Table 4 (which explores the impact of model specification) suggests that model misspecification in terms of student ability or the persistence of past educational inputs may be a greater source of bias in individual teacher effect estimates than omission of cross-subject teachers. If the traditional model were assumed to be correctly specified the decision to account for teacher spillovers may be relatively more trivial.

## *V. Public Policy Implications and Conclusions*

In the previous sections we have shown that model specification has a large influence on the estimated effect size of changes in teacher effectiveness and that models that include cross-subject teachers show cross-subject teachers to be statistically significant. A natural question is whether these findings have policy-relevant implications for the estimates of teacher effectiveness. For example, the findings we report above do not necessary imply that moving from one model specification to another would have a material impact on the performance *ranking* of teachers. We explore this issue in **Table 6**, which shows the percentage of teachers who fall into a quintile given one specification of the model and the same or a different quintile using an alternative specification. Quintile 1 represents the highest ranked teachers and Quintile 5 represents the lowest ranked teachers.

There is not an inconsiderable 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 to student fixed effects models. Were it the case that the model specification was irrelevant to a teacher's rank in the distribution, we would expect the diagonal elements of the matrix to each be 100 percent and the off-diagonal elements to be zero. This

clearly isn't the case. For instance, we observe (Table 6, Panel A) that about ten percent 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 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. These findings are less extreme when it comes to the import of including cross-subject teachers in the model (Table 6, Panel C). For instance, there are no teachers in the bottom quintile in the specification omitting cross-subject teachers who are found in either of the top two quintiles in the specification that includes cross-subject teachers.

Some suggest that value-added measures are not reliable enough to use; Hill (2009), for instance, argues, "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). On the other hand, there is little evidence that non-VAM teacher evaluation methods are rigorous, and that value-added estimates of teacher effectiveness are better predictors of future student achievement than other teacher credentials typically used for employment and compensation in the teacher labor market (Glazerman et al., 2010; Goldhaber and Hansen, 2010).

There is no right answer to the above line of debate about the use of VAM approaches to evaluate teachers. It is a normative question whether it might be worthwhile to accept some misclassification of teachers in order to have a more rigorous evaluation system. The findings we report in this paper, however, clearly highlight concerns about the use of VAM estimates at the high school level for the evaluation of individual teachers. In particular, we explain

differences between VAM specifications by relating assumptions about how students learn to how students are matched to teachers, a promising direction for future research is to study VAM estimates in an experimental setting where students are randomly matched to teachers. This has been done at the elementary level by Kane and Staiger (2008) and is currently being pursued at the high school level as well.[35] Our results also have implications for future academic research on value-added modeling. Removing student heterogeneity in ability by using contemporaneous student performance across a variety of subjects may be insufficient to generate unbiased teacher effect estimates if student ability varies considerably across subjects, particularly if this information is used to match students to teachers. We have shown that student ability may have a large component that is specific to a given subject and our results are consistent with the idea that this component is a factor in student teacher matching.

[35] The Measures of Effective Teaching study to be undertaken in New York City schools utilizes randomized classroom assignment; see (Medina, 2009).

**REFERENCES**


Aaronson, D., L. Barrow, and W. Sander. (2007). "Teacher and Student Achievement in the Chicago Public High Schools." Journal of Labor Economics 25(1): 95-135.

ACT (retrieved 11/1/2009). 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(3), pp:559-574.

Ballou, D (2005). Value-added assessment: Lessons from Tennessee. In R. Lissetz (Ed), *Value Added Models in Education: Theory and Applications*. Maple Grove, MN: JAM Press.

Bressoux, .P. and M. Bianco (2004). The Long-term teacher effects on pupils' learning gains, *Oxford Review of Education*, 30(3), pp. 327–345.

Clotfelter, C.T., H. Ladd, and J. Vigdor. (2007a) "Teacher Credentials and Student Achievement in High School: A Cross-Subject Analysis with Student Fixed Effects. Calder Working Paper 11.

Clotfelter, C. T., H. Ladd and J. Vigdor (2007b) "How and Why Do Teacher Credentials Matter for Student Achievement?" Calder Working Paper 2 and NBER Working Paper 12828.

Dee, T. S. and Cohodes, S. R. (2005). "Out-of-Field Teachers and Student Achievement: Evidence from 'Matched-Pairs' Comparisons." NBER Working Paper.

DeVellis, R. F. (2003). *Scale development: Theory and applications, Second Edition*. Thousand Oaks, CA: Sage Publications.

Ehrenberg, R. G., & Brewer, D. J. (1994). Do school and teacher characteristics matter? Evidence from high school and beyond. Economics of Education Review, 13, 1-17.

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. and M. Hansen. (2009). "Assessing the Potential of Using Value-Added Estimates of Teacher Job Performance for Making Tenure Decisions." CRPE Working Paper #2009-2

Goldhaber, D. and M. Hansen. (2010). "Is it Just a Bad Class? Assessing the Stability of Measured Teacher Performance." CEDR Working Paper 2010-3.

Goldhaber, Dan and Brewer, Dominic. "When Should We Reward Degrees for Teachers?" (1998). *Phi Delta Kappa*, 80(2): 134–138.

Goldhaber, D. and Brewer, D.. "Why Don't Schools and Teachers Seem to Matter? Assessing the Impact of Unobservables on Educational Productivity." (1997). *Journal of Human Resources*, 32(3): 505-523.

Goldhaber, D. and Brewer, D. "Does Teacher Certification Matter? High School Teacher Certification Status and Student Achievement." (2000). *Educational Evaluation and Policy Analysis*, 22: 129–145.

Goldhaber, D. "Teacher Quality and Teacher Pay Structure: What Do We Know, and What are the Options?" (2002). *Georgetown Public Policy Review*, 7(2): 81-94.

Goldschmidt, P., J.F. Martinez-Fernandez, D. Niemi, and E. L. Baker (2007). The Relationship Among Measures as Empirical Evidence of Validity: Incorporating Multiple Indicators of Achievement and School Context, *Educational Assessment,*12(3-4), 239–266

Goldschmidt, P., K.C. Choi, and F. Martinez, and J. Novak (2010). Using growth models to monitor school performance: comparing the effect of the metric and the assessment, *School Effectiveness and School Improvement,* 21(3), 337–357.

Gordon, R., T. J. Kane, and D. O. Staiger, "Identifying Effective Teachers Using Performance on the Job." Discussion Paper 2006-01, The Hamilton Project. April 2006.
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(3): 351–388.
Hanushek, E. A. (1986). "The Economics of Schooling - Production and Efficiency in Public-Schools." Journal of Economic Literature 24(3): 1141-1178.
Hanushek, E. A. (1997). "Assessing the Effects of School Resources on Student Performance: An Update." Educational Evaluation and Policy Analysis 19(2): 141–64.
Hanushek, E. A. (2004). "Why Public Schools Lose Teachers." Journal of Human Resources 39(2): 326–354.
Hanushek, E.A. (2009). "Teacher deselection." In *Creating a New Teaching Profession*, edited by Dan Goldhaber and Jane Hannaway. Washington, DC: Urban Institute Press.
Hanushek, E.A., and S. Rivkin (2010).

Heckman, J.J. (1995). Lessons from the Bell Curve, *The Journal of Political Economy*, v103,pp. 1091–1120. Jackson, C. K. (2009) "Student Demographics, Teacher Sorting, and Teacher Quality: Evidence From the End of School Desegregation" *The Journal of Labor Economics*. 27(2) 213–256.
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
Jackson, C.K., and E. Bruegmann. (2009). "Teaching Students and Teaching Each Other: The Importance of Peer Learning for Teachers." American Economic Journal: Applied Economics, 1(4), 1–27
Jacob, B., L. Lefgren, D. and Sims (2008). The Persistence of Teacher Induced Learning Gains, NBER Working Paper Series 14065, Cambridge, MA.
Kane, T. and D. O. Staiger (2008). Are Teacher Value Added Estimates Biased? An Experimental Validation of Non-Experimental Estimates, NBER Working Paper Series 14607, Cambridge, MA.
Koedel, C. (2009) An empirical analysis of teacher spillover effects in secondary in secondary school, *Economics of Education Review*, 28, pp. 682–692.
Koedel, C. and J. Betts (2007). Re-Examining the Role of Teacher Quality in the Educational Production Function, National Center on Performance Incentives, Working Paper 2007-03, Vanderbilt Peabody College, TN.
Konstantopoulos, S., & V. Chung (2011).The Persistence of Teacher Effects in Elementary Grades, *American Educational Research Journal* April 2011, Vol. 48, No. 2, pp. 361–386.
Koretz, D. (2002). Limitations in the use of achievement tests as measures of educators' productivity. *The Journal of Human Resources*, v. 374(4), pp.752–777.
Koretz, D. (2005). Alignment, High Stakes, and the Inflation of Test Scores. CSE Report 655. 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, Los Angeles.
Koretz, D. (2008). Measuring Up: What Educational Testing Really Tells Us. Harvard University Press.

Kukla-Acevedo, S. (2009). "Do teacher characteristics matter? New results on the effects of teacher preparation on student achievement." Economics of Education Review, 28(1) pp 49–57.

Leigh, A. (2009). Estimating Teacher Effects from Two-Year Changes in Student's Test Scores, Center for Economic Policy Research, Discussion Paper, Australian National University.

Lockwood, J., & 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., and Martinez, J.F. (2007). "The Sensitivity of Value-Added Teacher Effect Estimates to Different Mathematics Achievement Measures." Journal of Educational Measurement. 44(1): 47–67.

Medina, J. (2009). "A 2-Year Study to Learn What Makes Teachers Good." New York Times. September 1, 2009.

McCaffrey, D.F., J.R. Lockwood, D.M. Koretz, and L.S. Hamilton. (2004). "Models for Value-Added Modeling of Teacher Effects", *Journal of Educational and Behavioral Statistics*, 29(1), 2004

McCaffrey, D., T. Sass, J.R. Lockwood, and K. Mihaly (2009). The Inter-Temporal Variability of Teacher Effects Estimates, National Center on Performance Incentives, Working Paper 2009-03, Vanderbilt Peabody College, TN.

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: ILR.

Murnane, R.J., B.R. Phillips. (1981). "What do effective teachers of inner-city children have in common." Social Science Research. 10(1) pp 83-100.

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

Nye, B., S. Konstantopoulos, and L. V. Hedges (2004). How Large are Teacher Effects?, *Educational Evaluation and Policy Analysis*, 26(3), pp. 237-257.

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

Rivkin, S., E. A. Hanushek, and J.F. Kain. (2005). "Teachers, Schools and Academic Achievement." Econometrica 73(2): 417–458.

Rivkin, S. (2009). "The Estimation of Teacher Value Added as a Determinant of Performance Pay." In *Creating a New Teaching Profession*, edited by Dan Goldhaber and Jane Hannaway. Washington, DC: Urban Institute Press.

Rockoff, J. E. (2004). "The Impact of Individual Teachers on Students Achievement: Evidence from Panel Data." American Economic Review 94(2): 247–252.

Rothstein, J. (2010). "Teacher Quality in Educational Production: Tracking, Decay, and Student Achievement." Quarterly Journal of Economics, 125(1): 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.

Sass, T. R. (2008). "The Stability of Value-Added Measures of Teacher Quality and the Implications for Teacher Compensation Policy" Policy Brief 4, Washington, DC: The Urban Institute, Center for Analysis of Longitudinal Data in Education Research

Taub, G.E., R.G. Floyd, T.Z. Keith, and K.S. McGrew (2008). Effects of General and Broad Cognitive Abilities on Mathematics Achievement, *School Psychology Quarterly*, v23(2), pp. 187–198.

Todd, P.E., and K.I. Wolpin. (2003). "On the Specification and Estimation of the Production

Xu, Z., J. Hannaway, and C. Taylor. (2007). "Making a Difference? The Effects of Teach For America in High School." Calder working paper #17

Yen, W. M. (1986). The choice of scale for educational measurement: An IRT perspective. *Journal of Educational Measurement, 23*(4), 299–325.

## 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* [1] | | |
| 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 U.S. (2007-08)[b] |
| Less than a BA | 0% | 2% |
| BA/BS | 51% | 44% |
| MA/MS | 46% | 46% |
| EdD/PhD | 2% | 8% |
| Less than four years teaching experience | 18% | 19% |
| 4–9 years of teaching experience | 29% | 28% |
| 10 or more years of teaching experience | 53% | 53% |

[a] SOURCE: National Center for Education Statistics, Common Core of Data (CCD), "Public Elementary/Secondary School Universe Survey", 2007-08 v.1b

[b] SOURCE: U.S. Department of Education, National Center for Education Statistics, Common Core of Data (CCD), Schools and Staffing Survey (SASS), Table "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 "Percentage distribution of school teachers, by highest degree earned, school type, and selected school characteristics: 2007–08".

**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 | 12th Grade English | Geometry |
|---|---|---|---|---|---|---|---|---|
| Algebra I | 1 | | | | | | | |
| Algebra II | n/a<br>n/a<br>2 | 1<br><br>2 | | | | | | |
| Biology | 0.540<br>0.000<br>406 | 0.560<br>0.000<br>96 | 1<br><br>502 | | | | | |
| Chemistry | 0.344<br>0.092<br>25 | 0.619<br>0.000<br>512 | 0.924<br>0.008<br>6 | 1<br><br>1,626 | | | | |
| 10th Grade English | 0.254<br>0.001<br>160 | 0.609<br>0.000<br>247 | 0.649<br>0.000<br>738 | 0.503<br>0.000<br>296 | 1<br><br>1,441 | | | |
| 11th Grade English | 0.543<br>0.007<br>23 | 0.504<br>0.000<br>613 | 0.600<br>0.000<br>624 | 0.610<br>0.000<br>624 | 0.301<br>0.297<br>14 | 1<br><br>1,711 | | |
| Geometry | 0.008<br>0.977<br>14 | 0.359<br>0.051<br>30 | 0.587<br>0.000<br>643 | 0.575<br>0.000<br>233 | 0.460<br>0.000<br>833 | 0.225<br>0.002<br>183 | 0.311<br>0.114<br>27 | 1<br><br>1,963 |

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

**Panel B. Subject Pre-Test 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 |

**Table 3. Standard Deviation of Teacher Effect Estimates Under Different Model Specifications**

The standard deviation of teacher effectiveness is represented as follows:
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/NA |
| Chemistry Teachers | 26 | 0.442/0.281 | 0.197/0.035 | 0.156/NA |
| English 10 Teachers | 25 | 0.258/0.025 | 0.186/0.034 | 0.180/0.055 |
| English 11 Teachers | 34 | 0.426/0.294 | 0.206/0.086 | 0.172/0.115 |
| Geometry Teachers | 38 | 0.417/0.350 | 0.191/0.135 | 0.165/0.034 |

**Table 4. Spearman Correlations of Teacher Effect Estimates Across Model Specifications**

| | (1) Correlation Between Traditional and Student F.E. Models | | (2) Correlation Between Student F.E. Models and Student F.E. with Lagged Score Models | | (3) Correlation Between Student F.E. with Lagged Score and Traditional 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 | ** |
| English 10 Teachers | 0.293 | | 0.921 | ** | 0.245 | |
| English 11 Teachers | 0.674 | ** | 0.925 | ** | 0.529 | ** |
| Geometry Teachers | 0.464 | ** | 0.901 | ** | 0.677 | ** |

* and ** denote statistical significance at the 5 and 1 percent levels, respectively.

**Table 5. Effect of Cross-subject Teachers on Student Achievement**

| | | Traditional Model | Fixed Effects Model | Comprehensive Model | Spearman Correlations[a] |
|---|---|---|---|---|---|
| Algebra I | F | 2.30 | 5.44 | 3.51 | 0.892** |
| | p | 0.001 | 0.000 | 0.000 | |
| | N (Teachers) | 22 | 19 | 20 | |
| Algebra II | F | 1.30 | 10.21 | 5.88 | 0.791** |
| | p | 0.110 | 0.000 | 0.000 | |
| | N (Teachers) | 35 | 34 | 32 | |
| Biology | F | 2.30 | 5.61 | 21.77 | 0.921** |
| | p | 0.000 | 0.000 | 0.000 | |
| | N (Teachers) | 31 | 30 | 31 | |
| Chemistry | F | 1.77 | 2087.47 | 366.64 | 0.863** |
| | p | 0.009 | 0.000 | 0.000 | |
| | N (Teachers) | 26 | 25 | 25 | |
| 10th Grade English | F | 2.44 | 5.18 | 2.83 | 0.779** |
| | p | 0.000 | 0.000 | 0.000 | |
| | N (Teachers) | 34 | 35 | 35 | |
| 11th Grade English | F | 2.38 | 5.95 | 3.65 | 0.881** |
| | p | 0.000 | 0.000 | 0.000 | |
| | N (Teachers) | 33 | 32 | 32 | |
| Geometry | F | 1.28 | 6.52 | 5.55 | 0.762* |
| | p | 0.123 | 0.000 | 0.000 | |
| | N (Teachers) | 37 | 37 | 37 | |

Note: F-Statistic corresponds to a test of the hypothesis that all the cross-subject teacher effects are zero.

* and ** denote statistical significance at the 5 and 1 percent levels, respectively.

[a] Spearman Correlations of teacher effect estimates between a traditional model that includes cross-subject teachers and a traditional model without cross-subject teachers.

**Table 6. Transition Tables**

| Panel A | | Student Fixed Effects Model | | | | |
|---|---|---|---|---|---|---|
| | | 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 | 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% |
| Panel C | | Traditional Model Including Cross-subject Teachers | | | | |
| Traditional Model | 1 | 68.0% | 20.0% | 12.0% | 0.0% | 0.0% |
| | 2 | 21.7% | 47.8% | 21.7% | 6.5% | 2.2% |
| | 3 | 8.9% | 26.7% | 33.3% | 28.9% | 2.2% |
| | 4 | 4.3% | 2.2% | 21.7% | 47.8% | 23.9% |
| | 5 | 0.0% | 2.3% | 9.3% | 18.6% | 69.8% |