Program on Education Policy and Governance Working Papers Series

Impact of the Teacher Advancement Program on Student Test Score Gains: Findings from an Independent Appraisal

Matthew Springer
Vanderbilt University

Dale Ballou
Vanderbilt University

Art (Xiao) Peng
Vanderbilt University

PEPG 08-14

Preliminary draft
Please do not cite without permission

Prepared for the CESifo/PEPG joint conference
"Economic Incentives: Do They Work in Education"
Insights and Findings from Behavioral Research

CESifo Conference Center
Munich, Germany
May 16-17, 2008

Impact of the Teacher Advancement Program on Student Test Score Gains: Findings from an Independent Appraisal

Matthew G. Springer, Dale Ballou, and Art (Xiao) Peng[1]
Peabody College of Vanderbilt University
National Center on Performance Incentives

This version was prepared for the CESifo and Program on Education Policy conference titled *Economic Incentives: Do They Work in Education? Insights and Findings from Behavioral Research* in Munich, Germany May 16th – 17th, 2008.

Corresponding Author:

Matthew G. Springer
Research Assistant Professor of Public Policy and Education
Director, National Center of Performance Incentives
Peabody College of Vanderbilt University
Peabody #43, 230 Appleton Place
Nashville, Tennessee 37203
615.322.5538
matthew.g.springer@vanderbilt.edu

An earlier version of this paper was presented at the 28th Annual Association for Public Policy Analysis and Management Research Conference in November 2007 and at a research seminar at the Federal Research Bank of New York in December 2007. Readers are advised that result in this paper differ somewhat from those presented in the previous version, following revisions made in response to reviewer comments and suggestions. This version also differs from the paper presented at the National Center on Performance Incentives' 2008 research to policy forum in that this version includes an additional year of achievement data.

[1] This working paper was supported by the National Center on Performance Incentives (www.performanceincentives.org), which is funded by the United States Department of Education's Institute of Education Sciences (R305A06034). We appreciate helpful comments and suggestions from Steve Glazerman, F. Howard Nelson, Tamara Schiff, Gary Stark, and participants at the 28th Annual Meeting of the Association for Public Policy Analysis and Management, a Federal Reserve Bank of New York research seminar, and the National Center on Performance Incentives' 2008 research to policy forum. The authors also wish to acknowledge Kelly Fork, Rebekah Hutton, and Kurt Scheib for their research assistance and the Northwest Evaluation Association for providing data and technical support to conduct our analyses. Any errors remain the sole responsibility of the authors. The views expressed in this paper do not necessarily reflect those of sponsoring agencies or individuals acknowledged.

Abstract

This article presents findings from the first independent, third-party appraisal on the impact of the Teacher Advancement Program (TAP) on student test score gains in mathematics. TAP is a comprehensive school reform model designed to attract highly-effective teachers, improve instructional effectiveness, and elevate student achievement. We use a panel data set to estimate a TAP treatment effect by comparing student test score gains in mathematics in schools that participated in TAP with student test score gains in non-TAP schools. Ordinary least squares estimation reveals a positive TAP treatment effect on student test score gains in the elementary grades, with weaker but still positive point estimates in the secondary grades. When estimation methods control for selection bias, the positive effect remains at the elementary level, but most estimates for grades 6 through 10 turn negative. Our findings are qualified by the lack of information on the fidelity of implementation across TAP schools and on variation in features of TAP programs at the school level.

I. Introduction

A number of school districts and states are instituting performance incentive policies as a potential lever to enhance teacher effectiveness and school productivity. Performance incentive policies also are being used to recruit and retain more effective teachers. These policy innovations are driven, in part, by the fact that existing teacher remuneration practices are not closely related to student performance and schooling outcomes (Hanushek, 2003).[2]

The Teacher Advancement Program (TAP), a comprehensive school reform model providing teachers with opportunity to earn performance pay, has gained considerable attention in recent years. Developed in 1999 by Lowell Milken and other individuals at the Milken Family Foundation (MFF) to attract highly-effective teachers, improve instructional effectiveness, and elevate student achievement, TAP operates in more than 180 schools in 15 states and the District of Columbia. In the aggregate, there are approximately 5,000 teachers and 60,000 students in TAP schools across the United States (Milken Family Foundation, 2007).

TAP also figured prominently in the 2006 announcement of Teacher Incentive Fund (TIF) grantees. TIF, a federally-enacted direct discretionary grant program, funds development and implementation of principal and teacher performance incentive pay programs. Of the approximate $240 million awarded during fall 2006, $88.3 million (36.80 percent) went to districts and states that proposed to implement TAP.

Evaluations of TAP report generally positive findings. Studies have found positive effects on teachers' and schools' value added (Solmon, White, Cohen, and Woo, 2007) and student achievement gains (Schacter, et al., 2003; Schacter, Thum, Reifsneider, and Schiff, 2004). Solmon, White, Cohen, and Woo (2007) found furthermore that an equal or higher percentage of TAP schools make Adequate Yearly Progress under No Child Left Behind than other schools in their respective states, despite larger concentrations in TAP schools of students qualifying for free and reduced price lunch. All of these studies, however, have been conducted in partnership with MFF or the National Institute for Excellence in Teaching (NIET),[3] causing some to raise concerns about the independence of evaluators from stakeholders.

To our knowledge, the research reported here represents the first independent, third-party assessment of TAP. We use a panel data set to estimate a TAP treatment effect by comparing student test score gains in schools that participated in TAP with student test score gains in non-TAP schools. Our data set includes roughly 1,200 TAP and non-TAP schools from two states over a five-year period comprising the 2002-03 to 2006-07 school years. 35 schools implemented TAP at some point during this period in the two states under study.[4] Student test

[2] See Podgursky and Springer (2007) for a comprehensive review of teacher performance pay.

[3] In May 2006, it was announced Teacher Advancement Program Foundation developed into the National Institute for Excellence in Teaching (NIET) to further its mission of improving teacher quality.

[4] Our sample includes 32 TAP schools. We do not have student-level test score information on one TAP school and exclude two schools that abandoned the program shortly after adoption.

scores are available in mathematics two times per year in 2nd through 10th grades, allowing for a fall-to-spring gain score as the outcome of interest.

In our models TAP is usually represented by a simple binary indicator. While the coefficient on this variable is meant to measure the effect of TAP on student test score gains, as always in this kind of research, other factors might be confounded with TAP. This is of particular concern as TAP schools are self-selected. Thus, one might expect distinctive outcomes in TAP schools even in the program's absence. We address this concern in three ways. First, we include a variety of school and student characteristics in the model. Second, we use a school fixed effects estimator to control for unobserved characteristics of schools that may explain selection into TAP as well as achievement. Third, we use a two-step selection-correction estimator (as in Heckman, 1979) to remove selection bias.

The results we obtain when controlling for selection on unobservables stand in contrast to prior studies. Ordinary least squares estimation reveals a positive TAP treatment effect on student test score gains in mathematics in the elementary grades, with weaker but still positive point estimates in the secondary grades. When estimation methods control for selection bias, the positive effect remains at the elementary level, but most estimates for grades 6 through 10 turn negative and some significantly so.

While our study is the first to estimate a TAP treatment effect that controls for selection, it is important to acknowledge limitations of our study. The sample of TAP schools is small. We also lack information both on the fidelity of implementation and on variation in features of TAP programs at the school or grade level (e.g., minimum and maximum bonus sizes, percent of teachers voting in favor of TAP adoption, and so forth).[5] Furthermore, it is unknown whether our sample is representative of other schools and locations that will be or are actually implementing TAP across the nation. Finally, TAP is designed to attract highly-effective teachers, improve instructional effectiveness, and elevate student achievement. This study directly tests only the latter.

In the next section, we provide more detail on TAP. We follow with a review of relevant literature in section III. In sections IV and V we describe our analytic strategy and our data and sample, respectively. Findings are presented in section VI. Section VII discusses results and explores some alternative explanations to our findings.

II. Description of the Teacher Advancement Program

TAP's design has four components: 1) multiple career paths; 2) ongoing applied professional growth; 3) instructionally-focused accountability; and 4) performance-based compensation.[6]

[5] TAP typically requires teachers to vote whether to adopt the program at their school. However, in some instances, TAP is implemented without a vote.

[6] Both the Milken Family Foundation and the National Institute for Excellence in Teaching have produced a large number of resources about the TAP reform model. More details can be found at www.talented.teachers.org.

Multiple career paths create opportunities for teachers and specialists to advance professionally without leaving the classroom by becoming a career teacher, master teacher, or mentor teacher. Ongoing applied professional growth is encouraged by providing teachers collaboration time to develop and implement new instructional practices and curricula focused on increasing student learning. Professional growth occurs both individually and in groups of teachers (grade-level or subject-level). TAP-identified mentor and master teachers are engaged to facilitate discussion and planning and conduct classroom observations.

Table 1 provides a summary of TAP's assessment and compensation system for career, master, and mentor teachers. Teacher knowledge, skills, and responsibilities is the first indicator in TAP's assessment and compensation system. Fifty percent of a teacher's performance award is contingent on classroom observations. Four to six observations are done by certified evaluators who are master teachers, mentor teachers, or administrators. All evaluators conduct observations separately, not as a group. Teachers are hired as career, mentor, and master teachers on a competitive basis which includes formal interviews, classroom observations, evidence of instructional expertise, demonstrated leadership, and expertise in adult learning to name a few.

Teacher value added is the second indicator in TAP's assessment and compensation structure. Thirty percent of a teacher's performance award is based on value-added measurement of gains the teacher produces in his/her classroom's achievement. Based on the number of standard errors a teacher's estimated value added falls above or below the average estimate in the state,[7] teachers are rated as level one through level five.[8] If a teacher does not have direct instructional responsibilities for a group of students, or a teacher works in a non-tested subject or grade, this component of TAP's assessment and compensation structure is shifted to school-wide achievement gains.

School-wide achievement is the final indicator in TAP's assessment and compensation structure. Twenty percent of a teacher's individual performance award is dependent upon school-wide achievement. Like the classroom achievement score, school-wide performance is determined by how many standard errors above (or below) a school's value-added estimate falls from the average school-wide effect estimate in the state.[9] The school wide achievement award is equally distributed to all teachers.

NIET recommends allocating a minimum of $2,500 per teacher to a school's performance award fund. The performance award fund is then apportioned based on the ratio of the number of teachers in each career level (i.e., career, mentor, or master teacher) to the total number of teachers in the school. NIET's recommended compensation structure enables teachers to earn

[7] In select instances, the reference group is not the average teacher value-added effect estimate in the state.

[8] Levels are defined as follows: level one – two standard errors or more below the state average; level two—one standard error below the mean; level three – between one standard error below and one standard error above; level four—one standard error above the mean; level five—two standard errors above the mean.

[9] In select instances, the reference group is not the average school value-added effect estimate in the state.

anywhere from zero to $12,000 per year, though there can be variation across sites. Performance is judged against an absolute standard so as not to create competition among teachers for a fixed amount of awards. NIET estimates that the program costs approximately $400 per student, or about six percent of a school's operating budget.

III. Review of Relevant Literature

We know of only three studies that have analyzed the impact of TAP on student outcomes, all of them conducted or commissioned by MFF or NIET.

The first evaluation of TAP reported by MFF analyzed student achievement growth in four Arizona TAP schools relative to a similar set of non-TAP schools using data from the 2000-01 and 2001-02 school years (Schacter et al., 2002).[10] Schools were compared on the basis of a "gap reduction model." This model quantifies school performance using the relative decrease in the distance between a school's baseline percentile rank score on the Stanford 9 and a fixed achievement target defined by MFF. The formula for school *j* in year *t* appears in equation (1).

$$Y_{jt} = \frac{A_{jt} - A_{j,t-1}}{T - A_{j,t-1}} \qquad (1)$$

T was established as the 85th percentile for all schools in the state.[11] Clearly, two schools that made the same absolute progress ($A_{jt} - A_{j,t-1}$), but started with different initial values of $A_{j,t-1}$, will fare differently by this metric. Unfortunately, in selecting the sample of comparison schools, schools were not perfectly matched on $A_{j,t-1}$. In one case, the difference in $A_{j,t-1}$ between a TAP school and its matched comparison school was 17.5 percentile points. Moreover, the metric is perverse: a school that started just below the 85th percentile will achieve a much larger proportional reduction than a school with the same value of $A_{jt} - A_{j,t-1}$ but a value of $A_{j,t-1}$ farther from T. Thus, gains count most in schools that were initially doing best.

The same metric was used in a second study that examined 12 TAP schools in two states – six in Arizona and six in South Carolina (Schacter, Thum, Reifsneider, and Schiff, 2004). The gap reduction technique was used to evaluate TAP in the Arizona sample.[12] Differences in initial

[10] The TAP school sample included a total of 1,114 student observations during the 2000-01 school year and 1,277 observations during the 2001-02 school year. The comparison group sample included a total of 2,009 students during the 2000-01 school year and 1,372 students during the 2001-02 school year. The authors do not explain the 31.7 percent reduction in comparison group students, even though the number of comparison group schools remained constant across years. Furthermore, the matching strategy was contingent upon districts supplying MFF with student level data from matched comparison schools. If a school district was unwilling to provide data to MFF, MFF moved to the next best matched school until necessary data for comparison group schools was obtained.

[11] All TAP schools and schools in the matched comparison group were below the 85th percentile.

[12] A different approach was used in South Carolina because individual student achievement results were reported as performance levels (e.g., Below Basic, Basic, Proficient, and Advanced).

percentile ranks ($A_{j,t-1}$) between TAP schools and control schools ranged from -13 to 12 percentile points for the Arizona sample. In 57 percent of the individually matched cases, the non-TAP school had the larger denominator. In these instances, the average difference in $A_{j,t-1}$ was 7.44 percentile points.

The most recent study reported by MFF (Solmon, White, Cohen, and Woo, 2007), compared TAP to non-TAP teachers and schools using the SAS EVAAS methodology developed by William Sanders.[13] TAP teachers outperformed comparison group teachers 63 percent of the time, while TAP schools outperformed comparison group schools 57 percent of the time in mathematics and 67 percent of the time in reading.[14]

However, the Solmon and colleagues' (2007) report contains no discussion of how the comparison sample was chosen. It appears the sample of control group schools was one of convenience; that is, all non-TAP schools with a valid EVAAS score that are located in same state as a TAP school were included in the control group.

Finding an appropriate comparison group for TAP schools is a critical issue. The process schools go through to select into TAP raises a strong likelihood that these are distinctive schools. For example, Glazerman et al. (2006: p. 9) remarked that:

> Selection as a TAP school occurs via a competitive process. Typically, a state department of education or district superintendent invites schools to learn about TAP and apply for the program. Candidate TAP schools also need to show an ability to provide financial support for the program. Ultimately, selection as a TAP school depends on the ability of the schools to implement, fund, and sustain the program, as well as on demonstrated faculty support.

Our review of relevant studies on the impact of TAP on student outcomes highlights several limitations that warrant further research on the topic. First, all evaluations to date have been conducted by MFF. Second, two of the studies use an idiosyncratic metric to compare TAP and non-TAP schools without ensuring the two groups are treated even-handedly. Finally, none of these studies appears to identify an appropriate comparison group of schools.

IV. Analytical Strategy

---

[13] This study was based on a large number of observations than previous studies reported by the MFF. Their sample included a total of 2,947 teachers and 346 schools from six states. TAP teachers and schools comprised roughly 21 percent and 18 percent of the sample, respectively.

[14] Solmon et al. (2007) also analyzed adequate yearly progress results for the 2004-05 and 2005-06 school years in TAP schools as compared to the statewide average. In terms of adequate yearly progress comparison, the authors found that, "In most cases an equal or higher percentage of TAP schools in the six states make AYP than all schools in their states, despite TAP schools having more students receiving free or reduced-price lunch" (p.7).

This research seeks to estimate a TAP treatment effect by comparing student test score gains in schools that participated in TAP with student test score gains in non-TAP schools. Our general model of student achievement gains is:

$$\Delta Y_{ijt} = \mathrm{X}_{it}\beta + W_{jt}\delta + \alpha T_{jt} + u_j + u_{it} + u_{jt} + u_{ijt} \qquad (2)$$

where ΔY_{ijt} is the fall-to-spring test score gain in reading, mathematics, or language arts for student *i* attending school *j* in year *t*; X_{it} is a vector of student characteristics; W_{jt} is a vector of school characteristics; T_{jt} is a dummy variable indicating whether school *j* had become a TAP school by year *t*; u_j is the influence of unobserved time-invariant characteristics of school *j*; u_{it} is a student by year effect; u_{jt} is a school by year effect; and u_{ijt} is an independent error.[15] Denote $v_{ijt} = u_j + u_{jt} + u_{it} + u_{ijt}$. We assume X_{it} and W_{jt} are uncorrelated with v_{ijt}. This is an unrestrictive assumption, given that we place no causal interpretation on the coefficients of these variables.

To develop a model of TAP participation, we begin by examining participation patterns over time. These are summarized in Table 2. Participation in TAP has increased from 10 schools in 2002-03 school year to 50 schools in the 2007-08 school year. This is consistent with a model in which resistance to participation is high but declining. Only two schools in these states have abandoned TAP, suggesting that schools do not typically reconsider this decision, or that the factors that might lead them to reverse the decision have a negligible variance over time: once in, schools stay in TAP. These facts shape our specification of the participation model, as follows:

$$\mathrm{T}_j^* = Z_{j0}\gamma + e_j \qquad (3)$$

$$\mathrm{T}_{it} = 1 \text{ if } \mathrm{T}_j^* > c_s, \qquad \mathrm{s} \geq \mathrm{t} \qquad (3a)$$

$$\mathrm{T}_{it} = 0 \text{ if } \mathrm{T}_j^* < c_s, \qquad \mathrm{s} \leq \mathrm{t} \qquad (3b)$$

where Z_{j0} are observed characteristics of school *j* in baseline year 0 (2001-02); c_s is the perceived "cost" of participating in TAP as of year s, monotonically decreasing in s; and e_j is an unobserved school effect.

T* is a latent variable representing a school's disposition toward the TAP program. By assumption schools do not change in this regard, at least over our sample period: thus, the schools that are most favorably disposed toward TAP at the outset remain most favorably disposed. However, actual participation rates start out quite low, possibly due to teachers'

[15] The model omits a time-invariant student effect. Because we estimate separate equations for each grade level, only a few students (those retained in a grade) appear more than once in any estimation sample.

uncertainty and suspicions about the nature and effect of the program. Over time, these perceived "costs" of participation have declined, leading more of the schools with high values of T_j^* to join. We estimate c_s as the coefficient on a dummy variable for year s in an equation predicting the year in which a school adopts TAP. Note that once a school adopts TAP it continues to be identified as a TAP school for our purposes, even if it subsequently drops out of the program. While this does some violence to reality, allowing for separate treatment of drop-outs would considerably complicate the model with little benefit, given that so few schools have reversed this decision.

The assumption that participation depends only on characteristics in a baseline year may be unduly restrictive. A more conventional specification would allow for T* to vary over time; thus, $T^*_{jt} = Z_{jt}\gamma + e_{jt} + e_{j.}$ However, most of the candidate variables for Z change slowly over time, and as just noted, the variance of e_{jt} must be quite small relative to e_j or there would be more reversals of the participation decision. Nonetheless, it is possible that teachers decide to join TAP following an upturn in test scores that encourages them to think they will qualify for future bonuses. Given the noisiness of test results, this expectation is not likely to be met. As scores revert back to more normal levels, it will appear that TAP has been ineffective—a negative bias in the estimated treatment effect caused by mean reversion. We look for evidence of such a bias in sensitivity tests below.[16]

There are some natural restrictions on the relationships among the error terms that appear in the model. First, as the u_{jt} are deviations about u_j, $\text{cov}(u_j, u_{jt}) = \text{cov}(e_j, u_{jt}) = 0$. In addition, as we restrict the sample to students who remain in the same school for an entire year, the mean of u_{it} over students in a given school in a given year is absorbed in u_{jt}. Thus, $\text{cov}(u_{it}, u_{jt}) = \text{cov}(u_j, u_{it}) = \text{cov}(e_j, u_{it}) = 0$ as well.[17] We leave open the possibility that u_j is correlated with e_j; indeed, it is just this possibility that gives rise to the selection effects discussed above. Thus, how we estimate this model depends on additional assumptions about the relationships among these unobservables.

Assumption 1: Ordinary Least Squares Regression

[16] It is also possible that the reverse occurs, and that schools are more likely to join TAP after a bad year, when under pressure from district or state officials to improve. As already noted, a model that allows for transitory changes in test results (or anything else) to influence TAP participation must explain why so few participation decisions are reversed. However, there may be some inertia among participating schools that discourages teachers from revisiting a decision: reluctance to admit a previous error, aversion to reopening a contentious debate, pressure from higher up, and so forth.

[17] Despite $\text{cov}(u_{jt}, e_j) = 0$, it is still possible for u_{jt} to be correlated with T_{jt}, if students and their parents use the fact that a school participates in TAP as a signal of quality. We consider this quite unlikely, given how little the general public knows about TAP. In the school of education where we work, there are many persons who do not know what TAP is.

Assume that u_j is uncorrelated with e_j; that is, once we have conditioned on X and W, there are no selection effects. This assumption is plausible if W is a large set of school characteristics that includes Z as well as any other pre-existing differences between TAP and non-TAP schools that affect Y. Under Assumption 1, ordinary least squares estimation of (2) yields an unbiased estimate of α. Standard errors are corrected for clustering of students within schools.

Assumption 2: School-Fixed Effects Estimator

Alternatively, assume cov(e_j, u_j) $\neq 0$. Differencing out u_j through the inclusion of school fixed effects yields an unbiased estimate of α, which is identified through variation in outcomes over time within schools that enter the TAP program.

Assumption 3: Ordered-Probit Selection Correction Model

Finally, assume (ν_{ijt}, e_j) have a bivariate normal distribution with covariance $\sigma_{ve} \neq 0$.
We estimate equation (3) as an ordered probit model, obtaining an estimate of the expectation of e_j conditional on the decision to adopt TAP in year s. For schools that adopt TAP in year 1 of our sample (there were no TAP schools prior to that year in these states), this is $\mathrm{E}\left(e_j \middle| \mathrm{T}_j^* > c_1\right)$; for schools joining in the second year, this is $\mathrm{E}\left(e_j \middle| c_1 > \mathrm{T}_j^* > c_2\right)$, and so forth. For schools that remain outside TAP throughout the period, the term is $\mathrm{E}\left(e_j \middle| c_6 > \mathrm{T}_j^*\right)$.[18] We introduce this expectation into the structural achievement equation, which becomes:

$$\Delta Y_{ijt} = Y_{ijt} - Y_{ij,t-1} = \mathrm{X}_{it}\beta + W_{jt}\delta + \alpha \mathrm{T}_{jt} + \beta_\lambda \hat{\lambda}_{js} + \varepsilon_{ijt} \qquad (4)$$

where $\hat{\lambda}_{js}$ is our estimate of $\mathrm{E}\left(e_j \middle| c_{s-1} > \mathrm{T}_j^* > c_s\right)$ and β_λ is an estimate of σ_{ve}. The new error term, $\varepsilon_{ijt} = \nu_{ijt} - \beta_\lambda \hat{\lambda}_{js}$, is asymptotically uncorrelated with ΔY_{ijt} by construction. Estimation of (4) by least squares yields a consistent estimate of α.

β_λ is weakly identified through the fact that $\hat{\lambda}_{js}$ is a non-linear function of the variables in the participation model. Stronger identification requires that there be at least one variable in the participation model that is not in the structural achievement equation. In our case, we obtain this identification through the year-of-adoption effects, c_s. Note that we can still include year effects in the achievement equation without affecting identification, as $\hat{\lambda}_{js}$ is a function of the year in which a school joined TAP, not the current year.

[18] The first year for which we have achievement data is 2002-03 school year, the final school year 2006-07. However, we know which schools have adopted TAP through the 2007-08 school year. Under the assumptions above, all of these adoption decisions contain information helping to identify σ_{ve}. Thus the year-of-adoption index, s, ranges from 1 to 6.

This identification strategy is not robust to all alternative specifications of the achievement model. If there are cohort TAP effects – that is, how a school responds to TAP depends on the year in which it adopts TAP – and we seek to estimate them by including a TAP-cohort interaction in the model, then our identification strategy does not identify selection effects, as all schools in the same cohort with the same value of Z also have the same value of $\hat{\lambda}_{js}$. However, we are aware of no reason to suspect a priori that there are cohort effects, and their existence would appear to be a distinctly second-order concern compared to estimation of a TAP main effect.

Our model of student achievement gains includes student gender and race/ethnicity. School level controls include the percentages of students by race/ethnicity, the percentage of students eligible for the free price lunch program, average teacher salary, the student teacher ratio, and percentages of students scoring basic and below basic under the state's accountability program in 2001-02 school year, the last year prior to the introduction of TAP. We include controls for the percentage of students tested, in case mean scores were affected by the exclusion of some students from testing, but these variables were never significant.[19]

Mean gains vary substantially by grade. When pooled across all years, the average fall-to-spring gain score ranges from 2.66 to 13.79 points. The magnitude of the fall-to-spring gain decreases monotonically from low to high grades, with the average 8th grade gain less than half the size of the average 3rd grade gain. Effects of covariates may also differ by grade. We therefore estimate separate equations for each grade.

All models also include state by year effects to control for changes in the test, changes in how well aligned the test is with curricula, and student cohort effects. These variables also control for changes in the composition of the sample, as described in the next section under "sample".

Participation in TAP is a function of school characteristics in the baseline school year (2001-02): student-teacher ratio, average teacher salary, percentages of minority students and students eligible for the free price lunch program, and percentages of students scoring basic or below basic in English and in mathematics.

V. Data and Sample

Data

The primary data for this study are drawn from the Northwest Evaluation Association's Growth Research Database (GRD). GRD contains longitudinal student test score results from approximately 2,200 school districts in 45 states. The NWEA test is administered twice per year, allowing for construction of a fall-to-spring gain score for each student. All scores reference a

[19] The National Center for Education Statistics has not released the Common Core of Data (CCD) for the 2006-07 school year. We therefore imputed values at the school-level for the following variables for the 2006-07 school year in both State A and State B: average teacher salary, student teacher ratio, percent Asian, percent Hispanic, percent Black, and percent free price lunch eligible.

single cross-grade, equal-interval scale developed using a one parameter Raasch model (Kingsbury, 2003). GRD also contains a limited number of student characteristics, notably race/ethnicity, gender, grade, and date of birth.

We supplement GRD with publicly-available school report card data from state department of education websites and information from the National Center for Education Statistics' Common Core of Data (CCD). State school report cards include information on average teacher salary, student attendance, student-teacher ratio, and aggregate school performance on a state's high-stakes assessments. CCD data contain fiscal and non-fiscal information on schools and school districts, including students and school personnel.

Sample

Our sample includes roughly 1,200 schools from two states over a five year period comprising the 2002-03 to 2006-07 school years. The number of TAP schools in our sample increases from six in the 2002-03 school year to 32 in the 2006-07 school year, while the number of TAP student observations with valid fall and spring test scores in mathematics rises from 663 in 2002-03 school year to 8,496 in the 2006-07 school year.[20] Although all of the TAP schools in these states had contracts with NWEA, the NWEA tests are not the high-stakes exams used to determine which teachers earn bonuses. Thus, there is no particular reason for teachers to "teach to" the NWEA exams or otherwise manipulate scores on these tests. Indeed, because these exams are used mainly for diagnostic and formative assessments, teachers have every reason to see that an accurate reading is obtained from these exams on all of their students.

The number of districts contracting for testing services with NWEA increased at the same time as the number of TAP schools. As a result, the set of comparison schools varied across years. To illustrate, in Table 3 we present the number of TAP and non-TAP schools by state and year for grades 3 and 8. We also present average mathematics scores (spring) and fall-to-spring gains. Gain scores in second grade trended upward in the non-TAP schools. The opposite trend prevailed in 8th grade. The same contrast is evident between other elementary and secondary grades. Though trends are less clear in the much smaller number of TAP schools, there are some sizeable differences across years that could be related to the entry of new TAP schools. These year-to-year differences are controlled for in our model's state-by-year effects.

Testing dates vary considerably by school. Average time elapsed between fall and spring testing is about 212 calendar days in TAP schools, with a standard deviation of slightly more than three weeks. Average time elapsed for non-TAP schools is about 194 calendar days, with a standard deviation of 27.78 days. Recognizing gains are positively correlated with time elapsed since the previous test administration, we include this variable in all model specifications.

Because familiarity with a test is generally associated with rising scores, we expect that the more frequently a school has used NWEA tests in the past, the higher scores will be. We define

[20] As shown in Table 2, there were 10 TAP schools in these states in 2002-03, but for 4 of these schools our data start in a later year.

NWEA cohort based on the year that a school first begins using NWEA tests. Dummy variables for the 2003, 2004, 2005, and 2006 cohorts are included in the model.

Table 4 summarizes the means and standard deviations of key school and student variables. TAP schools in State A have a greater percentage of Black students (60.59 vs. 37.34) compared to other schools in the state. The same holds true for free lunch status students (60.99 vs. 43.08) and students scoring below basic on the high-stakes assessment (43.15 vs. 28.79). Other school and student covariates tend to be very similar between TAP and non-TAP schools in State A.

TAP schools in State B tend to have a greater percentage of Hispanic students (48.36 vs. 31.67) when compared to the state-wide average. Non-TAP schools in State B have higher percentages of Black (3.53 vs. .47) and free lunch status (32.30 vs. 22.77) students and more students scoring below basic on the state high-stakes assessment (21.20 vs. 16.01). Other school and student covariates tend to be very similar between TAP and non-TAP schools in State B.

Insert Table 4 Here

Table 4 also summarizes student test score information. TAP schools have modestly larger test score gains when compared to the average test score gain in their respective state. TAP schools in both State A and State B have over three weeks more time between the fall and spring administration of the NWEA test. Nearly all students were tested in both fall and spring, irrespective of whether they attended a TAP school or not (99.02 vs. 99.22).

VI. Results

Table 5 reports OLS regression estimates with robust standard errors to correct for clustering of students within schools. At all grade levels, there is a positive association between TAP and a student's fall-to-spring test score gain. Although the positive coefficients for the 7th through 9th grade-level models fail to attain conventional levels of statistical significance, the dominant impression is positive. The largest effect is in 2nd grade (2.74 points), though differences of between 1 and 2 points are more common. Given that the standard deviation of fall to spring test score gains is approximately 7 to 8 points in the elementary and middle school grades, statistically significant effect sizes range from 12 (grade 3) to 34 percent (grade 2).

Insert Table 5 Here

The bias that results from using non-randomly selected samples to estimate a TAP treatment effect is a real concern. As explained above, we first use a school fixed effects estimator to control for unobserved characteristics of schools that may explain selection into TAP as well as achievement. We then implement a two-step selection-correction estimator as a second way of controlling for selection bias.

Table 6 reports estimates from the school fixed effects models. After differencing out time invariant school characteristics that may explain selection into TAP, a clear division arises between elementary and secondary grades. At the elementary level (grades 2 through 5), the

TAP effect continues to be positive, although coefficients are generally somewhat smaller than previously reported using OLS regression and significant in some instances, not all. However, coefficients in the 6th, 7th, 9th, and 10th grade-level models are now negative and statistically significant in 7th, 9th, and 10th grades. The coefficient for 8th grade-level model remained statistically insignificant.

Insert Table 6 Here

It may be the case that our school fixed effects results are sensitive to the fact that 11 TAP schools, equivalent to 40 percent of our TAP school sample, do not change TAP status during the sample period. These schools make no contribution to the estimated TAP effect in the school fixed effects model. We checked the possibility that this accounts for the difference between the OLS and fixed effects estimates by dropping these TAP schools from the sample and re-estimating the OLS regressions. Results were very similar to those reported in Table 5, indicating that the difference between OLS and fixed effects results was due to the inclusion of school effects, not to the loss of TAP school observations.

Table 7 reports results from the two-step selection correction model. In the first-step ordered probit model (not shown), the only significant school-level variables were the percentage of minority students and the percentage of students scoring at the basic level in English. The positive coefficient on percentage minority suggests that schools have been more likely to join TAP when under pressure to raise achievement of traditionally underperforming groups. The second result is more difficult to interpret, given that an increase in the percentage of students at the basic level implies a decrease at both extremes (below basic and proficient). The year of adoption effects (the c_s in equation 3) were all significant.

Results in the achievement equation are qualitatively similar to school fixed effect estimates reported in Table 6. Estimated TAP effects in the elementary grades are significantly positive in grades 2, 4, and 5. This is not the case in the 6th grade-level model and higher, and in 10th grade, the coefficient is significantly negative.

Insert Table 7 Here

The coefficients on $\hat{\lambda}_{js}$ confirm the presence of selection bias in the upper grades. In the 6th, 7th, 9th, and 10th grade-level models, these coefficients are positive and statistically significant, indicating a tendency for schools with above average outcomes to adopt TAP. Additionally, there are no instances of a negative TAP selection effect.

Both the fixed effects estimates and the selection-correction estimates rest on the assumption that selection into TAP is a function of time-invariant characteristics of the school. If this is not the case and participation is influenced by transitory changes in test scores, estimated TAP effects could be biased by regression to the mean. We test for this by including an indicator for new TAP schools—schools in their first year in the program. The new-TAP indicator could also pick up implementation problems and first-year bugs in starting up. Results are displayed in Table 8.

The new-TAP indicator is statistically significant and negative in the OLS estimates in the 9th grade-level model, and all other estimates on new-TAP are not different from zero. When using the two-step selection-correction model new-TAP is marginally significant and positive in the 6th and 7th grade-level models, while being statistically significant and negative in the 9th grade-level model. Positive and significant coefficients are found in the 6th and 7th grade-level models when using a school fixed effect estimator. The only negative and significant coefficient is found in 2nd grade in the school fixed effect model.

As displayed in Table 8, the sign on the TAP coefficients are similar to those shown in previous tables. They tend to be positive and significant in the OLS estimates in the 2nd through 6th, while estimates on the TAP coefficient in the 7th through 9th grade-level models are not different from zero at conventional levels. The estimate in the 10th grade-level model is positive and significant, but the sign changes and is not different from zero in the school fixed effect and two-step selection-correction approaches. We do find a positive and significant effect in the 2nd, 4th, and 5th grade-level fixed effect models, and negative and significant effects in grades 6, 7, and 9. We see a similar pattern in the selection-correction model; however, estimates in grades 6 and 9 are no longer statistically significant. It appears TAP elementary schools are more effective in select grades in years following implementation, while findings in the upper-grades are mixed.

Insert Table 8 Here

VI. Conclusion

This study has presented findings from study of the impact of the Teacher Advancement Program (TAP) on student test score gains in mathematics. We have used a panel data set to estimate a TAP treatment effect by comparing student test score gains in schools that participated in TAP with student test score gains in non-TAP schools. Ordinary least squares estimation revealed a significant, positive TAP treatment effect on student test score gains in the grades 2 through 6 and 10. Point estimates were still positive in grades 7 through 9, though no estimates were statistically different from zero.

When controlling for selection into TAP, either through the use of a school fixed effects estimator or a two-step selection correction model, the estimates in grades 6 through 10 typically turn negative, frequently significantly so. While the tests that furnished the data for this study are not the high-stakes exams on which teacher bonuses are based, this does not account for the difference between our findings and those of earlier investigators, inasmuch as we reproduce their results when estimating an achievement equation using OLS regression techniques. It is only when we control for selection into TAP that we obtain markedly different findings in the higher grades.[21]

[21] Our estimates remained qualitatively similar when using fall test score as an independent variable. We also estimated all models with one TAP school dropped from the regression dataset to check whether an outlier school or grade was influencing our estimates. This was done 32 times, once for each TAP school. All estimates remained qualitatively similar to those reported.

Given the small number of schools in this study, the failure of TAP to produce positive outcomes at the middle and high school level may have been due to idiosyncratic failures in program implementation in these schools. However, the explanation is not simply that the secondary schools in this study were not as effective as the elementary schools (as a result, say, of poor leadership). We re-estimated the model by OLS regression after including an indicator for pre-TAP status (taking the value 1 in the years before a school joined the program and zero in all other instances). The sign on the coefficient on pre-TAP is generally positive but insignificant in the elementary grades (Appendix Table 1). It is positive in all of the higher grades, significantly so in all but one (grade 8), and larger in magnitude than in the lower grades. These estimates suggest TAP middle and high schools were outperforming their non-TAP comparison group before they started implementing the TAP program.

It is possible that the way teachers and schools respond to TAP does not work as well in the middle and high school grades. Exhorting students to try harder on tests, for example, may succeed with younger children who are eager to please their teachers, but not with adolescents who are more likely to differentiate between low- and high-stakes assessments. It is also possible that TAP incentives work best in schools where most teachers are doing essentially the same job, but that differences in the way instructors of core subjects are treated from other teachers produces acrimony and a breakdown in teamwork at the secondary school level. We do not know that either of these explanations is correct, and offer them only as illustrative of differences between elementary and secondary schools that could account for our findings. The latter of which may be the less plausible explanation given TAP's emphasis on collaboration across grades and disciplines, and the fact that the great majority of schools remain TAP schools once they opt into the program.

It is important to acknowledge several methodological limitations. The sample of TAP schools is small. The numbers of student test score observations in 2nd, 9th and 10th grades are far fewer than other grades in our data panel. We also lack information on the fidelity of implementation and on variation in features of TAP programs at the school level (e.g., minimum and maximum bonus sizes, percent of teachers voting in favor of TAP adoption, and so forth). Furthermore, it is unclear whether our sample is representative of other schools and locations across the United States that will be or are actually implementing TAP.

Finally, we have investigated only one aspect of the TAP reform model – the impact of TAP on student achievement. TAP is a comprehensive school reform model designed to attract highly-effective teachers, improve instructional effectiveness, and elevate student achievement. While student achievement is ultimately the outcome of interest, we have not determined whether TAP has altered teacher recruitment and retention or instructional practices. Clearly, these are important components of TAP and important areas for future research.

References

Glazerman, S., Silva, T., Addy, N., Avellar, S., Max, J., McKie, A., Natzke, B., Puma, M., Wolf, P., and Greszler, R.U. (2006). Options for Studying Teacher Pay Reform Using Natural Experiments. Washington, DC: United States Department of Education's Institute of Education Sciences.

Hanushek, E.A. (2003). The Failure of Input-Based Resource Policies. *The Economic Journal*, 113, F64-F68.

Heckman, J. (1979). Sample Selection Bias as a Specification Error. *Econometrica*, 47: 153-161.

Kingsbury, G.G. (2003). A Long-Term Study of the Stability of Item Parameter Estimates. Paper presented at the annual meeting of the American Educational Research Association. Chicago, IL.

Podgursky, M. and Springer, M.G. (2007). Teacher Performance Pay: A Review. *Journal of Policy Analysis and Management*, 26(4), 909 – 949.

Schacter, J., Schiff, T., Thum, Y.M., Fagnano, C., Bendotti, M., Solmon, L., Firetag, K., and Milken, L. (2002). The Impact of the Teacher Advancement Program on Student Achievement, Teacher Attitudes, and Job Satisfaction. Santa Monica, CA: Milken Family Foundation.

Schacter, J., Thum, Y.M., Reifsneider, D., and Schiff, T. (2004). The Teacher Advancement Program Report Two: Year Three Results from Arizona and Year One Results from South Carolina TAP Schools. Santa Monica, CA: Milken Family Foundation.

Solmon, L.C., White, J.T., Cohen, D., and Woo, D. (2007). The Effectiveness of the Teacher Advancement Program. Santa Monica, CA: Milken Family Foundation.

Milken Family Foundation (2007). More than 60 Schools Join Teacher Advancement Program: TAP Enjoys Greatest Growth Yet, Gaining Ground in Urban Center like Philly, Chicago. Accessed September 29, 2007 from http://www.talentedteachers.org/newsroom.taf?page=tapnews_article291.

Milken Family Foundation (2004). Teacher Advancement Program: Teacher Evaluation and Performance Award Guide. Santa Monica, CA: Milken Family Foundation.

TABLE I: SELECT CHARACTERISTICS OF TAP'S ASSESSMENT AND COMPENSATION SYSTEM

Assessment Components	Career Track		
	Career/Specialist Teacher	**Mentor Teacher**	**Master Teacher**
Knowledge, Skills and Responsibilities			
Evaluators	Mentor review; Self review; Master teacher review; Administrator review	Mentor review; Self review; Master teacher review; Administrator review	Teachers' review; Self review; Administrator review
Measurement Instruments	Portfolio documentation; Observation; Interview process	Portfolio documentation; Observation; Interview process	Portfolio documentation; Observation; Interview process
Unit of Analysis	Teacher	Teacher	Teacher
Percentage of award pool designated for this component	50%	50%	50%
Teacher Value-Added			
Measurement Instrument	Standardized assessment	Standardized assessment	Standardized assessment
Unit of Analysis	Classroom	Classroom	Classroom
*Percentage of award pool designated for this component**	30%	30%	30%
School Value-Added			
Measurement Instrument	Standardized assessment	Standardized assessment	Standardized assessment
Unit of Analysis	School	School	School
*Percentage of award pool designated for this component**	20%	20%	20%
Career Track Bonus			
Amount	$0	$2,500 - $4,500	$6,000 - $12,000

* If a teacher does not have direct instructional responsibilities for a group of students, or a teacher works in a non-tested subject or grade, this assessment component is shifted to school achievement gains. In these instances, the percentage of award pool designated for the school value-added component is 50 percent for that teacher.

SOURCE: Adapted from Milken Family Foundation (2004). *Teacher Advancement Program: Teacher Evaluation and Performance Award Guide* . Santa Monica, CA: Milken Family Foundation, p. 7.

TABLE II: TAP PARTICIPATION, STATES A AND B COMBINED

	Joined	Abandoned	Total
School Year			
2002-03	10	…	10
2003-04	5	1	14
2004-05	9	0	23
2005-06	4	1	26
2006-07	7	0	33*
2007-08	18	0	51

* Our sample includes a total of 32 TAP schools because we do not have access to student level test score information for one TAP school.

TABLE III: SAMPLE MAKE-UP

Non-TAP Schools	Year	Grade 3				Grade 8			
		No. Schools	No. Students	Mean Spring Score	Mean Gain	No. Schools	No. Students	Mean Spring Score	Mean Gain
State A	*2002-03*	57	3395	204.4	9.2	19	3274	236	4.6
	2003-04	129	9835	203.8	9.8	59	10177	235.2	4.7
	2004-05	299	21934	203.7	9.9	133	21958	234.2	3.9
	2005-06	395	31213	203.4	9.9	176	32595	233.4	3.8
	2006-07	483	39214	202.1	9.9	214	38856	234.2	3.5
State B	*2002-03*	176	5875	196.6	9.5	76	5667	230.3	5.2
	2003-04	246	8429	199.1	10.6	92	7723	231.8	5.9
	2004-05	161	5938	199.4	11.6	71	4307	231.6	5.4
	2005-06	220	8457	201.1	11.6	108	6049	231.6	4.8
	2006-07	310	13154	201.8	11.5	150	10922	233.9	4.4

Non-TAP Schools	Year	Grade 3				Grade 8			
		No. Schools	No. Students	Mean Spring Score	Mean Gain	No. Schools	No. Students	Mean Spring Score	Mean Gain
State A	*2002-03*	0	…	…	…	0	…	…	…
	2003-04	2	188	205.6	13.5	0	…	…	…
	2004-05	3	212	203	10.3	2	372	231.7	2.5
	2005-06	5	352	200.9	10.3	5	816	229.9	5.1
	2006-07	9	572	196.2	9.5	6	968	229.1	2.7
State B	*2002-03*	3	121	201.1	13.1	1	52	228.3	6.2
	2003-04	6	232	199.3	13.9	3	209	240	8.5
	2004-05	8	340	203	13.9	4	292	237.1	8.8
	2005-06	8	340	202.8	13.2	4	344	239	6.7
	2006-07	8	350	200.7	13.4	4	326	233.1	8.8

TABLE IV: SELECT SAMPLE STATISTICS

	State A		State B		State A and State B	
	TAP	**Non-TAP**	**TAP**	**Non-TAP**	**TAP**	**Non-TAP**
School Variables						
Average Teacher Salary / 100	41.4661 (2.0402)	42.2509 (2.5576)	44.0304 (2.8856)	42.2640 (5.9942)	42.6290 (2.7713)	42.2546 (3.8609)
Student Teacher Ratio	14.2311 (1.9967)	15.1572 (1.8309)	14.3514 (1.4694)	16.5263 (7.9787)	14.2857 (1.7781)	15.5475 (4.5748)
Percent Asian	0.5810 (0.5971)	1.3449 (1.4289)	0.7598 (0.5599)	2.2480 (2.7150)	0.6621 (0.5873)	1.6023 (1.9306)
Percent Hispanic	2.8726 (3.5905)	4.1328 (4.9403)	48.3527 (16.0123)	31.6667 (23.2051)	23.4972 (25.2184)	11.9814 (18.0401)
Percent Black	60.5920 (19.1620)	37.3409 (23.5781)	0.4713 (0.3869)	3.5373 (6.9244)	33.3280 (33.1145)	27.7051 (25.3772)
Percent Free Lunch	60.9929 (15.8674)	43.0831 (19.8725)	22.7651 (12.0029)	32.2977 (21.1330)	43.6571 (23.7722)	40.0087 (20.8172)
Percent Tested	99.7181 (0.7331)	99.3529 (3.0407)	98.1580 (4.2613)	98.8468 (4.5189)	99.0181 (3.0075)	99.2174 (3.5052)
Percent Below Basic (Math)	43.1514 (12.3013)	28.7932 (13.7996)	16.0086 (9.0382)	21.2033 (14.1963)	30.3854 (17.3813)	26.4862 (14.3524)
Percent Basic (Math)	36.9646 (5.0693)	39.2737 (5.5177)	32.8451 (7.0237)	33.5400 (9.3029)	35.0271 (6.4063)	37.5309 (7.3792)

	State A		State B		State A and State B	
	TAP	**Non-TAP**	**TAP**	**Non-TAP**	**TAP**	**Non-TAP**
Student Variables						
Male	0.5057 (0.5000)	0.5087 (0.4999)	0.5132 (0.4998)	0.5086 (0.4999)	0.5091 (0.4999)	0.5087 (0.4999)
Black	0.5902 (0.4918)	0.3527 (0.4778)	0.0040 (0.0629)	0.0356 (0.1852)	0.3243 (0.4681)	0.2631 (0.4403)
Hispanic	0.0280 (0.1651)	0.0397 (0.1952)	0.4816 (0.4997)	0.2924 (0.4549)	0.2337 (0.4232)	0.1110 (0.3142)
Asian	0.0051 (0.0709)	0.0113 (0.1056)	0.0079 (0.0883)	0.0299 (0.1702)	0.0063 (0.0793)	0.0165 (0.1275)
American Indian / Native Alaskan	0.0026 (0.0513)	0.0023 (0.0476)	0.0045 (0.0669)	0.0173 (0.1302)	0.0035 (0.0589)	0.0065 (0.0804)
Other	0.0007 (0.0257)	0.0041 (0.0640)	0.0018 (0.0420)	0.0425 (0.2018)	0.0012 (0.0340)	0.0150 (0.1214)
Time Elapsed between Fall and Spring Assessment	195.9135 (19.7490)	185.5276 (22.3877)	231.7004 (10.3309)	217.8205 (26.6779)	212.1424 (24.0616)	194.6451 (27.7839)

TABLE IV: SELECT SAMPLE STATISTICS (CONTINUED)

	State A		State B		State A and State B	
	TAP	**Non-TAP**	**TAP**	**Non-TAP**	**TAP**	**Non-TAP**
Dependent Variable						
Fall-to-Spring Test Score Gain						
Grade 2	13.4558 (7.6417)	11.7771 (7.5815)	18.1887 (8.0098)	13.6778 (8.4851)	16.0003 (8.1878)	12.0861 (7.7673)
Grade 3	10.4077 (7.6431)	9.8799 (7.4116)	13.5360 (7.3501)	11.0645 (7.7462)	12.0059 (7.6549)	10.2161 (7.5270)
Grade 4	9.2324 (7.8399)	7.4809 (7.4646)	11.3943 (7.3740)	9.0749 (7.5767)	10.1238 (7.7238)	7.9269 (7.5302)
Grade 5	8.7361 (7.8454)	6.9709 (7.5764)	11.0993 (7.3505)	8.8039 (7.6136)	9.7172 (7.7309)	7.4929 (7.6319)
Grade 6	5.6734 (8.1351)	4.9464 (7.8125)	9.1642 (7.6204)	6.6641 (7.6506)	6.9052 (8.1292)	5.3768 (7.8078)
Grade 7	3.8654 (8.2977)	4.3785 (7.9728)	7.9489 (7.3002)	5.9036 (7.8694)	5.3470 (8.1881)	4.7557 (7.9745)
Grade 8	3.5834 (8.5710)	3.8140 (8.0361)	8.0592 (7.7116)	5.0560 (8.0566)	5.2034 (8.5443)	4.1182 (8.0588)
Grade 9	1.1235 (7.9740)	1.5039 (9.4487)	3.3779 (8.1998)	2.5171 (9.0442)	2.6697 (8.1939)	1.7887 (9.3479)
Grade 10	0.2601 (8.4438)	1.6383 (9.4792)	4.4981 (10.2352)	1.7711 (9.5198)	3.7524 (10.0702)	1.7009 (9.4985)

Standard errors in parentheses

TABLE V: ORDINARY LEAST SQUARES MODELS

	Grade								
	2	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
Independent Variables									
TAP	2.7346*** (0.4415)	0.9390** (0.3370)	1.8631*** (0.2861)	2.0502*** (0.2943)	0.9795** (0.3576)	0.2587 (0.3747)	0.9061 (0.4801)	0.7925 (0.7609)	2.5040** (0.8613)
Time Elapsed between Fall and Spring Assessment	0.0529*** (0.0036)	0.0399*** (0.0032)	0.0379*** (0.0032)	0.0319*** (0.0031)	0.0227*** (0.0032)	0.0203*** (0.0032)	0.0106* (0.0042)	0.0082 (0.0065)	-0.0038 (0.0078)
Average Teacher Salary / 100	-0.0435 (0.0265)	0.0416* (0.0193)	0.0072 (0.0177)	0.0221 (0.0173)	0.0276 (0.0199)	0.0402* (0.0196)	0.0380 (0.0241)	0.0125 (0.0347)	0.0505 (0.0450)
Student-Teacher Ratio	-0.0311 (0.0575)	-0.0419 (0.0351)	-0.0673 (0.0415)	-0.0589 (0.0349)	-0.1426*** (0.0387)	-0.0169 (0.0376)	0.0346 (0.0432)	-0.1064 (0.0683)	-0.1013 (0.0808)
Percent Below Basic (Math)	-0.0362** (0.0123)	-0.0291** (0.0102)	-0.0347*** (0.0085)	-0.0464*** (0.0092)	-0.0358*** (0.0090)	-0.0322*** (0.0077)	-0.0306** (0.0098)	-0.0174 (0.0149)	-0.0249 (0.0199)
Percent Basic (Math)	-0.0374* (0.0155)	-0.0327** (0.0118)	-0.0151 (0.0104)	-0.0281** (0.0095)	-0.0144 (0.0124)	-0.0084 (0.0136)	-0.0033 (0.0158)	-0.0019 (0.0217)	-0.0051 (0.0229)
Intercept	6.7402** (2.2433)	2.9889 (1.6465)	1.3522 (1.5350)	2.3420 (1.5111)	5.6296** (1.9445)	2.6072 (1.7828)	-0.6726 (2.1752)	4.4272 (2.6015)	3.0198 (3.3433)
R^2	0.0551	0.0344	0.0343	0.0324	0.0232	0.0185	0.0135	0.0139	0.0123
N	90555	127474	128567	135480	135781	132893	129642	25468	16849

******* estimates statistically significant from zero at the 10%, 5%, and 1% levels, respectively.

Models include year fixed effects, NWEA cohort fixed effects, school-level controls for race/ethnicity, special education, and free lunch status, and student-level controls for race/ethnicity and gender.

Standard errors reported in parentheses.

Standard errors corrected for clustering of students within-schools.

TABLE VI: SCHOOL FIXED EFFECTS MODELS

	Grade								
	2	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
Independent Variables									
TAP	1.5750** (0.5222)	0.6369 (0.4330)	1.0553** (0.3956)	1.2305** (0.3942)	-0.4019 (0.3464)	-1.4043*** (0.3962)	0.5164 (0.4318)	-1.9929** (0.6131)	-0.9084 (0.7016)
Time Elapsed between Fall and Spring Assessment	0.0536*** (0.0017)	0.0456*** (0.0015)	0.0352*** (0.0015)	0.0320*** (0.0015)	0.0234*** (0.0016)	0.0201*** (0.0016)	0.0113*** (0.0017)	0.0125*** (0.0037)	-0.0138** (0.0052)
Average Teacher Salary / 100	-0.0309 (0.0379)	0.0370 (0.0262)	-0.0113 (0.0268)	-0.0296 (0.0257)	-0.0127 (0.0331)	0.0731* (0.0342)	-0.0262 (0.0354)	0.1974* (0.0872)	0.0372 (0.1144)
Student-Teacher Ratio	-0.0369 (0.0511)	-0.0840* (0.0364)	-0.0898* (0.0359)	-0.1180*** (0.0348)	-0.0970** (0.0371)	-0.1067* (0.0415)	-0.1469*** (0.0428)	-0.2673** (0.0850)	-0.1361 (0.1007)
Intercept	0.6377 (2.3330)	0.1050 (1.5763)	2.2188 (1.6164)	4.0115** (1.5305)	3.4396 (1.8662)	1.9072 (1.9819)	0.2409 (2.0449)	-0.4437 (4.9267)	3.9459 (6.1731)
R^2	0.0239	0.0158	0.0104	0.0082	0.0047	0.0040	0.0026	0.0078	0.0053
N	90555	127474	128567	135480	135781	132893	129642	25468	16849

*'**'*** estimates statistically significant from zero at the 10%, 5%, and 1% levels, respectively.

Models include year fixed effects, NWEA cohort fixed effects, school fixed effects, school-level controls for race/ethnicity, special education, and free lunch status, and student-level controls for race/ethnicity and gender.

Standard errors reported in parentheses.

TABLE VII: ORDERED-PROBIT SELECTION CORRECTION MODELS

	Grade								
	2	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
Independent Variables									
TAP	2.4625*** (0.7244)	0.6588 (0.7940)	1.7536*** (0.4749)	1.6340*** (0.4349)	-0.2205 (0.5258)	-0.7866 (0.5201)	0.3238 (0.5521)	-1.3673* (0.6467)	-0.9110 (0.5563)
Time Elapsed between Fall and Spring Assessment	0.0529*** (0.0036)	0.0399*** (0.0032)	0.0379*** (0.0032)	0.0320*** (0.0031)	0.0227*** (0.0032)	0.0202*** (0.0032)	0.0105* (0.0042)	0.0087 (0.0064)	-0.0024 (0.0073)
Average Teacher Salary / 100	-0.0432 (0.0266)	0.0415* (0.0194)	0.0072 (0.0177)	0.0219 (0.0173)	0.0264 (0.0199)	0.0395* (0.0196)	0.0377 (0.0241)	0.0101 (0.0345)	0.0425 (0.0433)
Student-Teacher Ratio	-0.0308 (0.0575)	-0.0420 (0.0351)	-0.0672 (0.0415)	-0.0584 (0.0349)	-0.1371*** (0.0384)	-0.0119 (0.0374)	0.0374 (0.0432)	-0.0819 (0.0671)	-0.0698 (0.0760)
Percent Below Basic (Math)	-0.0363** (0.0123)	-0.0292** (0.0102)	-0.0348*** (0.0085)	-0.0463*** (0.0092)	-0.0343*** (0.0090)	-0.0305*** (0.0078)	-0.0296** (0.0099)	-0.0160 (0.0147)	-0.0207 (0.0199)
Percent Basic (Math)	-0.0376* (0.0155)	-0.0328** (0.0118)	-0.0152 (0.0103)	-0.0284** (0.0095)	-0.0155 (0.0121)	-0.0097 (0.0135)	-0.0040 (0.0158)	-0.0011 (0.0205)	0.0019 (0.0217)
Mill's Ratio	0.1078 (0.2513)	0.1118 (0.2831)	0.0470 (0.1606)	0.1785 (0.1377)	0.5476** (0.1766)	0.4441* (0.1786)	0.2478 (0.2336)	1.1669*** (0.2560)	2.1087*** (0.4865)
Intercept	6.7139** (2.2363)	3.0004 (1.6465)	1.3549 (1.5350)	2.3486 (1.5107)	5.3547** (1.8862)	2.2668 (1.7124)	-0.8036 (2.1464)	1.9700 (2.2055)	-0.0828 (2.6017)
R^2	0.0551	0.0344	0.0343	0.0324	0.0235	0.0186	0.0136	0.0153	0.0152
N	90555	127474	128567	135480	135781	132893	129642	25468	16849

*,**,*** estimates statistically significant from zero at the 10%, 5%, and 1% levels, respectively.

Models include year fixed effects, NWEA cohort fixed effects, school-level controls for race/ethnicity, special education, and free lunch status, and student-level controls for race/ethnicity and gender.

Standard errors reported in parentheses.

Standard errors corrected for clustering of students within-schools.

TABLE VIII: OLS, SCHOOL FIXED EFFECTS, AND TWO-STAGE SELECTION-CORRECTION MODELS WITH *NEW-TAP* INDICATOR

Model		Grade								
	Independent Variables	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
OLS †	*TAP*	2.9122*** (0.5109)	1.0168** (0.3466)	1.8289*** (0.2952)	2.1801*** (0.3283)	0.7703* (0.3160)	0.1204 (0.3557)	0.9649 (0.5359)	1.1431 (0.7667)	2.2726** (0.7042)
	New-TAP	-1.0173 (0.8171)	-0.3849 (0.8201)	0.1400 (0.5230)	-0.5292 (0.4882)	0.7063 (0.4995)	0.7707 (0.5222)	-0.3186 (0.7744)	-1.1443** (0.3596)	0.7094 (1.7239)
School Fixed Effect ‡	*TAP*	2.8470*** (0.6004)	0.3401 (0.5000)	0.9122* (0.4428)	1.2724** (0.4446)	-1.4022*** (0.4105)	-2.2928*** (0.4552)	0.7216 (0.4893)	-1.7330** (0.6557)	-1.0983 (0.7485)
	New-TAP	-2.1345*** (0.4974)	0.5086 (0.4283)	0.2538 (0.3526)	-0.0730 (0.3578)	1.6308*** (0.3592)	1.6966*** (0.4281)	-0.3846 (0.4313)	-0.7691 (0.6883)	0.5328 (0.7319)
Two-Step Selection-Correction †	*TAP*	2.7202*** (0.7691)	0.7491 (0.8599)	1.6975*** (0.4706)	1.7789*** (0.4702)	-0.6612 (0.5346)	-1.0915* (0.5501)	0.3688 (0.6845)	-1.0203 (0.6623)	-1.1791 (1.0241)
	New-TAP	-0.9689 (0.8192)	-0.3275 (0.8465)	0.1690 (0.5126)	-0.4416 (0.4787)	1.0835* (0.4919)	1.1018* (0.5038)	-0.1587 (0.8312)	-1.1215** (0.3740)	0.7930 (1.7703)

*'**'*** estimates statistically significant from zero at the 10%, 5%, and 1% levels, respectively.

† Models include year fixed effects, NWEA cohort fixed effects, school-level controls for race/ethnicity, special education, free lunch status, time elapsed between fall and spring assessment, average teacher salary, student-teacher ratio, percentage of students scoring below basic in math, and percentage of students scoring basic in math, and student-level controls for race/ethnicity and gender. The two-step selection-correction model also includes the Mill's ratio.

‡ Models include year fixed effects, NWEA cohort fixed effects, school fixed effects, school-level controls for race/ethnicity, special education, free lunch status, time elapsed between fall and spring assessment, average teacher salary, and student-teacher ratio, and student-level controls for race/ethnicity and gender.

†,‡ Standard errors reported in parentheses.

† Standard errors corrected for clustering of students within-schools.

APPENDIX A: ORDINARY LEAST SQUARES MODELS USING *PRE-TAP* INDICATOR

	Grade								
Dependent Variables	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
Independent Variables									
TAP	2.7332*** (0.4427)	0.9403** (0.3369)	1.8651*** (0.2882)	2.0605*** (0.2952)	1.0376** (0.3490)	0.3004 (0.3655)	0.9405 (0.4793)	0.8317 (0.7305)	3.1762*** (0.8083)
Pre-TAP	-0.0796 (0.4538)	0.0833 (0.6179)	0.0528 (0.2868)	0.2521 (0.2515)	1.0379** (0.3640)	0.9125* (0.3811)	0.7522 (0.4613)	1.7393*** (0.4565)	3.8225*** (1.0368)
Time Elapsed between Fall and Spring Assessment	0.0529*** (0.0036)	0.0399*** (0.0032)	0.0379*** (0.0032)	0.0320*** (0.0031)	0.0228*** (0.0032)	0.0201*** (0.0032)	0.0104* (0.0042)	0.0086 (0.0065)	-0.0023 (0.0073)
Average Teacher Salary / 100	-0.0434 (0.0265)	0.0415* (0.0194)	0.0072 (0.0177)	0.0218 (0.0173)	0.0262 (0.0199)	0.0391* (0.0196)	0.0372 (0.0241)	0.0115 (0.0346)	0.0440 (0.0430)
Student-Teacher Ratio	-0.0315 (0.0575)	-0.0417 (0.0355)	-0.0671 (0.0414)	-0.0580 (0.0349)	-0.1342*** (0.0382)	-0.0073 (0.0375)	0.0424 (0.0434)	-0.0834 (0.0681)	-0.0732 (0.0777)
Percent Below Basic (Math)	-0.0361** (0.0123)	-0.0293** (0.0103)	-0.0349*** (0.0086)	-0.0469*** (0.0092)	-0.0362*** (0.0089)	-0.0317*** (0.0077)	-0.0301** (0.0098)	-0.0164 (0.0147)	-0.0199 (0.0205)
Percent Basic (Math)	-0.0375* (0.0155)	-0.0326** (0.0117)	-0.0150 (0.0104)	-0.0279** (0.0095)	-0.0124 (0.0120)	-0.0073 (0.0133)	-0.0023 (0.0155)	-0.0016 (0.0208)	0.0013 (0.0226)
Intercept	6.7601** (2.2392)	2.9837 (1.6475)	1.3487 (1.5349)	2.3218 (1.5101)	5.1955** (1.8820)	2.1086 (1.7063)	-0.9742 (2.1265)	2.3881 (2.2653)	-0.1285 (2.5894)
R^2	0.0551	0.0344	0.0343	0.0324	0.0235	0.0187	0.0137	0.0151	0.0147
N	90555	127474	128567	135480	135781	132893	129642	25468	16849

*,**,*** estimates statistically significant from zero at the 10%, 5%, and 1% levels, respectively.

Models include year fixed effects, NWEA cohort fixed effects, school-level controls for race/ethnicity, special education, and free lunch status, and student-level controls for race/ethnicity and gender.

Standard errors reported in parentheses.

Standard errors corrected for clustering of students within-schools.