Draft working paper
Comments welcome, please do not cite without permission of the authors.

Using Student Performance Data to Identify Effective Classroom Practices

John H. Tyler*

Eric S. Taylor***

Thomas J. Kane**

Amy L. Wooten***

December 31, 2009

Recent research has confirmed both the importance of teachers in producing student achievement growth and in the variability across teachers in the ability to do that. The findings of this research raise the stakes on our ability to identify effective teachers and identify effective teaching practices. This paper combines information from classroom-based observations and measures of teachers' ability to improve student achievement as a step toward addressing these challenges. We show that classroom based measures of teaching effectiveness are related in substantial ways to student achievement. Our results also offer information on which types of classroom practice are most effective at raising achievement.

* Brown University and NBER

** Harvard University and NBER

*** Harvard University

The authors would like to thank Douglas Staiger, Ron Ferguson, and participates at the Brown University seminar on Race and Inequality, the Program on Education Policy and Governance Colloquium at Harvard University, and the Speaker Series at the Institute for Research on Education Policy and Practice at Stanford University for helpful comments on previous drafts of this paper. We also acknowledge the Joyce Foundation for their generous support of this project, as well as the cooperation and support of the Cincinnati Public Schools.

1. Introduction

Many people, experiences and structures contribute an individual student's achievement. The contribution of teachers, however, has increasingly become a focal point as a number of studies have found large differences in teachers' effectiveness at increasing student achievement. For example, Gordon, Kane and Staiger (2006) find that a student assigned to a top-quartile ranked teacher in terms of promoting achievement growth scored 10 percentile points higher on standardized tests of achievement than otherwise similar students assigned a bottom-quartile ranked teacher. Other researchers have found similar-sized variation in teacher effects on student achievement (Aaronson, Borrow and Sander (2003), Rivkin, Hanushek and Kain (2005), Rockoff (2004), Kane, Rockoff and Staiger (2005)).

Unfortunately, beyond experience in the classroom, little is known about which skills, characteristics, and practices of a teacher cause the observed large differences. This knowledge gap is critical since it severely curtails the policy and management levers available for creating an effective teacher workforce. In this paper we provide initial evidence on the relationship between classroom management and instructional skills and gains in student achievement. We do so using data from the Cincinnati Public School (CPS) system which, like many other systems, maintains historical data on student achievement and teacher assignments; but which, unlike other systems, also maintains detailed data on classroom observations as part of its long-running Teacher Evaluation System (TES). We find that the classroom management and instructional skills measured by TES meaningfully predict student achievement growth. Additionally, we show that some specific skills merit additional attention by teachers, administrators, and policy makers interested in helping teachers increase their contribution to their students' achievement.

2. Measuring Teachers' Classroom Practices in Cincinnati

Systems that provide detailed measurement and evaluation of teacher practices are rare. An examination of evaluation practices in twelve districts in four states by the New Teacher Project found that evaluations of teachers "are short and infrequent—most are based on two or fewer classroom observations totaling 60 minutes or less—[and] conducted by untrained administrators" (Weisberg et al. (2009), p. 2006). In wider examination of teacher evaluation systems Toch and Rothman (2008) conclude that while our knowledge of how to effectively and fairly evaluate teachers has grown substantially in the past decades, the "vast majority of districts" still do not employ a credible system of measuring the quality of teachers' work. Cincinnati's TES program is an exception to this generalization. During the 1999-2000 school year Cincinnati Public Schools field tested the TES system that utilizes trained evaluators, a specified and research-based evaluation rubric, and usually includes at least four classroom observations of teachers during a year, only one of which is announced.

The foundation of the TES system is a set of practices and behaviors set forth in Charlotte Danielson's *Enhancing Professional Practice: A Framework for Teaching*. The rubric associated with the "Danielson framework" includes four domains, and within each domain, teachers are evaluated against a set of standards, which themselves are subdivided into elements. Each element has language that describes performance at each level of the rubric: Distinguished, Proficient, Basic, and Unsatisfactory, with evaluators assigning respective scores of 4, 3, 2, and 1 to these rubric levels. Cincinnati Public Schools maintains detailed records for each TES evaluation, including scores from each classroom observation and each portfolio review that contribute to the final score. Our data contain records on 1,830 teacher TES evaluations covering 2000-01 through 2007-08 with a high of 292 in 2006-07 and a low of 112 in 2000-01. Each teacher was observed in the classroom between one and eight times; 97 percent were observed between two and six times.

Our analysis will focus on the domains 2 and 3 in the TES system, the domains that are associated with classroom practice and that involve classroom observations by TES evaluators. The focus of domain 2 is "Creating an Environment for Learning" and the focus for domain 3 is "Teaching for Learning."[1] There are three standards in domain 2 covering practices such as "the teacher creates an inclusive and caring environment," "the teacher establishes effective routines and procedures…and manages transitions to maximize instructional time," and "the student manages and monitors student behavior to maximize instructional time." There are five standards in domain 3 covering practices such as "the teacher communicates standards-based instructional objectives and high expectations," "the teacher demonstrates content knowledge," "the teacher uses…instructional activities that promote conceptual understanding, "the teacher engages

[1] The focus of domain 1 is "Planning and Preparing for Student Learning" and the focus of domain 2 is "Professionalism." Teachers are evaluated in these two domains based upon artifacts such as lesson plans, written communications to parents, earned professional development certificates, etc.

students in discourse and uses thought-provoking questions," and "the teacher provides timely, constructive feedback to students."[2]

Under each of the eight standards in domains 2 and 3 there are (potentially) multiple "elements," and it is at the element level that the rubric language used by TES evaluators resides. Thus, a single TES score sheet from one observation of a teacher can have multiple (element) scores for a given standard, and a given teacher will have multiple TES score sheets (one for each classroom observation) in a given year that she has been under TES evaluation.[3] Our analysis uses all of this information to compute a year-end average standard level score for each of the eight standards in domains 2 and 3 for that teacher and that year.

3. Measuring Student Achievement Growth in Cincinnati

Paralleling the TES program years, we have panel data on Cincinnati students for the 2000-01 through 2008-09 school years. The student-by-year observations include information on the student's gender, race or ethnicity, English proficiency status, participation in special education or gifted and talented programs, class and teacher assignments by subject, and, when applicable, standardized test scores.

Between 2000-01 and 2008-09 Cincinnati students, in general, took end of year exams in reading and math in third through eighth grades. However, in earlier years the testing program did not cover all grades, and over the course of 2003-04 to 2005-06 the state switched tests from the State Proficiency Test (SPT) and its companion the Off Grade Proficiency Test (OGPT) to the Ohio Achievement Test (OAT). In all cases we standardize (mean zero, standard deviation one) test scores by grade and year. Across all tested grades and years we have math test scores for 93 percent of students (ranging from 83 percent to 97 percent in any particular grade and year) and reading scores for 94 percent of students (ranging from 83 percent to 98 percent in any particular grade and year).

[2] This language from the TES rubric are meant to be illustrative and do not represent the entirety of the rubric language in any of the examples.

[3] The actual TES score sheets in the CPS files do not have element scores per se, but they do contain the rubric language that the TES evaluator used in describing the observed practices of the teacher under observation. Given the virtual one-to-one correspondence between the rubric language and the 4,3,2, 1 scores associated with the language one can "score" a TES observation sheet. For this project we hired a team of retired CPS teachers to read and score all of historical TES observation sheets.

Our empirical strategy requires both an outcome test (e.g., end of year test in year t) and a baseline test (e.g., end of year test in year t-1). Thus, our analysis sample will exclude some entire grade-by-year cohorts who were not tested in year t or t-1. For example, the largest gap is in fifth-grade math where students were not tested in the years 2001-02 through 2004-05. This gap also excludes sixth-grade students in 2002-03 through 2005-06. We are able to close some third-grade gaps using 2nd grade math and reading tests administered in 2000-01 through 2002-03, and a reading test administered to 3rd graders in the fall beginning in 2003-04.

Cincinnati Public Schools also maintains records of individual students' class schedules that include the teacher, course, and section.[4] Using these data we identified a math (and separately a reading) class and teacher for each student each school year. For the 2003-04 school year and subsequent years we identified a math teacher and class for 97 percent of tested students in grade 3-8, and a reading teacher and class for 96 percent of the same population.[5] For the 2000-01 through 2002-03 school years the available class schedule data are more limited. In these earlier years teacher and section information is mostly absent; indeed it would be entirely absent but for the efforts of prior researchers studying the TES program (Holtzapple (2003)). To facilitate that prior analysis, a previous research team identified student rosters for a number of teachers evaluated by TES. Thus we can identify a math and reading teacher for selected students in 2000-01 through 2002-03.

4. Empirical Strategy

The goal is to estimate the relationship between teachers' TES scores and the ability of teachers to promote student achievement growth. Constructing a model for estimating the parameters of this relationship involves modeling the relationship between unobserved teaching skills and measured classroom practices and in specifying the potential tradeoffs between using different years of measured classroom practices relative to when achievement is measured. In the

[4] Cincinnati's historical class schedule data retain each student's last class assignment for each course each year. This structure does not allow us to identify students who had more than one teacher or class during the year (or semester). Thus, for example, if a student originally enrolled in Mr. Smith's Pre-algebra class, but later transferred to Ms. Jones Pre-algebra class the available data record Ms. Jones and the appropriate section number.

[5] Infrequently a student's record indicates one teacher and class for reading, and a different teacher or class for other English language arts subjects (e.g., spelling, writing). In such cases we use the reading teacher given the test content. Students for whom we could not identify a class were almost always missing from the class schedule data entirely, or, much less frequently, did not have a class listed in the specific subject.

Appendix we develop the model relating latent, unobserved teaching skills to student achievement growth, and we describe the use of TES scores as potentially error prone measures of these latent traits. This development also discusses the issues involved in choosing which year to measure TES scores relative to which year student achievement is measured. A summary of the Appendix discussion is that:

1. there is potential bias in estimating the relationship between achievement growth and TES scores using any combination of TES and achievement years,
2. while using TES scores from the same year that achievement is measured—meaning that both measures are based on the same class of students—may *a priori* seem like the most intuitive solution, there are reasons that this contemporaneous solution may, in fact, not be the best choice, and
3. using TES scores from the year *following* the measurement of student achievement likely results in the most generalizable.

Based on the reasoning developed in the Appendix we will focus our discussion on results that use TES scores measured in the year immediately following the measurement of student achievement, but we will show results using all possible combinations of achievement and TES years. To estimate the relationship between a teacher's observed classroom practices and that teacher's ability to promote student achievement growth we fit equation 1 (based on Appendix equation A.3b) to the data:

$$(1) \qquad A_{ijkt} = \alpha + TES_{Jk,t+1}\gamma + \sum_{m} \rho^{m} (Exp^{m}_{k,t+1} * -n) + A_{i,t-1}\beta + X_{it}\delta + v_{ijkt}$$

where *i* indexes students, *j* and *J* index classes ($j \neq J$), *k* indexes teachers, and *t* indexes year, and *v* is an error term that may be correlated with *TES* (see the Appendix for this discussion). A_{ijkt} is the end of year math (reading) test score for student *i* taught by teacher *k* in class *j* during school year *t*. The vector $A_{ik,t-1}$ captures the student's prior achievement including the main effect of the prior year math (reading) test score, the score interacted with each grade-level, and fixed effects for each test (i.e., grade-by-year fixed effects). When the baseline score was missing for a student, we imputed $A_{ik,t-1}$ with the grade-by-year mean, and included an indicator for missing baseline score. A vector of student-level controls, X_{it}, includes separate indicators for student (i) gender, (ii) race or ethnicity, and whether, in our observed data, the student was ever (iii) retained in grade or participating in (iv) special education, (v) gifted, or (vi) limited English proficient programs. $TES_{Jk,t+1}$ is a vector of TES measures of the observed classroom practices of teacher *k* in class *J* in year *t+1*.

To this point we have not discussed in detail the composition of the $TES_{Jk,t+1}$ vector. One intuitive approach would be to simply include the eight TES standards scores from domains 2 and 3. In practice, however, the scores across these eight standards are highly correlated so that estimates of the γ s tend to be unstable and hard to interpret.[6] To address this situation we use the first three principal components from a principal components analysis of the eight standards in domains 2 and 3. These three components explain 87 percent of the variance of the eight standard scores, and a scree plot of the eigenvalues of the standard scores correlation matrix suggests retaining at most three components. In this analysis all eight of the standards load about equally on the first principal component. The second principal component is a contrast between the scores in domains 2 and the scores in domain 3. The third principal component is a contrast between the score on standard 3.4 and a combination of the scores in standards 2.2, 3.1 and 3.2.

Our interpretation of these principal components is that the first principal component captures the general importance of all eight behaviors and practices measured in domains 2 and 3. A contrast between the scores in domains 2 and 3—the second principal component—is a contrast between the type of *classroom environment* a teacher has created as recorded by the TES evaluator (domain 2) and the extent to which an evaluator observes a teacher engaging in *teaching practices* that are believed to be related to student learning (domain 3). Conceptually, the third principal component is a contrast between two types of teaching. The first type of teaching can be described as a pedagogical style that is focused on engaging students in discourse and exploring and extending the students' content knowledge through thought-provoking questioning. One might call this *inquiry-based teaching.* This is contrasted in the third component with teaching that focuses on classroom management routines, on conveying standards-based instructional objectives to the students, and on teaching in which the teacher

[6] The correlations between the eight standards range between 0.619 and 0.813 and estimating equation 1 using these 8 standards as the *TES* vector results in only two statistically significant coefficient estimates and several wrong signed (negative) coefficient estimates.

demonstrates content-specific pedagogical knowledge in teaching these objectives. One might call this *routinized standards and content focused teaching*.

Instead of using the component loadings that result from the principal components analysis to form linear component scores, we have elected to use their counterparts constructed from simple functions of the TES standard score variables. To capture the essence of the first principal component we use a teacher's average score across all eight standards. To capture the second we subtract the average of a teacher's domain 3 standard scores from the average of her domain 2 standard scores. For the third we subtract the average of standards 2.2, 3.1, and 3.2 from a teacher's score on standard 3.4. Figures 1a, 1b, and 1c display the distribution, mean, and standard deviation for each of the three principal-component-based measures.

The correlation between the each of the three principle components and the constructed counterparts we use are 0.999, 0.981, and 0.947 respectively. At the same time, the correlations among the three constructed component variables are, as expected, relatively low ($\rho_{1,2} = 0.110$, $\rho_{1,3} = 0.049$, $\rho_{2,3} = -0.107$). All of the analyses that follow use these constructed component variables as the elements of $TES_{Jk,t+n}$. Additionally, we always include a fixed effect for the year in which the TES evaluation was conducted.

5. Results and Discussion

In the analysis that follows we ask: (1) to what extent do TES scores predict student achievement growth, and (2) which classroom practices measured by the TES process are the most effective at promoting student achievement? Table 1 has the first answers to these questions and reports the relationship between TES scores and student achievement growth as specified in Equation 1. In Table 1 an average TES score increase of one is associated with a student achievement gain of about one-sixth of a standard deviation in math and one-fifth in reading. A one point increase in the average scores across the eight standards represents an increase of about two standard deviations (see Figure 1a). Meanwhile, a teacher who scores higher on "classroom environment" (Domain 2) relative to "classroom practices" (Domain 3) is predicted to produce additional student gains; with coefficients of 0.25 standard deviations in math and 0.15 in reading. Last, a teacher who scores higher on *inquiry-based teaching* (Standard 3.4) relative to *routinized* *standards and content focused teaching* (Standards 2.2, 3.1 and 3.2) is predicted to produce student gains in reading but not in math.[7]

To place these results in the context of the TES system, the estimates on the first principal component suggest that a student assigned a teacher whose average scores placed her in the "Distinguished" category would, by the end of the school year, score more than one-fifth of a standard deviation higher in reading than her peer in a class taught by a "Proficient" teacher. The estimates on the second and third principal components in Table 1 require some interpretation.

The literal interpretation on the second component is that controlling for the average TES score, a teacher whose domain 2 average is one point higher than her domain 3 average would generate student achievement gains in math that are 0.25 of a standard deviation higher than a teacher whose average scores in these two domains are the same. The similar estimate for reading achievement is 0.15 of a standard deviation. That is, the correct interpretation of the estimated coefficients on the second principal component is that it is the *contrast* between the domain 2 and domain 3 averages that matter. Likewise, when it comes to the third principal component it is a *contrast* in teaching styles and emphasis that matters, at least when it comes to reading achievement gains.

One interpretation of the estimated effects of the second and third principal components on student achievement gains is as follows.[8] The contrasts in these principal components can be thought of as measures of the relative emphases teachers place on the different things they do in class *while they are being observed* by TES evaluators. Thus, the second component can be viewed as the relative importance a teacher places on the climate of the classroom versus an emphasis on the exact instructional practices in which she is engaged on the day she is being observed. Taken literally, the estimates on the second component suggest that given two classrooms whose teachers have the same overall average scores across domains 2 (classroom environment) and 3 (instructional practices), the students in the classroom where the TES evaluator rates the classroom environment to be better than the instructional practices of the teacher are expected to learn more than the students in a classroom where the classroom environment and instructional practices of the teacher are rated about equally by the TES

[7] When we restrict the sample to teachers for whom we have both math and reading scores in years *t* and *t-1*, the results are similar but most similar for the first overall TES measure.

[8] We thank Ron Ferguson for his very helpful insights on these interpretations and this section is largely the product of discussions and correspondence with him on this topic.

evaluator. For example, it might be that the students in the first class were observed to be better behaved, more respectful to each other and the teacher, and spending more time on task than the students in the second class, but the quality of the pedagogy was judged to be lower in the first class than the second. The estimates in Table 1 suggest the students of the first teacher will learn more than the students of the second teacher. One possible explanation for this result is that Cincinnati might be operating in the range of the education production function where increases in classroom environment inputs such as keeping kids on task have bigger payoffs to student achievement than increases to inputs associated with instructional practice such as the extent to which teachers "communicate standards-based instructional objectives" to students. Unfortunately, we have no data that would allow exploration of this possibility.

As stated earlier, the third principal component is a contrast between what we call *inquiry-based* instruction and *routinized content and standards-based* instruction. This contrast suggests that at least when being observed teachers may be making a tradeoff between placing an emphasis on engaging students in discussion and taking the class time necessary to do that, and placing an emphasis on "managing transitions to maximize instructional time," "communicating standards-based instructional objectives," and demonstrating their own content knowledge "by using content specific instructional strategies." That is, it may not be possible to do everything during the class period in which a teacher is being observed. In particular, if it takes time to engage students via inquiry and the give and take of discussion, there may be fewer opportunities for a teacher to demonstrate other instructional practices that are in the TES rubrics. The estimates in Table 1 suggest that to the extent that this is the case, then teachers observed making a tradeoff in favor of inquiry-based instruction tend to produce higher student achievement in reading but not in math.

The discussion over the exact meaning of the estimates on the second and third components in Table 1 should not obscure the overarching message of the table. Namely, that TES scores are an important predictor of student achievement growth. In particular, while some of the classroom practices measured by the TES process appear to be more important than others, a teacher's TES average across domains 2 and 3 is an important predictor of how well that teacher's students will perform. To provide a sense of how important, if fadeout is minimal, a core of "Distinguished" teachers might well close the black-white achievement gap—often estimated at one standard deviation—in five to six years relative to the same students being taught by a core of "Proficient" teachers.

We next turn to the sensitivity of our estimates to our choice of using TES scores from year *t+1*. Table 2 shows that our point estimates change somewhat when using TES in years other than *t+1*. Most notably, the relationship between the domains 2 and 3 contrast and achievement does not appear in other years (except for the "any following year" reading estimate). By contrast, the coefficients for the overall TES score remain fairly consistent.

A second regularity in Table 2 is that a teacher's overall TES score is most strongly associated with achievement gains for the students he taught during the year of the TES evaluation (i.e., 0.27 in math, 0.26 in reading). This stronger association need not be unexpected because of the reasons discussed in the Appendix having to do with the correlation between contemporaneous measures of TES and student achievement growth particular to a classroom environment.

Table 3 reports the results of formal comparisons of the coefficients in Table 2, in particular comparisons between using TES scores from year *t + 1* and all other available TES years. We note that none of the estimates on the average TES score are different and that it is only in reading for the *t + 1* to *t-1* comparison that we find a statistically significant difference between the estimates of the third TES component. In summary, Table 3 suggests that our estimate of the relationship between the average TES score and student achievement growth is quite robust to the choice of TES year.

6. Conclusion

Our results provide some of the strongest evidence to date that classroom observation measures capture elements of teaching that are related to student achievement. Our estimates show a positive and non-trivial relationship between TES scores and student achievement growth. Our main results from Table 1 indicate that moving from, say, an overall TES rating of "Basic" to "Proficient" or from "Proficient" to "Distinguished" is associated with student achievement gains of about one-sixth to one-fifth of a standard deviation. Though moving from "Proficient" to "Distinguished" on the TES scale may be more difficult than a casual reading of the rubric's evaluative language would suggest.

Relating observed classroom practices to achievement growth offers some insight regarding what types of classroom practices may be important in increasing student achievement. First, we show that a teacher's overall score is important. Our results predict that policies and programs that help a teacher get better on all eight "teaching practice" and "classroom environment" skills measured by TES will lead to student achievement gains. Second, given teachers who have similar proficiency in "teaching practices" (measured in TES domain 3), helping teachers improve their "classroom environment" management (measured in TES domain 2) will likely also generate higher student achievement. Third, given two teachers who are equally adept at "content and standards focused teaching," the teacher who adds "inquiry-based pedagogy" practices will generate higher reading achievement, but not higher math achievement. Teachers working to improve their practice should consider their current performance in these areas.

While our results demonstrate relationships between practices measured in TES and student achievement growth, we cannot exclude relationships with practices not measured by TES nor do we intend to suggest that other TES measures should necessarily be discarded. First, it is unclear whether the relationships we observed would hold if the TES rubric elements, those in domains 1 and 4, were no longer measured or discussed. Second, a district may value outcomes for its teachers and students beyond growth in standardized test scores. This latter decision deserves serious discussion, but is beyond the scope of our analysis.

References

Aaronson, D., Barrow, L., and Sander, W. (2003). "Teachers and Student Achievement in the Chicago Public High Schools." Chicago: Federal Reserve Bank of Chicago.

Gordon, Robert, Thomas J. Kane and Douglas O. Staiger. (2006). "Identifying Effective Teachers Using Performance on the Job" Hamilton Project Discussion Paper, Published by the Brookings Institution.

Hanushek, Eric A. (1971). "Teacher characteristics and gains in student achievement; estimation using micro data". *American Economic Review*, 61:280-288.

Holtzapple, Elizabeth. (2003). "Criterion-Related Validity Evidence for a Standards-Based Teacher Evaluation System," *Journal of Peronnel Evaluation in Education,* 17(3): 207-219.

Kane, Thomas J. and Douglas O. Staiger. (2008). "Estimating Teacher Impacts on Student Achievement: An Experimental Evaluation." *NBER working paper* #14601, December 2008.

Rivkin, Steven, Eric Hanushek and John Kain. (2005). "Teachers, Schools and Academic Achievement" *Econometrica,* 73(2):417-458.

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

Toch, Thomas and Robert Rothman. (2008) "Rush to Judgment: Teacher Evaluation in Public Education in *Education Sector Reports,* January 2008.

Weisberg, Daniel, Susan Sexton, Jennifer Mulhern, and David Keeling. (2009). *The Widget Effect: Our National Failure to Acknowledge and Act on Differences in Teacher Effectiveness*. Report for The New Teacher Project, Downloaded from http://widgeteffect.org/.

Appendix

Over the course of a career, each teacher develops a set of classroom management and instructional skills. In any particular school year, an individual teacher's collection of skills is a function of several factors including her pre- and in-service training, performance evaluations, peers and administrators, and the quantity and characteristics of classes and students taught to date. In our notation teacher *k*'s present skills employed, but unmeasured, in school year *t* are represented by the vector Λ_{kt}. We are interested in estimating the relationships, ω, formalized in Equation A.1, between the elements of Λ_{kt} and A_{ijkt}, the achievement of student *i* in class *j* taught by teacher *k* in school year *t*, net of student *i*'s prior achievement, $A_{i,t-1}$, and observable characteristics, *X*, of student *i* that might affect achievement,

$$\text{(A.1)} \quad A_{ijkt} = \alpha + \Lambda_{kt}\omega + A_{i,t-1}\beta + X_{it}\delta + +v_{ijkt}$$

While a teacher's true Λ_{kt} is unobserved, one could sample a teacher's practices by visiting his classroom. Records of such observations, including the extensive TES data, are potentially useful, even if error prone, measures of Λ_{kt}. In Equation A.2 we formalize this relationship using the vector $TES_{jk,t+n}$ to represent a teacher *k*'s TES scores observed in classroom *j* during school year *t*+*n*.

$$\text{(A.2)} \quad TES_{jk,t+n} = \Lambda_{kt}\delta + \sum_{m} \phi^{m}(Exp^{m}_{k,t+n} * -n) + w_{jk,t+n} + u_{k,t+n}\text{, } where\ \ n \le 0 \le n$$

Beyond a direct relationship to a teacher's true practices, Λ_{kt}, a teacher's measured practices, $TES_{jk,t+n}$, are determined by three additional factors. The first and second are sources of error: $w_{jk,t+n}$ representing error related to the class of students, *j*, in which the teacher is observed, and $u_{k,t+n}$ representing residual idiosyncratic error.

The third arises because we may not have—or may choose not to use—TES observation scores from the school year under study; that is the *t* in Equation A.1 may not equal *t*+*n* (i.e., $n \neq 0$). To the extent an additional year of experience improves a teacher's classroom skills, past (or future) classroom observation scores will diverge from the true practices and skills a teacher presently employs. The series of terms $(Exp^{m}_{k,t+n} * -n)$, indexed by *m,* are intended to capture the difference in a teacher's classroom experience between the year she is observed for TES, year *t*+*n*, and the year in which we are interested in knowing Λ_{kt}, year *t*. We might have simply included the number of years since (or before) the TES observation, *n*; extant evidence suggests, however, that the returns to experience for teachers are non-linear (see Kane, Rockoff and Staiger (2006) for a review). Thus we allow the effect of *n* to vary depending on the quantity of experience teacher *k* had at the time of the TES observation, the *m* indicator variables $Exp^{m}_{k,t+n}$.

Rearranging terms in Equation A.2 and substituting into A.1 we get Equation A.3.

$$\text{(A.3)} \quad \begin{aligned} &A_{ijkt} = \alpha + TES_{Jk,t+n}\gamma + \sum_{m} \rho^{m}(Exp^{m}_{k,t+n} * -n) + A_{i,t-1}\beta + X_{it}\delta + \eta(w_{Jk,t+n} + u_{k,t+n}) + v_{ijkt} \\ &and \quad j = J \ \ if \ and \ only \ if \ \ n = 0. \end{aligned}$$

Stating Equation A.3 allows us to evaluate options for the data we will use to estimate γ and other parameters and it makes explicit the ideas that achievement *A* may be measured in a different year than *TES* ($n \neq 0$), and that if this is the case then the class in which student *i*'s achievement, *A*, is measured is different than the class in which teacher *k*'s classroom practices, *TES*, are observed ($j \neq J$). For discussion we define three options for when we might measure *TES* relative to *A*, though they are not necessarily mutually exclusive. Specifically, we can predict student achievement, A_{ijkt}, as a function of the teacher's TES scores measured in: (i) the contemporaneous school year[9], *n=0,* (ii) some previous school year, *n<0*, or (iii) some future school year; that is, *n>0*. Each of these three options requires different assumptions about the error terms, and thus brings different potential biases in estimating γ. We summarize these assumptions in Table A.1.

[9] In theory option (ii) and (iii) could be done with two different classes taught in the same school year, but the TES data do not allow us to pursue this approach.

Table A.1: Assumptions Regarding Error Correlation

	Option 1: $n=0$	Option 2: $n<0$	Option 3: $n>0$
$A_{ijkt} \perp u_{k,t+n}$	Yes	Yes	Yes
$A_{ijkt} \perp w_{Jk,t+n}$	No	No	Yes
$TES_{Jk,t+n} \perp v_{ijkt}$	Yes	Yes	No
$TES_{Jk,t+n} \perp w_{Jk,t+n}$	No	No	No

Option one may (n = 0, and $j = J$), *a prioi*, be the most intuitive option. However, given the contemporaneous measurement of *A* and *TES* in this option, unobserved class characteristics, for example the level of social cohesion among the students, may independently affect both a TES observer's measurement *and* student achievement.[10] To the extent this is the case, our estimates of γ will be biased. Our concerns regarding options one and two are structurally similar, but the mechanisms are different. Even though option two uses two separate classes of students ($j \neq J$), a teacher's particular past classes may affect his current students' achievement *through him* in ways independent of the average gains from experience. Under option three, we are no longer concerned with potential correlation between A_{ijkt} and $w_{Jk,t+n}$ because class *J* occurs in the future relative to class *j*. We are, however, concerned with the effect of a teacher's past classes on her future TES scores, again in ways not captured by the average gains from experience.

Recognizing that we lack measures of the potential bias that would indicate a strong preference for one of these options, we proceed as follows. First, we report our main estimates of γ separately under each option. It turns out that the point estimates are very similar. Second, we focus the bulk of our discussion on results from the third option, and specifically n = 1, student

[10] To see why consider an example of two classes, class A and class B, in which an evaluator is measuring TES standard 3.4: "The teacher engages students in discourse and uses thought-provoking questions aligned with the lesson objectives to explore and extend content knowledge." Assume for this example that the teachers in those two classes have identical Λ s. Class A is a representative sample of the school's students, but class B is composed of students who are unusually socially cohesive. Even in this case where the teachers in both classes have identical underlying teaching skills, class B may be more likely to exhibit to an observer the ideal described in standard 3.4. Thus the characteristics of class B introduce error in our attempt to measure a teacher's true ability to use questions and foster conversation across *all classes* he taught that school year. Additionally, the same unusual social cohesion in class B's may also result in positive peer effects that raise achievement independently of the teacher's contribution.

achievement as a function of a teacher's TES scores measured the following school year. Notice that if we choose $n = 1$ then $A_{ijkt} \perp u_{k,t+1}$ and $A_{ijkt} \perp w_{Jk,t+1}$ based on the assumptions in Table A.1 so that equation A.3 can be rewritten as

$$\text{(A.3b)}\ \ A_{ijkt} = \alpha + TES_{Jk,t+1}\gamma + \sum_{m} \rho^{m} (Exp^{m}_{k,t+1} * -n) + A_{i,t-1}\beta + X_{it}\delta + v_{ijkt}$$

We chose the third option in part given the greater potential for the generalizability of our results. One way to think of the first and second options is that they study classes where the teacher has participated in the TES process—a *process* that may uniquely change a teacher's classroom management and instructional practices. The change may be additive, or detrimental, or may simply make teachers more homogeneous in terms of their practice. By contrast, teachers who will participate in TES in the future, as in option three, may still be a selected sample, but their pre-TES-participation practices are likely closer to the average teacher than teachers who have already been through TES.

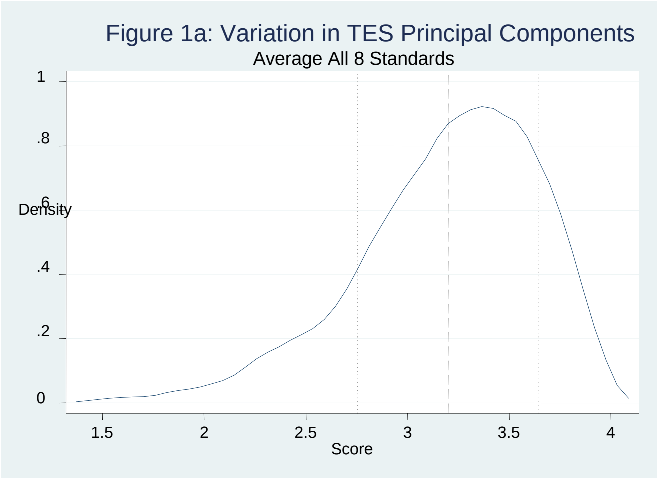

Figure 1a: Variation in TES Principal Components

Average All 8 Standards

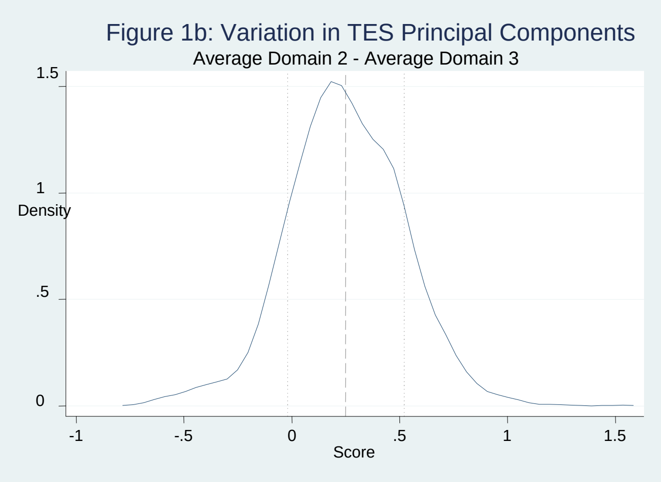

Figure 1b: Variation in TES Principal Components

Average Domain 2 - Average Domain 3

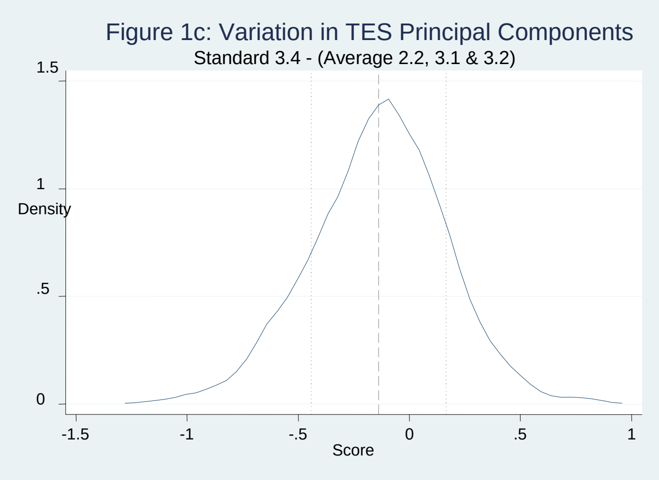

Figure 1c: Variation in TES Principal Components

Standard 3.4 - (Average 2.2, 3.1 & 3.2)

Table 1: Estimates of the Relationship Between Student Test Scores & TES Score Principal Components

	Math	Reading
Average All 8 Standards	0.171*	0.212***
	(0.071)	(0.052)
Average Domain 2 - Average Domain 3	0.249**	0.147*
	(0.086)	(0.066)
Standard 3.4 - (Average 2.2, 3.1 & 3.2)	-0.050	0.150*
	(0.102)	(0.068)
TES Year Fixed Effects	Y	Y
Teacher Experience Terms	Y	Y
Student-level Covariates	Y	Y
R-squared	0.530	0.506
Student Sample	3,791	5,739
Teacher Sample	100	206

Note: Each column represents a separate student-level specification. Student achievement measured in the year just prior to the TES evaluation was completed. School years 2000-01 through 2008-09. Clustered (teacher) standard errors in parentheses. ***p<0.001, **p<0.01, *p<0.05, +p<0.1.

Table 2: Estimates of the Relationship Between Student Test Scores in Varrying Years & TES Scores

			Math		
			Teacher's TES Score Observed in:		
	Any Previous Year (t-n)	Previous Year (t-1)	Same Year (t)	Following Year (t+1)	Any Following Year (t+n)
Average All 8 Standards	0.207***	0.246***	0.272***	0.171*	0.192***
	(0.059)	(0.061)	(0.062)	(0.071)	(0.043)
Average Domain 2 - Average Domain 3	0.067	0.126	0.047	0.249**	-0.016
	(0.067)	(0.083)	(0.100)	(0.086)	(0.061)
Standard 3.4 - (Average 2.2, 3.1 & 3.2)	-0.061	-0.124	0.001	-0.050	-0.043
	(0.068)	(0.087)	(0.085)	(0.102)	(0.064)
TES Year Fixed Effects	Y	Y	Y	Y	Y
Teacher Experience Terms	Y	Y	Y	Y	Y
Student-level Covariates	Y	Y	Y	Y	Y
R-squared	0.530	0.543	0.570	0.530	0.494
Student Sample	15,676	5,836	6,086	3,791	15,251
Teacher Sample	168	122	156	100	306

			Reading		
			Teacher's TES Score Observed in:		
	Any Previous Year (t-n)	Previous Year (t-1)	Same Year (t)	Following Year (t+1)	Any Following Year (t+n)
Average All 8 Standards	0.180***	0.204**	0.261***	0.212***	0.200***
	(0.046)	(0.066)	(0.047)	(0.052)	(0.032)
Average Domain 2 - Average Domain 3	0.032	0.002	0.063	0.147*	0.080+
	(0.059)	(0.067)	(0.061)	(0.066)	(0.046)
Standard 3.4 - (Average 2.2, 3.1 & 3.2)	0.099*	0.001	0.063	0.150*	0.110*
	(0.048)	(0.065)	(0.058)	(0.068)	(0.043)
TES Year Fixed Effects	Y	Y	Y	Y	Y
Teacher Experience Terms	Y	Y	Y	Y	Y
Student-level Covariates	Y	Y	Y	Y	Y
R-squared	0.545	0.558	0.551	0.506	0.490
Student Sample	17,375	6,136	7,522	5,739	19,393
Teacher Sample	278	191	257	206	395

Note: Each column represents a separate student-level specification. School years 2000-01 through 2008-09. Clustered (teacher) standard errors in parentheses. ***p<0.001, **p<0.01, *p<0.05, +p<0.1.

Table 3: Difference in Coefficients Under Alternative Specifications of When TES Scores Were Observed (Test of Equality p-value in Parentheses)

	Math			
	Coefficient Compared to "Following Year (t+1)"			
	Any Previous Year (t-n)	Previous Year (t-1)	Same Year (t)	Any Following Year (t+n)
Average All 8 Standards	-0.036	-0.075	-0.101	-0.021
	(0.687)	(0.398)	(0.189)	(0.734)
Average Domain 2 - Average Domain 3	0.181+	0.123	0.202+	0.264**
	(0.071)	(0.252)	(0.090)	(0.002)
Standard 3.4 - (Average 2.2, 3.1 & 3.2)	0.011	0.074	-0.051	-0.007
	(0.917)	(0.481)	(0.648)	(0.934)

	Reading			
	Coefficient Compared to "Following Year (t+1)"			
	Any Previous Year (t-n)	Previous Year (t-1)	Same Year (t)	Any Following Year (t+n)
Average All 8 Standards	0.032	0.008	-0.049	0.012
	(0.619)	(0.912)	(0.439)	(0.772)
Average Domain 2 - Average Domain 3	0.115	0.145+	0.084	0.068
	(0.156)	(0.098)	(0.304)	(0.221)
Standard 3.4 - (Average 2.2, 3.1 & 3.2)	0.051	0.149*	0.087	0.040
	(0.464)	(0.045)	(0.212)	(0.442)

Note: Each cell reports the difference between coefficients from two specifications: (i) using TES scores from the "Following Year (t+1)" minus (ii) using TES scores from the year(s) noted in the column heading. The p-value from a test of equality of coefficients is reported in parentheses. ***p<0.001, **p<0.01, *p<0.05, +p<0.1.