

## Using Student Performance Data to Identify Effective Classroom Practices

John H. Tyler\*

Eric Taylor \*\*\*

Thomas J. Kane\*\*

Amy Wooten \*\*\*

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. These findings 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 support for a "hybrid" teacher evaluation system that would use information from both classroom observations and student test scores to identify effective teachers. 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

## 1. Introduction

Research over the past decade has confirmed and quantified what most parents seem to know intuitively: teachers are perhaps the most important school-based input into the academic achievement of K-12 students (Rockoff (2004), Rivkin et al. (2005), Aaronson et al. (2007)). Given this, our ability to identify which teachers are effective and which are not is important to efforts aimed at improving the nation's schools. Research also shows, however, that there is little in a teacher's personnel file that would help administrators identify who are the best and worst teachers in any given school or school district (Rivkin et al. (2005)). What to do then? There are currently two methods that can be used for evaluating teachers. The most common method by far uses measures based on some combination of classroom observations of teachers, examinations of a teacher's products such as lesson plans, and evidence of a teacher's "professionalism". More recently some school districts have begun to show interest in so called "value-added" measures of teacher effectiveness that are based on gains in student test scores that under certain conditions are attributable to the teacher.

Each of these two methods has its own strengths and weaknesses. The observational measures can theoretically be used to evaluate teachers at any grade level and subject and can provide valuable information on areas where a teacher needs improvement. At this point, however, we lack a strong body of evidence linking these measures to actual student achievement. On the other hand, value-added measures of teacher effectiveness are, by their very construction, directly linked to student achievement, but they have their own problems. First, since these measures rely on annual tests that are comparable across years, only one-half to two-thirds of the nation's teachers teach grades and subjects that allow for the construction of value-added scores. Second, value-added scores can only rank teachers in a school or school district, they provide no information regarding what low scoring teachers should do to become "high value-added" teachers.

Given the importance of identifying effective teachers and the shortcomings of the two current teacher evaluation methods, some observers have called for a "hybrid" teacher evaluation model that would utilize information from both the current methods (see for example Gordon, Kane, and Staiger (2006) and Toch and Rothman (2008)). We address that challenge in this paper using unique data from the Cincinnati Public School (CPS) system to answer the following

questions: what classroom practices are most important in promoting student achievement as measured by growth in student achievement, and how do you best use this information to develop a hybrid teacher evaluation model?

## 2. Literature Review

### 2.1 Practice-Based Measures of Teacher Effectiveness

Practice-based measures of teacher effectiveness, evaluating a teacher by observing their work and their products of work, have a long history in the U.S. Prior to the early 1980s teacher evaluation was largely the responsibility of district and school administrators. Teacher evaluations in this era were primarily used for retaining or dismissing teachers rather than to provide information that teachers could use to improve their practice, and the criteria that administrators used for evaluation often lacked a solid research basis (Stronge and Tucker (2003), Medley, Coker and Soar (1984)). The beginning of current models of teacher evaluation can be traced to the early 1980s when “peer review” evaluation systems were launched in Toledo, Ohio and Rochester, New York (Kahlenberg (2007)). The introduction of peer review systems where teachers are evaluated by fellow teachers led to at least three important innovations in teacher evaluation: expansion of who conducts evaluations to include a teacher’s peers, the provision of feedback to teachers for the purpose of improving instruction, and a move toward evaluation criteria more tightly linked to research on what is believed to constitute good teaching practice. These advances were slow to spread, however, in part because of the resistance of the nation’s largest teachers’ union, the National Education Association (NEA) (Kahlenberg (2007)). Thus, it was only in 1997 that NEA delegates voted to drop their long-standing opposition to peer-review evaluation (Escamilla, Clarke, and Linn (2000)).

In recent years the standards-based accountability movement and a desire for increased professionalism by teacher organizations have both resulted in pressures to identify and implement high quality evaluation models. Nevertheless, Toch and Rothman (2008) point out 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.

The literature suggests that such credible systems should be based on clear, objective standards of practice; be conducted by multiple, trained evaluators; and consider multiple sources of data that is collected over time (Donaldson (2009), Goe and Croft (2009), Toch and Rothman (2008), Danielson and McGreal (2000)). Supporting the view of Toch and Rothman that most current evaluation systems fail to employ these best practices, 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). This last point highlights concerns with the evaluators themselves in many of today’s systems. They often lack training and administrators who conduct evaluations are often required to evaluate teachers who work in a subject area that is beyond their own expertise.

In spite of the potential shortcomings in practice, there are well developed models that districts can implement including standards-based evaluation. Standards-based evaluation of teaching attempts to assess teacher effectiveness based on agreed upon standards of what teachers should know and be able to do. Further, standards-based evaluation systems often are designed to provide teachers with useful feedback in order to support development of teacher practice (Heneman et al. (2006)). Used as a method of formative evaluation, observation enables evaluators to provide teachers with immediate and specific feedback about the quality of instruction in their classroom, which may be particularly useful for novice teachers (Goe and Croft (2009)).

On the other hand, standards-based performance assessments that rely on evaluator observations also have important limitations. Kimball and Milanowski (2009) suggest that evaluators play a significant role in the implementation of performance assessments and note that three characteristics of evaluators may influence the outcomes of evaluation: motivation, attitude and skill. Clearly, the purpose of evaluation is to assess teacher performance, not to reflect evaluator characteristics. Another concern related to the use of observational evaluations is the cost. Conducting observations of appropriate frequency and duration requires significant human and financial capital (Goe and Croft (2009), Papay et al., (2009), Toch and Rothman, (2008)).

The challenge then for districts like Cincinnati is to choose a well developed evaluation system that fits the goals and challenges of the district, and then devote the necessary time and resources to successfully implement and support the system. Successful and meaningful teacher evaluation then requires the support of teachers, the teachers' union, the school administration, and the community.

## 2.2 Student Test Score Measures of Teacher Effectiveness

Education researchers have long been interested in measuring a teacher's contribution to student achievement (for example Armour (1971), Hanushek (1976), Murnane and Phillips (1981), Sanders and Rivers (1996), Rockoff (2004), Rivkin, Hanushek and Kain (2005), Gordon, Kane and Staiger (2006)). While empirical strategies differ somewhat, a common objective is to isolate an estimate of a teacher's contribution to student achievement separate from the student, class, school, and other contributors.

Researchers have made considerable progress in the empirical methods of estimating a teacher's contribution to student achievement. Several strategies are now widely practiced; for example, modeling *growth* in achievement as opposed to achievement *levels*, and taking into account the hierarchical structure of school systems (McCaffrey, Lockwood, Koretz and Hamilton (2003)). This progress owes much to the proliferation of student achievement data (particularly under No Child Left Behind regulations), and advances in the software used to estimate models (e.g., hierarchical and Bayesian approaches). Nevertheless, a number of important statistical and interpretive questions remain (Todd and Wolpin (2003), McCaffrey, Lockwood, Koretz, Louis and Hamilton (2004), Raudenbush (2004)).

Researchers have long recognized the possibility that non-random assignment of students to teachers could distort measures of teacher effectiveness. Some teachers, the ubiquitous example states, are assigned better students who would have achieved highly in many different classrooms. Some researchers have questioned whether a teacher's specific contribution can be accurately estimated under such non-random assignment (Rothstein (2009)). Other researchers, recognizing the potential for bias, are more optimistic (Koedel and Betts (2009)). One recent study compared, for a small sample of teachers in Los Angeles, each teacher's (i) value added

estimate using observational methods to (ii) the achievement growth of a class randomly assigned to the teacher in an experiment. While only one case, that study found that the observational measures predicted random assignment measures with little bias—as long as the observational models controlled for each student’s prior achievement (Kane and Staiger (2008)).

In a number of studies the teacher effects seem to fade out over time (McCaffrey, Lockwood, Koretz, Louis and Hamilton (2004), Kane and Staiger (2008), Jacob, Lefgren and Sims (2008), Rothstein (2009)). That is, achievement growth this year does not appear to carry over 100 percent to next year. Hypotheses for fade out range from artifacts of empirical strategy to the heterogeneity of teacher quality within schools to the relevance of skills gained this year for skills tested next year (Kane and Staiger (2008)). Understanding the causes and structure of fade out is an emerging area of inquiry.

Additional issues and areas of innovation include the relevance of psychometric choices in the underlying standardized tests (Doran, Bates and Phillips (2009)), the quality of administrative data used to identify the student-class-teacher-school hierarchy, and alternatives to standard ordinary least squares (OLS) and hierarchical linear modeling (HLM) approaches for modeling achievement growth (Betebenner (2007)).

Except for rare opportunities (like the Los Angeles experiment mentioned above) researchers have few external criteria with which to compare these test score based measures. A few recent studies have found a relationship between teacher value added and more traditional classroom observation scores and administrator ratings ((Jacob and Lefgren (2008), Rockoff, Staiger, Kane and Taylor (2009)); suggestive evidence even if the traditional measures have their own shortcomings. And while the methodological development of teacher value added measures continues, these measures are now commonly used in a wide variety of settings from program and policy research (Gordon, Kane and Staiger (2006), Boyd, Grossman, Lankford, Loeb and Wyckoff (2008a) and (2008b), Rockoff, Jacob, Kane, Staiger (2008)) to school system management (e.g., New York City’s Teacher Data reports, and differential compensation programs in Dallas and Denver).

### 3. Observation-Based Teacher Evaluation in Cincinnati

### 3.1 Cincinnati's Teacher Evaluation System

Classroom observation data in our study come from Cincinnati's Teacher Evaluation System (TES) program. In 1997, the collective bargaining agreement between the Cincinnati Federation of Teachers and the Cincinnati Public Schools called for development of a revised system of teacher evaluation. In response Cincinnati Public Schools field tested in 1999-2000 the TES system that uses trained evaluators, a specified and research-based evaluation rubric, and includes multiple classroom observations of teachers during a year.<sup>1</sup>

For a teacher undergoing the TES process, there are generally a minimum of four evaluations conducted periodically throughout the school year by trained peer evaluators and school level administrators. In order to serve as a peer evaluator, a qualified "lead teacher" must complete extensive training that includes guidance and practice on how to collect and record evidence, and they must accurately score a videotaped teaching exercise prior to beginning their term as a peer evaluator. All new teachers are required to participate in TES during their first year in the district, and again pursuant to achieving career status.<sup>2</sup> Career status teachers are required to participate in TES every fifth year.

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, fifteen standards and 32 elements that describe the practices, skills, and characteristics that effective teachers should possess and employ. The domains cover four job categories including preparation, classroom management, pedagogical and content knowledge and application, and collegial responsibilities and engagement. The four domains over which a teacher is evaluated are: (Domain 1) Planning and Preparing for Student Learning, (Domain 2) Creating an Environment for Student Learning, (Domain 3) Teaching for Student Learning, and (Domain 4) Professionalism.

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

---

<sup>1</sup> As discussed later, some veteran teachers are only evaluated once during the school year.

<sup>2</sup> In Cincinnati, Career Status is analogous with full licensure. Teachers who have achieved full status are not subject to yearly contract renewal requirements. Essentially, career status teachers have tenure. [This footnote needs more clarification.]

assigning respective scores of 4, 3, 2, and 1 to these rubric levels.<sup>3</sup> As an example, consider the standard and element language provided for Standard 3.2 which resides in Domain 3, “Teaching for Student Learning”:

		Distinguished (4)	Proficient (3)	Basic (2)	Unsatisfactory (1)
3.2  The teacher demonstrates content knowledge by using content specific <u>Instructional strategies</u> .	A. Instructional Strategies & Content Knowledge	<ul style="list-style-type: none"> <li>Teacher <u>routinely</u> uses a broad range of multiple <u>instructional strategies</u> that are effective and appropriate to the content.</li> <li>Teacher conveys accurate content knowledge, including standards-based content knowledge.</li> </ul>	<ul style="list-style-type: none"> <li>Teacher uses <u>instructional strategies</u> that are effective and appropriate to the content.</li> <li>Teacher conveys accurate content knowledge, including standards-based content knowledge.</li> </ul>	<ul style="list-style-type: none"> <li>Teacher uses a limited range of <u>instructional strategies</u> that are effective and appropriate to the content.</li> <li>Teacher conveys some minor content inaccuracies that do not contribute to making the content incomprehensible to the students.</li> </ul>	<ul style="list-style-type: none"> <li>Teacher uses <u>instructional strategies</u> that are ineffective and/or inappropriate to the content.</li> <li>Teacher conveys content inaccuracies that contributes to making the content incomprehensible to the students.</li> </ul>

Standard 3.2 has only one element “Instructional Strategies & Content Knowledge,” which, in turn, has two components (the bullet-level items). A teacher will be evaluated on both components within the element and the result will be a standard-level score for that observation. For example, if an evaluator records that a teacher provides accurate information to students in a way that supports learning then that teacher would receive a score of 3 from the evaluator for that observation. Data from classroom observations are used in evaluating a teacher on domains 2 and 3, while evidence for domains 1 and 4 comes from the collection of documents such as lesson plans and goes into a portfolio that is reviewed by the evaluators. Only the first observation in an evaluation cycle is announced, the remaining observations may be

<sup>3</sup> The complete TES rubric is available on the Cincinnati Public Schools website: <http://www.cps-k12.org/employment/tchreval/stndsrburics.pdf>.

unannounced, and evaluators are required to submit the evaluation report to the teacher being evaluated within ten working days of the observation.

At the end of the year evaluators consider evidence from all observations and submitted evidence for a given teacher in arriving at a final formal standard score for each of the fifteen standards within domains 1-4. These end-of-year scores are based on a “preponderance of the evidence” and can take into account improvement in observed practice over the year and thus are not necessarily simple averages of the scores that a teacher received across all observations for the year. Once final standard scores are determined, evaluators use those scores to determine final Domain level scores, which are very close to, but not exactly, the simple average of the standard scores within each domain.<sup>4</sup> In their final end-of-year report teachers are provided with the final domain-level scores.

### 3.2 Data from the TES System

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.

While the only TES “scores” in the CPS personnel files are the end-of-year standard and domain scores, all of the score sheets for each observation of a teacher going back to 2000-01 are on file, and each score sheet contains the rubric language the evaluator used to score each element for a given observation of a given teacher. Because the rubric language maps, with very few exceptions, 1-to-1 onto numeric scores, we have been able to use the district’s files to create a historic file of CPS teachers’ TES scores at the element level for all teacher observations from 2000-01 through 2007-08. Teachers in the data will have scores in domains 2 and 3 that respond to each time they were observed in an evaluation year. On the other hand, because the rubrics for domains 1 and 4 are not based on classroom observation, each teacher will only have a single final standard score for the TES standards in these domains.

---

<sup>4</sup> The final domain scores are computed using the computational table found in Appendix XX.

Based on this file our present analysis employs two constructions of TES scores. The first is simply the final standard and domain level scores recorded in Cincinnati’s records. These are the formal scores reported to the teacher and used, where applicable, for consequential decisions. As described earlier, the standard scores are not a fixed mathematical function (e.g., mean) of the individual observation scores and each final domain score is close to, but not exactly the mean of the standard scores.

The second construction, and the primary focus of our analysis, uses the average of the individual classroom observation scores. This second construction will differ from the first to the extent that lead teachers apply (implicitly or explicitly) differential weights to some behaviors or observations when selecting a final standard score. In this second construction, we first calculated standard level scores for each observation by averaging all the individual behavior and practice scores within an element and then averaging the elements of each standard. Mathematically,

$$(3.1) \quad y_{so} = \frac{1}{m} \sum_{e=1}^m \left( \frac{1}{n} \sum_{b=1}^n x_{beso} \right)_{eso}$$

where  $y_{so}$  is the score (1-4) for standard,  $s$ , as measured during observation,  $o$ . Each  $x$  represents an individual score (1-4) selected by the evaluator as a result of observation,  $o$ , for behavior,  $b$ , which is a component of element,  $e$ , and standard,  $s$ . We then averaged these observation-specific standard scores,  $y_{so}$ , across all observations to obtain a single score for each standard summarizing the entire TES evaluation for a given teacher. Again mathematically,

$$(3.2) \quad \bar{y}_s = \frac{1}{l} \sum_{o=1}^l y_{so}$$

These grand average scores,  $\bar{y}_s$ , are the focus of our analysis. We do, however, explore how our main results differ when only selected observations are included, i.e., the average of just the lead teacher’s observations, the administrator’s observation<sup>5</sup>, the lead teacher’s first observation, and

---

<sup>5</sup> In a few cases teachers were observed more than once by their administrator. In these cases we used the average of the administrator observations.

the lead teacher's final observation. Since the classroom observation component of TES is only relevant in domains 2 and 3, our analysis will focus primarily on these domains.<sup>6</sup>

One additional characteristic of the TES data is important to note. Cincinnati updated the TES rubric twice during the period under study. TES evaluators used the original version from 2000-01 through 2002-03, a second version for 2003-04 and 2004-05, and the current version beginning in 2005-06. All three versions measured the same constructs using essentially the same language to describe behaviors and practices. The main differences between versions were in how the behaviors and practices were grouped into elements, standards, and domains. We restructured data from versions one and two to match the grouping structure of the current version (see Appendix A). For example, standard 3.1 in the current rubric is a combination of standards 3.1 and 1.2 in the previous version. Nevertheless, we use year fixed effects in our regressions specifications to help control the residual differences.

#### 4. Student Test Score Measures of Teacher Effectiveness in Cincinnati

##### 4.1 Student and Class Data in Cincinnati

As discussed earlier, there is substantial and growing interest in using student test scores to measure teacher effectiveness. This approach requires the regular testing of students and the ability in administrative data to link students and their test scores to teachers. Data made available to us in Cincinnati satisfy these requirements.

Paralleling the TES program years, we have panel data on Cincinnati students for the 2000-01 through 2008-09 school years. When our data begin in 2000-01 Cincinnati enrolled approximately 21,000 students in grades 3-8, but enrollment had fallen over 30 percent to approximately 14,500 by 2008-09 (Ohio Department of Education, 2009). The student-by-year observations include information on the student's gender, race or ethnicity, English proficiency

---

<sup>6</sup> In work not included here we show that information from these two domains dominates that from domains 1 and 4 when it comes to predicting teacher effectiveness.

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.

Table 1 details the specific grades and years when reading and math tests were administered. Across all 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).

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 2<sup>nd</sup> grade math and reading tests administered in 2000-01 through 2002-03, and a reading test administered to 3<sup>rd</sup> graders in the fall beginning in 2003-04. The bolded cells in Table 1 indicate outcome tests that can be paired with a baseline test.

Cincinnati Public Schools also maintains records of individual students' class schedules that include the teacher, course, and section.<sup>7</sup> 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.<sup>8</sup> For the

---

<sup>7</sup> 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.

<sup>8</sup> Infrequently a student's records would indicate 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

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. This partial data is, however, useful for our empirical approach (more in the following section) and so we include the earlier years.

#### 4.2 Constructing Test Score-Based Measures of Teacher Effectiveness

To estimate a teacher's effectiveness in raising student achievement we estimate a predicted test score for each student, calculate the difference between each student's predicted score and actual score, and take a weighted average of the differences across all of a given teacher's students. These teacher averages are commonly known as *value-added* scores.

In the first step, we estimate the following student-level regression for math achievement (and separately for reading):

$$(4.1) \quad A_{it} = \beta A_{it-1} + X_i \alpha + v_{ijkt}, \text{ with } v_{ijkt} = \mu_k + \theta_{jk} + \varepsilon_{ikt}$$

where  $A_{it}$  is the end of year math (reading) test score for student,  $i$ , taught by teacher,  $k$ , in class,  $j$ , in year,  $t$ ; and  $A_{it-1}$  is the student's prior year math (reading) test score. We also include a term for  $A_{it-1}$  interacted with each grade level. When the baseline score was missing for a student, we imputed  $A_{it-1}$  with the grade-by-year mean, and included an indicator for missing baseline score. The vector of additional controls,  $X_i$ , includes separate indicators for student (i) gender and (ii) race or ethnicity; 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; and (vii) each grade-year combination. The residual,  $v_{ijkt}$ , was assumed to vary as a function of the teacher,  $\mu_k$ , the specific class in which the teacher and student interacted,  $\theta_{jk}$ , and the student,  $\varepsilon_{ikt}$ .

---

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.

First, we estimated Equation 4.1 using OLS and including students from all grades (3-8) and years (2000-01 through 2008-09) in our data. Second, following previous work by Kane and Staiger (2008), we calculated the average residual,  $\bar{v}_{jk}$ , for each class,<sup>9</sup> and an estimate of the precision (inverse of the variance) of each class average:

$$(4.2) \quad p_{jk} = \frac{1}{\hat{\sigma}_{\theta}^2 + \left( \frac{\hat{\sigma}_{\varepsilon}^2}{n_{jk}} \right)}, \text{ with}$$

$$\hat{\sigma}_{\varepsilon}^2 = \text{Var}(v_{ijkt} - \bar{v}_{jk}), \quad \hat{\sigma}_{\mu}^2 = \text{Cov}(\bar{v}_{jk}, \bar{v}_{j-1k}) \text{ and } \hat{\sigma}_{\theta}^2 = \text{Var}(v_{ijkt}) - \hat{\sigma}_{\mu}^2 - \hat{\sigma}_{\varepsilon}^2$$

Third, for each teacher we calculated a weighted average of that teacher's various class averages. Each class was weighted by its precision,  $p_{jk}$ ; classes with more students had greater precision (smaller variance) and thus received more weight. However, we excluded classes that were unusually small (less than 5 students) or unusually large (more than 50 students) in this third step.

Our present analysis uses three different value added score constructions. All three follow the process described in the previous paragraph, and differ only in which class averages (from step 2) are included in the teacher's average (in step 3). The teacher's average is our value added score. The first value added score includes only the class or classes taught during the same year in which the teacher had a TES evaluation. We use these single-year same-year value added scores when exploring the relationship between a teacher's TES growth and value added growth over time. To the extent that TES observations measure performance in specific class of students, as opposed to performance of the teacher independent of the class, part of the estimated relationship between TES scores and value added scores will be due to the specific class not the teacher.

The second and third value added score constructions both exclude any class from the year in which the teacher participated in TES. Our second construction, and the primary focus of our analysis, is the weighted average of all classes a teacher taught the school year immediately

---

<sup>9</sup> For the school years 2000-01 through 2002-03 section identifiers were not available. Course identifiers were available. In these cases our second step calculated the average of all students taking the same course from a given teacher in a given year.

*before* the TES year. And the third construction is the weighted average of all classes a teacher taught the school year immediately *after* the TES year.

## 5. Value Added Measures and Classroom Practice

As a first step in exploring the relationship between value-added measures and measured classroom practices we asked a simple question. Without knowing which type of teacher they are evaluating, do TES evaluators record differences in the classroom practices of high and low value-added teachers?<sup>10</sup> Table 2 displays for each TES standard in domains 2 and 3 the results of t-tests of the difference in mean standard scores between (1) teachers in the upper quartile of value added versus those in the lowest value-added quartile and (2) upper quartile teachers versus teachers in the second quartile of value added.

The results in Table 2 are stark. The first row of Panel A in Table 2 shows that based on the formal end-of-year summary scores that were assigned by TES evaluators, upper quartile teachers of math value added had statistically significantly higher mean TES standard scores than lowest quartile math value added teachers in all of the standards in domains 2 and 3 except for 2.1 and 3.1. The second row in Panel A shows that the same comparisons between upper quartile teachers and second quartile math value added teachers were statistically different in all standards except 2.3, and 3.2. The third and fourth rows of the table show results from the same comparisons using the reading value added scores to group teachers into the value added quartiles. The results in Panel B use the TES scores we constructed using the averages of all observations through the year.

Table 2 tells a simple but important story: according to TES evaluators, high value added teachers are teaching differently than lower value added teachers. We note that this relationship is the same as that found by Boyd et al. (2009) in a pilot experiment they conducted in the New York City school district where trained evaluators were randomly assigned a high value added teacher and a second quartile value added teacher. The evaluators did not know about the

---

<sup>10</sup> While certain CPS teachers may have reputations of being “good” or “struggling” teachers and CPS evaluators may be aware of these reputations, it is highly unlikely that evaluators know where in the value-added distribution are the teachers they are observing.

experiment and were basing their evaluations on a different set of evaluation rubrics than that used by the TES system in Cincinnati. In this experimental setting Boyd et al. report results that are very similar to those we report in Table 2. Based on a series of comparisons between upper quartile and second quartile value added teachers they state that “Even with a small sample we find consistent evidence that high value added teachers have different instructional practices than low value added teachers.” Their small sample size does not allow them to say with confidence just what instructional practices lead to student achievement gains. While this is an issue they will address in a larger study based on their pilot, it is this exactly this issue that is the focus of this paper and to which we now turn.

## 6. A Model Relating Classroom Practice to Teacher Effectiveness

In our model of teaching effectiveness teachers at time  $t$  have an unobserved ability  $\lambda$  to foster learning. While in theory  $\lambda$  could be different for different types of students, to keep the model tractable we prefer to think of  $\lambda$  as a teacher’s ability to teach a potentially diverse classroom of students and is thus a measure of the teacher’s ability to foster average learning gains across the classroom. In our model  $\lambda$  is a scalar that is a function of potentially many different attributes of teaching.

$$(6.1) \quad \lambda_{kt} = f(x_{1k}, x_{2k}, \dots, x_{nk})$$

where  $k$  indexes teachers and the  $x$ ’s are attributes such as the ability to create an orderly environment for learning, the ability to communicate learning expectations to students, the possession of sufficient content knowledge to teach a subject well, etc. We view our measure of a teacher’s ability to raise student achievement at time  $t$  as an error prone measure of  $\lambda$ :

$$(6.2) \quad \bar{v}_{kt} = \lambda_{kt} + u_{kt}$$

In our data the  $x$ 's are measured by the TES standards in domains 2 and 3, and assuming that they are additive and linearly related to  $\lambda$ , we have

$$(6.3) \quad \lambda_{kt} = \alpha + \beta_1 S1_{kt} + \beta_2 S2_{kt} + \dots + \beta_8 S8_{kt} + w_{kt}$$

where  $S1, S2, \dots, S8$  are the eight standards that compose domains 2 and 3 in the TES evaluation system,  $\alpha$  is a constant teaching ability component that is common across all teachers that is not captured by the TES standards, and  $w_{kt}$  is the teacher-specific component of teaching ability that is not captured by the TES standards.

Substituting our observed measure of teaching ability for the unobserved  $\lambda$  we have

$$(6.4) \quad \bar{v}_{kt} = \alpha + \beta_1 S1_{kt} + \beta_2 S2_{kt} + \dots + \beta_8 S8_{kt} + \varepsilon_{kt} \text{ with } \varepsilon_{kt} = w_{kt} + u_{kt}$$

We assume that  $u_{kt}$  is uncorrelated with a teacher's scores on the TES standards. That is  $u \perp S_k$  for  $k = 1-8$ .

In Equation 6.4 the coefficients on the TES standards give the weight of each standard in predicting teacher effectiveness as measured by value added scores. With sufficient data one could estimate Equation 6.4 and use estimates of these coefficients as the optimal weights on each of the standards in a predictive model relating TES scores to teacher effectiveness as measured by a teacher's valued added score.

Currently districts, like Cincinnati, that use classroom observation systems to evaluate their teachers rely on an ad hoc weighting scheme in assigning value to the many practices and behaviors that are being observed, evaluated, and scored. For example, in Cincinnati the end-of-year domain scores that hold consequences for teachers are approximately equal to the simple arithmetic mean of the end-of-year standard scores. This makes sense if each of the TES standards is equally important in fostering student learning. Of course, if some teaching practices and behaviors are more important than others in promoting student learning then equal weighting of the standards is not the optimal design for an evaluation system. Resolving how the different classroom practices in a TES-type system should be optimally "weighted" is a step toward a "hybrid model" of teacher evaluation since in our model, this resolution combines information

on “teacher quality” from value added measures with “teacher quality” as captured by classroom observations.

A direct approach to deriving optimal predictive weights for the TES standards in domains 2 and 3 would be to use the estimated coefficients from fitting Equation 6.4 to data. Table 3 presents these estimates for both math and reading value-added scores. Each column represents a separate regression with a different dependent variable: a teacher’s effect on student achievement the year before TES participation, the year of TES participation, and the year following TES participation.

<Table 3 about here>

Of course, one problem with this approach is that the standard scores across the eight standards are all relatively highly correlated, ranging between 0.619 and 0.813. These relatively high correlations combined with the relatively small sample sizes make many of the estimates in Table 3 unstable, and ultimately hamper our ability to interpret the estimated coefficients from these regressions as meaningful or “optimal” weights for the standard scores.

In an attempt to simplify and impose some structure on the underlying teaching behaviors and practices that the TES standard scores are attempting to measure, we conducted a principal components analysis of the eight standard scores in domains 2 and 3. The first three principal components resulting from that analysis 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 3.1 and 3.2.

Our interpretation is that the first principal component captures the general importance of all eight behaviors and practices; teachers who score high on one of the eight tend to score high on the other seven. 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 conveying standards-based instructional objectives to the students and allows the teacher to demonstrate content-specific pedagogical knowledge in teaching these objectives. One might call this *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 3.1 and 3.2 from a teacher's score on standard 3.4. The correlation between the each of the three principle components and their constructed counterparts 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.

## 7. Results and Discussion

### 7.1 Relationship Between TES and Student Achievement Growth

We find that a teacher's TES scores contain information valuable for predicting student achievement growth. Table 4 reports the relationship between TES scores and value added.<sup>11</sup> Each column represents a separate OLS regression where a teacher's value added in a given time period relative to the TES year is the dependent variable and the three TES measures,

---

<sup>11</sup> Samples sizes vary somewhat because we cannot estimate value added for all teachers in all years. This limitation is due in part to gaps in the testing program during the study period, and in part because teachers new to the district participate in TES. We find generally similar results when we restrict the sample to teachers for whom we can estimate value added in all three periods, though they are estimated much less precisely. The results are most similar for the first overall TES measure.

constructed to capture the principal components, are the independent variables of interest. We include school year fixed effects in all regressions.

An overall TES score that is one point higher (i.e., an increase in the average score across all eight standards of about two standard deviations) is associated with a value added score measured *in the year before TES evaluation* that is 0.18 student standard deviations higher in math, and 0.28 in reading. Scoring higher on “classroom environment” (Domain 2) relative to “classroom practices” (Domain 3) is also associated with previous-year student gain coefficients similar in size for math and reading: 0.18 and 0.17 respectively. Last, scoring higher on *inquiry-based teaching* (Standard 3.4) relative to *standards and content focused teaching* (Standards 3.1 and 3.2) is associated with student gains in reading but not in math. In math the coefficient’s sign is opposite that for reading, but not significant.

These results suggest that a student assigned a “Distinguished” teacher would, by the end of the school year, score more than one-quarter of a standard deviation higher in reading than her peer in a class taught by a “Proficient” teacher. If, in particular, that teacher is “Distinguished” in managing classroom environment and inquiry-based teaching, the student would gain up to another one-third of a standard deviation. In math the estimates are smaller, but still impressive. If fadeout is minimal, a school staffed by “Distinguished” teachers might well close the black-white achievement gap—often estimated at one standard deviation—in a matter of years.

Not only do the behaviors and practices measured by TES predict non-trivial gains in student achievement, they also explain a good portion of the variation in achievement growth as measured by value added. The R-squared statistics indicate that the three TES score components explain 0.24 of the variation in math value added and 0.29 in reading.

The results, however, change somewhat when predicting student growth in years other than  $t-1$ . For reading, we find the same pattern of results when using value added that was measured in the TES evaluation year and year following TES though the relationships diminish, especially in the year following, and often become insignificant. For math, we find similar, though somewhat smaller, results in the year following TES, but much different results in the TES year. In the TES year the coefficient on a teacher’s overall TES score is much larger (0.24 student standard deviations), the coefficient on the second component becomes small and

insignificant, and the coefficient on the 3<sup>rd</sup> component changes sign but remains insignificant. These differences need not be unexpected, if both the TES measures and test scores measures capture variation particular to the classroom and not the teacher (e.g., a random assignment of a particularly difficult and disruptive student), we would expect some of the observed association in the TES year to be due the class. In the remainder of this paper we generally focus on student test score growth measured the year prior to the TES year as it is this measure that most closely fits the spirit of our model relating value added scores to TES scores.

The TES measures used in Table 4 are based on an average of all observations made by all evaluators during their TES year. Table 5 explores how the results from Table 4 differ when we use TES measures based on selected observations and evaluators. With a few exceptions, the results are generally robust to alternative combinations of the TES observation scores. The TES scores based on an average of all observations, however, are the most strongly associated with growth in student achievement.<sup>12</sup> Using the formal TES scores assigned by the lead teacher at the conclusion of the year produces results similar to, if slightly weaker than, the simple average of all observations. The dominance of these two scores lends support to the TES system's investment in multiple observations, at least when judged on the criteria of predicting student growth.

The individual observations, it appears, capture some variation in specific practices. For reading, individually the lead teacher's first and last observations are similarly associated with value added except for the TES measure that contrasts "inquiry-based teaching" to "content and standards focused teaching." There is also similarity between the first and last observations in math except, notably, for the TES measure that contrasts measures of "classroom environment" (Domain 2) and "classroom practices" (Domain 3). Averaging across all of a lead teacher's observations, however, appears to provide slightly more predictive value in the two TES measures which contrast standards rather than simply averaging all the standards.

In many evaluations systems, there is skepticism about what administrator observations add to the effort. The TES scores based just on the administrator's observation(s) have the weakest association when compared to all our other specifications. The administrator single

---

<sup>12</sup> Lead teachers provide all but one of the observations, thus, not surprisingly, and average of lead teacher observations produces results very similar to the average of all observations.

observation may, however, be better compared with the lead teacher's own *first* observation. In that pair wise comparison results are very similar.<sup>13</sup> Administrators, whom TES invests in through substantial training, may come to conclusions not unlike the lead teachers given additional observations. Unfortunately we lack data to test this hypothesis directly. Additionally, it is notable that the overall average of observations—which includes the administrator's observation—does appear a stronger predictor in some regards than just the average of the lead teacher's observations.

In Table 6 we separate elementary and middle grades (i.e., 3-5 and 6-8 in our sample). In math our three TES measures predict student achievement growth more strongly in elementary grades than in middle grades. In reading the coefficient on our overall TES measure is somewhat larger for middle grades, but the other two TES measures are stronger predictors in elementary grades. For both subjects the difference between elementary and middle grades is relatively large for the TES measure that contrasts measures of “classroom environment” (Domain 2) with measures of “classroom practices” (Domain 3). It is unclear what drives these differences. The teaching practices valuable in elementary grades may not, as some would argue, be equally valuable in middle grades and *visa versa*. It is possible, however, that elementary and middle teachers in our sample differ on other unmeasured characteristics (e.g., experience if the district's hiring needs varied between grade levels over the study period).

We find these results encouraging first steps for the identification of classroom practices associated with increased growth in student achievement. However, the heterogeneity across subjects and grade levels, suggests caution in extending these relationships to other subjects and to high school settings.

## 7.2 TES Growth

---

<sup>13</sup> Not all teachers were observed and scored by an administrator. When we restrict the sample to just teachers with an administrator's observation the results are similar to those reported, except the coefficient on the overall TES measure for reading. In that instance the coefficient drops by about 20 percent in each case. Thus the magnitude of the difference reported in Table 5 between administrators and other constructions of the overall TES measure should not be over interpreted.

While a teacher's TES scores from a *single year* contain information valuable for predicting student achievement growth, we now turn to the question of *changes over time* in a teacher's TES scores. A first order question is whether TES scores change over time. TES scores do, on average, increase over time. In our data, which spans 2000-01 to 2007-08, 354 teachers participated in TES twice. The average change in our overall TES measure was 0.32 points (s.d.=0.45) which is about three quarters of a standard deviation.<sup>14</sup>

One potential mechanism for TES growth is that teachers become more skilled with experience on the job and TES is able to measure this growth. A growing literature suggests that a teacher's effectiveness—as measured by growth in student achievement—improves during the first few years in the classroom, but levels off after that (see Kane, Rockoff and Staiger (2006) for a review). We find somewhat of an association between TES and experience. Table 7 reports the mean and standard deviation of our overall TES measure by years of experience. In this pooled cross-sectional sample, the average increases more from zero to three years than after the third year. The difference between 2.86 in year zero and 3.20 in year three is about three-quarters of a standard deviation, or, alternatively stated, one-third of the distance between “Proficient” and “Distinguished.” The correlation between years of experience and TES score is 0.34 in years zero to three and 0.12 in years four plus.

Over time teachers, individually and generally, have presumably become more familiar with the TES rubric and the behaviors and practices it advocates. Increased awareness of those behaviors could lead to growth in TES scores, either because the teachers invest in learning and adopting TES behaviors permanently or because teachers strategically demonstrate TES-like behaviors when under observation. The TES data alone cannot differentiate between these possibilities.

Our measures of teachers' effect on student test scores provide an opportunity to estimate the relationship between changes in value added over time and changes in TES scores over time. Unfortunately our sample is limited for this purpose. While more than 350 teachers have participated in TES twice, we can only measure value added at both points for 21 reading teachers and just 12 math teachers. Table 8 reports results for reading using this selected sample.

---

<sup>14</sup> While the amount of time between TES evaluations varied from one to seven years (with an overwhelming mode of three years) the average change was fairly constant no matter the intervening time period.

Each column represents a separate OLS regression where a teacher's value added from the year of his TES evaluation is the dependent variable and our overall TES measure is the independent variable of interest. Each teacher contributes one observation for each year they participated in TES. Thus column A is the same specification as the "TES Year" column in Table 4 except that here we only include the one TES measure. Our empirical strategy is to compare this specification with and without teacher fixed effects; thus identifying the coefficient based on the within teacher variation.

Compared with the entire sample in Table 4 the results in column A are smaller, and not significant, for this selected sample. When we add a teacher fixed effect the coefficient moves closer to zero, but we cannot reject ( $p=0.497$ ) that the coefficients from these two estimates are equal. If we retain the teacher fixed effect but remove the year fixed effect the coefficient is similar to column A, but estimated even less precisely.

Taken together these two results—(i) the relationship between TES scores and experience and (ii) the relationship between TES growth and value added growth—suggest that moving a teacher one entire rubric level (e.g., from "Proficient" to "Distinguished") may be more difficult than simply reading the rubric would suggest. Our sample is, however, extremely limited. As the sample of teachers in Cincinnati's data and the data of other district's builds, we will be better equipped to address the question of growth over time.

### 7.3 Predicting Future Impact on Student Achievement

One motivation for a hybrid approach to teacher evaluation is that combining information from student achievement growth measures and classroom observation measures may provide better predictions of future teacher effectiveness than either would singly. Cincinnati's combined TES and student achievement data allow us an opportunity to test this hypothesis.

In grades three through five, the data include 35 math teachers whose classes we observe both the year before and the year after the teacher's TES evaluation year. In reading 99 teachers meet the same criteria. Using this sample, we first estimated a value added score for each teacher

using only classes taught *before* the year of TES participation. In this step, our student-level specification

$$(7.1) \quad A_{it} = \beta A_{it-1} + X_i \alpha + \gamma TES_{k(t=0)} + v_{ijkt}, \text{ with } v_{ijkt} = \mu_k + \theta_{jk} + \varepsilon_{ijkt}$$

is identical to Equation 4.1 with two exceptions: (i) we restricted the sample such that  $t < 0$  where  $t=0$  represents the year of TES participation, and (ii) we added a control for the overall TES measure of the student's teacher,  $TES_{k(t=0)}$ , as observed during the year of TES evaluation. The addition of  $TES_{k(t=0)}$  allowed us to estimate  $\mu_k$  separate from the effect of  $TES_{k(t=0)}$  on student achievement,  $\gamma$ . We used Hierarchical Linear Models (HLM) to estimate Equation 7.1 with nested random effects for teachers and classes. HLM provides empirical Bayes estimates of the teacher random effects,  $\hat{\mu}_{k(t<0)}$ , that are the best linear unbiased predictions. These empirical Bayes estimates are “shrunk”; that is they account for differences in the reliability of the estimates from teacher to teacher by shrinking less reliable estimates toward the mean (Raudenbush and Bryk (2002)). This shrinkage reduces random measurement error that is associated with the class- and student-levels, a desirable property since we use these random effects,  $\hat{\mu}_{k(t<0)}$ , as predictors of future value added.

Second, we estimated a second value added score for each teacher this time using only classes taught *after* the year of TES participation. In this step, our student-level specification was identical to Equation 4.1 with the sample restricted such that  $t > 0$ . As before we estimated Equation 4.1 using OLS, calculated the average student residual for each class, and calculated the weighted average of the class averages for each teacher,  $\bar{v}_{k(t=1)}$ . In this case only the class or classes from the year following TES participation were included a teacher's average.

Using the products of steps one and two, we estimated the following teacher-level specifications

$$(7.2.1) \quad \bar{v}_{k(t=1)} = \alpha \hat{\mu}_{k(t<0)} + \varepsilon_k$$

$$(7.2.2) \quad \bar{v}_{k(t=1)} = \alpha \hat{\mu}_{k(t<0)} + \beta (\hat{\gamma} * TES_{k(t=0)}) + \varepsilon_k$$

$$(7.2.3) \bar{v}_{k(t=1)} = \theta(\hat{\mu}_{k(t<0)} + (\hat{\gamma} * TES_{k(t=0)})) + \varepsilon_k$$

$$(7.2.4) \bar{v}_{k(t=1)} = \beta(\hat{\gamma} * TES_{k(t=0)}) + \varepsilon_k$$

where  $\bar{v}_{k(t=1)}$  is our estimate of a teacher’s effectiveness in raising student achievement in a “future” period,  $\hat{\mu}_{k(t<0)}$  is our estimate of a teacher’s effectiveness in raising student achievement in a “past” period, and  $\hat{\gamma} * TES_{k(t=0)}$  is the predicted contribution of to student achievement captured by the TES score for teacher  $k$ .

Table 9 reports the results of estimating Equations 7.2.1-7.2.4 using OLS. For both reading and math, the estimates of 7.2.1 suggest a teacher’s effectiveness in raising student achievement in the past alone is an unbiased predictor of that teacher’s effectiveness in raising student achievement in the future. Since the estimates from pre-TES years have been “shrunk” to account for random sources of measurement error, we would expect a coefficient of one if there were no bias using those estimates as predictors. Both coefficients are very close to one, and we cannot reject the hypothesis that they are indeed one.

Adding information from a teacher’s TES evaluation increases, for this sample, the accuracy of predicting the teacher’s future effectiveness in raising student achievement (as measured by the relative r-squared). This is true in both reading and math, and true whether the TES contribution is included as a separate predictor or summed with the prior teacher effect. Adding TES information may introduce some marginal bias, but we cannot reject that the coefficients in 7.2.3 are different from 7.2.1.

The predicted contribution of TES to achievement growth itself, shown as a regressor alone in 7.2.4, is by itself a relatively weaker predictor of future teacher effectiveness in raising student achievement. The coefficient in reading is significantly different from zero, but we cannot reject that it is different from one with the same confidence. The coefficient in math is estimated much less precisely with a confidence interval that includes both zero and one.<sup>15</sup>

---

<sup>15</sup> These results may appear to contradict our main results reported in Table 4. For the broader sample used in Table 4 our overall TES measure significantly predicted value added in the year after TES for math but not for reading. While here in Table 9 for this more limited sample the opposite is true. When we re-estimate the relevant specification in Table 4 using just this limited sample the results for math and reading are indeed switched—reading

## 8. 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 showed a positive and non-trivial relationship between TES scores and value added scores. Our main results from Table 4 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 a quarter of a standard deviation. Though moving up the TES scale may be more difficult than a casual reading of the rubric 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.

Yet 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 other TES rubric elements were no longer measured or discussed. Second, a district may value outcomes for its teachers and

---

is significantly predicted but not math—matching the pattern in Table 9. Given the incongruity depending on the sample, caution should be taken in interpretations that rely on combining findings from Tables 4 and 9.

students beyond growth in standardized test scores. This latter decision deserves serious discussion, but is beyond the scope of our analysis.

Last, these results support the notion of combined or hybrid measures of teacher effectiveness for predicting future student achievement. This is true when classroom observation scores are brought into a model that previously only had student achievement measures. It is also true, perhaps more so, when student achievement measures are added to a model that only had classroom observation data. A teacher's past student achievement gains are a good predictor of future achievement gains, but measuring classroom practice likely improves the prediction. Teachers or administrators considering their future prospects for success should be open to including both forms of measuring past effectiveness.

## References

- Armour, David. T. (1976). *Analysis of the school preferred reading program in selected Los Angeles minority schools*. R-2007-LAUSD. (Santa Monica, CA: Rand Corporation).
- Betebenner, Damian W. (2007). Estimation of Student Growth Percentiles for the Colorado Student Assessment Program. Colorado Department of Education. Accessed at [http://www.cde.state.co.us/cdeassess/documents/res\\_eval/FinalLongitudinalGrowthTAPReport.pdf](http://www.cde.state.co.us/cdeassess/documents/res_eval/FinalLongitudinalGrowthTAPReport.pdf) on September 11, 2009.
- Boyd, Donald, Pamela Grossman, Hamilton Lankford, Susanna Loeb and James Wyckoff. (2008a). "Who Leaves? Teacher Attrition and Student Achievement." *NBER working paper* #14022, May 2008.
- Boyd, Donald, Pamela Grossman, Hamilton Lankford, Susanna Loeb and James Wyckoff. (2008b). "Teacher Preparation and Student Achievement." *NBER working paper* #14314, September 2008.
- 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 Personnel Evaluation in Education*, 17(3): 207-219.
- Jacob, B.A., and Lefgren, L.J. (2008). "Principals as Agents: Subjective Performance Measurement in Education" *Journal of Labor Economics* 26(1): 101-136.
- Jacob, B.A., L. Lefgren, and D. Sims. (2008). "The Persistence of Teacher-Induced Learning Gains," *NBER working paper* #14065, June 2008.
- Kahlenberg, Richard D. (2007). *Tough Liberal: Albert Shanker and the Battles Over Schools, Unions, Race, and Democracy*. Columbia University Press, NY.
- Kane, Thomas J. and Douglas O. Staiger. (2008). "Estimating Teacher Impacts on Student Achievement: An Experimental Evaluation." *NBER working paper* #14601, December 2008.
- Koedel, Cory and Julian R. Betts. (2009). "Does Student Sorting Invalidate Value-Added Models of Teacher Effectiveness? An Extended Analysis of the Rothstein Critique." University of Missouri working paper, July 2009.
- McCaffrey, Daniel, J.R. Lockwood, Daniel Koretz and Laura Hamilton. (2003).

*Evaluating Value-Added Models for Teacher Accountability.* (Santa Monica, CA: Rand Corporation).

McCaffrey, Daniel F., J. R. Lockwood, Daniel Koretz, Thomas A. Louis, Laura Hamilton. (2004). "Models for Value-Added Modeling of Teacher Effects" *Journal of Educational and Behavioral Statistics*, 29(1):67-101.

Medley, Donald M., Homer Coker, and Robert S. Soar (1984). *Measurement-Based Evaluation of Teacher Performance: An Empirical Approach.* New York: Longman.

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

Ohio Department of Education. 2009. Publicly reported enrollment data accessed at <http://www.ode.state.oh.us/GD/Templates/Pages/ODE/ODEDetail.aspx?page=3&TopicRelationID=396&ContentID=12261&Content=65241> on September 11, 2009.

Raudenbush, Stephen W. (2004). "What Are Value-Added Models Estimating and What Does This Imply for Statistical Practice?" *Journal of Educational and Behavioral Statistics*, 29(1):121-129.

Raudenbush, Stephen W. and A.S. Bryk (2002), *Hierarchical Linear Models: Applications and Data Analysis Methods*, Newbury Park, CA: Sage Publications.

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.

Rockoff, Jonah E., Brian A. Jacob, Thomas J. Kane, Douglas O. Staiger. (2008). "Can You Recognize an Effective Teacher When You Recruit One?" *NBER working paper #14485*, November 2008.

Rothstein, J., (2009). "Teacher Quality in Educational Production: Tracking, Decay, and Student Achievement," Princeton University Working Paper, May 2009.

Sanders, William L. and June C. Rivers. (1996). "Cumulative and Residual Effects of Teachers on Future Student Academic Achievement" *Research Progress Report* University of Tennessee Value-Added Research and Assessment Center.

Stronge, J.H. and P.D. Tucker. (2003) *Handbook on Teacher Evaluation: Assessing and Improving Performance.* Larchmont, N.Y.: Eye on Education.

Todd, P.E. and Wolpin, K.I. (2003). "On the Specification And Estimation of the Production Function for Cognitive Achievement," *Economic Journal*, 113(1): 3-33.

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

## Appendix

### TES Computation Tables for Assigning End of Year Domain Scores

Domains 1 & 2		Domain 3		Domain 4	
Total Standard Points	Rubric Score for the Domain	Total Standard Points	Rubric Score for the Domain	Total Standard Points	Rubric Score for the Domain
3	1	5	1	4	1
4	1	6	1	5	1
5	2	7	1	6	1
6	2	8	2	7	2
7	2	9	2	8	2
8	3	10	2	9	2
9	3	11	2	10	2
10	3	12	2	11	3
11	4	13	2	12	3
12	4	14	3	13	3
		15	3	14	3
		16	3	15	4
		17	3	16	4
		18	4		
		19	4		
		20	4		

**Table 1: Testing Program 2000-01 through 2008-09**

	(a) Reading Grade Level						
	2nd	3rd	4th	5th	6th	7th	8th
2000-01	OGPT	OGPT	SPT	OGPT	SPT	OGPT	OGPT
2001-02	OGPT	<b>OGPT</b>	<b>SPT</b>		<b>SPT</b>	<b>OGPT</b>	
2002-03	OGPT	<b>OGPT</b>	<b>SPT</b>		SPT	<b>OGPT</b>	
2003-04		<b>OAT</b>	<b>SPT</b>		SPT		
2004-05		<b>OAT</b>	<b>OAT</b>	<b>OAT</b>	SPT		OAT
2005-06		<b>OAT</b>	<b>OAT</b>	<b>OAT</b>	<b>OAT</b>	<b>OAT</b>	<b>OAT</b>
2006-07		<b>OAT</b>	<b>OAT</b>	<b>OAT</b>	<b>OAT</b>	<b>OAT</b>	<b>OAT</b>
2007-08		<b>OAT</b>	<b>OAT</b>	<b>OAT</b>	<b>OAT</b>	<b>OAT</b>	<b>OAT</b>
2008-09		<b>OAT</b>	<b>OAT</b>	<b>OAT</b>	<b>OAT</b>	<b>OAT</b>	<b>OAT</b>

	(b) Math Grade Level						
	2nd	3rd	4th	5th	6th	7th	8th
2000-01	OGPT	OGPT	SPT	OGPT	SPT	OGPT	OGPT
2001-02	OGPT	<b>OGPT</b>	<b>SPT</b>		<b>SPT</b>	<b>OGPT</b>	<b>OGPT</b>
2002-03	OGPT	<b>OGPT</b>	<b>SPT</b>		SPT	<b>OGPT</b>	
2003-04			<b>SPT</b>		SPT		
2004-05		OAT	SPT		SPT	<b>OAT</b>	OAT
2005-06		OAT	<b>OAT</b>	<b>OAT</b>	OAT	<b>OAT</b>	<b>OAT</b>
2006-07		OAT	<b>OAT</b>	<b>OAT</b>	<b>OAT</b>	<b>OAT</b>	<b>OAT</b>
2007-08		OAT	<b>OAT</b>	<b>OAT</b>	<b>OAT</b>	<b>OAT</b>	<b>OAT</b>
2008-09		OAT	<b>OAT</b>	<b>OAT</b>	<b>OAT</b>	<b>OAT</b>	<b>OAT</b>

Note: Tests listed are the Ohio State Proficiency Test (SPT) its companion Off Grade Proficiency Test (OGPT) and the replacement Ohio Achievement Test (OAT). Bolded cells indicate the the end of year outcome test score can be matched with a baseline test score from the prior school year (or prior fall in the case of 3rd grade reading since 2004-05).



**Table 3: Estimates of the Relationship Between TES Scores & Value Added**

	All Grades (3-8)					
	Previous Year	Math TES Year	Following Year	Previous Year	Reading TES Year	Following Year
Standard 2.1	-0.086 (0.125)	-0.031 (0.107)	0.018 (0.081)	0.096 (0.100)	0.118 (0.075)	-0.095 (0.087)
Standard 2.2	0.236+ (0.131)	0.147 (0.107)	0.172* (0.077)	-0.004 (0.079)	0.018 (0.082)	0.111 (0.075)
Standard 2.3	0.126 (0.086)	0.059 (0.094)	-0.029 (0.058)	0.161* (0.078)	0.019 (0.050)	0.076 (0.061)
Standard 3.1	-0.083 (0.103)	-0.081 (0.111)	-0.045 (0.083)	-0.126 (0.087)	-0.185* (0.073)	-0.104 (0.079)
Standard 3.2	0.11 (0.111)	0.115 (0.109)	-0.042 (0.069)	-0.053 (0.085)	-0.041 (0.089)	-0.158* (0.068)
Standard 3.3	0.046 (0.168)	-0.255 (0.164)	0.115 (0.118)	-0.1 (0.117)	0.104 (0.096)	0.210* (0.092)
Standard 3.4	-0.131 (0.132)	0.093 (0.092)	-0.071 (0.087)	0.179+ (0.098)	0.002 (0.062)	-0.097 (0.069)
Standard 3.5	-0.026 (0.112)	0.190+ (0.108)	0.033 (0.096)	0.113 (0.118)	0.207** (0.073)	0.122+ (0.069)
Year Fixed Effects	Y	Y	Y	Y	Y	Y
R-squared	0.279	0.252	0.179	0.308	0.141	0.099
Sample Size	82	105	123	179	222	218

Note: TES scores from 2001-02 through 2007-08, and value-added scores based on the same years plus 2008-09. Robust standard errors in parentheses. \*\*\*p<0.001, \*\*p<0.01, \*p<0.05, +p<0.1.

**Table 4: Estimates of the Relationship Between TES Scores & Value Added**

	All Grades (3-8)					
	Math			Reading		
	Previous Year (t-1)	Math TES Year (t)	Following Year (t+1)	Previous Year (t-1)	Reading TES Year (t)	Following Year (t+1)
Average All 8 Standards	0.178** (0.065)	0.237*** (0.045)	0.131** (0.047)	0.275*** (0.053)	0.193*** (0.040)	0.059 (0.048)
Average Domain 2 - Average Domain 3	0.182+ (0.108)	0.004 (0.075)	0.119* (0.060)	0.170*** (0.052)	0.091 (0.058)	0.053 (0.058)
Standard 3.4 - (Average of 3.1 and 3.2)	-0.098 (0.108)	0.083 (0.075)	0.03 (0.069)	0.170* (0.079)	0.073 (0.049)	0.021 (0.057)
Year Fixed Effects	Y	Y	Y	Y	Y	Y
R-squared	0.243	0.232	0.182	0.291	0.097	0.049
Sample Size	83	155	159	179	273	262

Note: TES scores from 2001-02 through 2007-08, and value-added scores based on the same years plus 2008-09. Robust standard errors in parentheses. \*\*\*p<0.001, \*\*p<0.01, \*p<0.05, +p<0.1.

**Table 5: Prior Year Value Added & Variations on TES Scores**

	(A) Math					
	All Observations	Formal Scores	Lead Teacher			Administrator Observation
			Average	First Obs	Last Obs	
Average All 8 Standards	0.178** (0.065)	0.153** (0.055)	0.167** (0.064)	0.156* (0.062)	0.143* (0.060)	0.126* (0.056)
Average Domain 2 - Average Domain 3	0.182+ (0.108)	0.087 (0.062)	0.206* (0.096)	0.057 (0.074)	0.229** (0.085)	0.026 (0.107)
Standard 3.4 - (Average of 3.1 and 3.2)	-0.098 (0.108)	-0.032 (0.053)	-0.109 (0.106)	-0.066 (0.057)	0.018 (0.061)	0.064 (0.072)
Year Fixed Effects	Y	Y	Y	Y	Y	Y
R-squared	0.243	0.231	0.259	0.225	0.259	0.229
Sample Size	83	83	83	83	83	63

  

	(B) Reading					
	All Observations	Formal Scores	Lead Teacher			Administrator Observation
			Average	First Obs	Last Obs	
Average All 8 Standards	0.275*** (0.053)	0.243*** (0.045)	0.266*** (0.055)	0.235*** (0.059)	0.224*** (0.047)	0.168** (0.053)
Average Domain 2 - Average Domain 3	0.170*** (0.052)	0.122*** (0.033)	0.141** (0.050)	0.091* (0.040)	0.100* (0.050)	0.126* (0.054)
Standard 3.4 - (Average of 3.1 and 3.2)	0.170* (0.079)	0.119** (0.042)	0.153* (0.066)	0.013 (0.040)	0.091* (0.044)	0.036 (0.064)
Year Fixed Effects	Y	Y	Y	Y	Y	Y
R-squared	0.291	0.305	0.276	0.216	0.252	0.188
Sample Size	179	179	179	179	179	140

Note: TES scores from 2001-02 through 2007-08, and value-added scores based on the same years plus 2008-09. Robust standard errors in parentheses. \*\*\*p<0.001, \*\*p<0.01, \*p<0.05, +p<0.1.

**Table 6: Prior Year Value Added & TES Scores by Grade Level**

	(A) Math			(B) Reading		
	All Grades	Elementary Grades (3-5)	Middle Grades (6-8)	All Grades	Elementary Grades (3-5)	Middle Grades (6-8)
Average All 8 Standards	0.178** (0.065)	0.232* (0.110)	0.115 (0.102)	0.275*** (0.053)	0.268*** (0.060)	0.319** (0.115)
Average Domain 2 - Average Domain 3	0.182+ (0.108)	0.302+ (0.170)	0.179 (0.136)	0.170*** (0.052)	0.219*** (0.057)	-0.068 (0.174)
Standard 3.4 - (Average of 3.1 and 3.2)	-0.098 (0.108)	-0.203 (0.143)	0.000 (0.161)	0.170* (0.079)	0.168+ (0.093)	0.200 (0.133)
Year Fixed Effects	Y	Y	Y	Y	Y	Y
R-squared	0.243	0.263	0.173	0.291	0.294	0.295
Sample Size	83	49	36	179	152	36

Note: TES scores from 2001-02 through 2007-08, and value-added scores based on the same years plus 2008-09. Robust standard errors in parentheses. \*\*\*p<0.001, \*\*p<0.01, \*p<0.05, +p<0.1.

**Table 7: TES & Teaching Experience**

	Overall TES Measure		
	Mean	S.D.	N=
0 years experience (first year)	2.86	(0.389)	185
1 year experience	2.97	(0.484)	67
2 years experience	3.04	(0.423)	79
3 years experience	3.20	(0.401)	245
4 years experience	3.17	(0.435)	108
5 to 9 years experience	3.23	(0.414)	303
10 to 14 years experience	3.26	(0.467)	329
15 to 19 years experience	3.33	(0.435)	233
20 or more years experience	3.33	(0.430)	151

Note: All teachers evaluated by TES from 2000-01 through 2007-08.

**Table 8: Changes in TES Scores & Value-added Over Time, Reading**

	(A)	(B)	(C)
Average All 8 Standards	0.061 (0.085)	0.005 (0.176)	0.079 (0.129)
Teacher Fixed Effects	N	Y	Y
Year Fixed Effects	Y	Y	N
R-squared	0.108	0.52	0.428
Sample Size	43	43	43

Note: Teachers with TES scores more than once between 2002-2008 and value-added estimates in the same year. Clustered standard errors in parentheses. \*\*\* $p < 0.001$ , \*\* $p < 0.01$ , \* $p < 0.05$ , + $p < 0.1$ .

**Table 9: Predicting Teacher Effects the Year Following TES Using Prior Estimates of Teacher Effects and TES Scores**

	Teacher Effect Estimate in Year Following TES							
	Math				Reading			
	(7.2.1)	(7.2.2)	(7.2.3)	(7.2.4)	(7.2.1)	(7.2.2)	(7.2.3)	(7.2.4)
(1) Teacher Effect Estimate in Pre-TES Years	0.986*	0.997*			0.954***	0.952***		
	(0.413)	(0.398)			(0.184)	(0.174)		
(2) Predicted TES Contribution in Pre-TES Years		0.572		0.539		0.598*		0.604+
		(0.598)		(0.586)		(0.295)		(0.335)
(3) 1 + 2			0.841*				0.827***	
			(0.388)				(0.144)	
R-squared	0.139	0.166	0.156	0.024	0.127	0.155	0.148	0.028
Sample Size	35	35	35	35	99	99	99	99

Note: Teachers of grades 3-5. Prior value-added estimated using data in years before TES evaluation. Future value-added estimated using data in years following TES evaluation. Robust standard errors in parentheses. \*\*\*p<0.001, \*\*p<0.01, \*p<0.05, +p<0.1.