Identifying Effective Classroom Practices Using Student Achievement Data

Thomas J. Kane
Eric S. Taylor
John H. Tyler
Amy L. Wooten

ABSTRACT

Research continues to find large differences in student achievement gains across teachers' classrooms. The variability in teacher effectiveness raises the stakes on identifying effective teachers and teaching practices. This paper combines data from classroom observations of teaching practices and measures of teachers' ability to improve student achievement as one contribution to these questions. We find that observation measures of teaching effectiveness are substantively related to student achievement growth and that some observed teaching practices predict achievement more than other practices. Our results provide information for both individual teacher development efforts, and the design of teacher evaluation systems.

I. Introduction

More than three decades ago, researchers including Hanushek (1971) and Murnane and Phillips (1981) began reporting large differences in student achievement gains in different teachers' classrooms. The literature on teacher effectiveness has undergone a resurgence in recent years as school districts and state governments have begun to track achievement gains of similar students assigned to

Authors are listed alphabetically. Thomas J. Kane is professor of education and economics at the Harvard Graduate School of Education, Eric S. Taylor is a research analyst at the Center for Education Policy Research at Harvard University, John H. Tyler is an associate professor of education, public policy, and economics at Brown University, and Amy L. Wooten is a doctoral candidate at the Harvard Graduate School of Education. *The authors would like to thank Douglas Staiger, Ron Ferguson, two anonymous reviewers, and seminar participates at Brown University, Harvard University, and Stanford University for helpful comments on previous drafts of this paper. They also gratefully acknowledge the Joyce Foundation for their generous support of this project, as well as the cooperation and support of the Cincinnati Public Schools.*
[Submitted March 2010; accepted September 2010]
ISSN 022-166X E-ISSN 1548-8004

different teachers (Aaronson, Borrow, and Sander 2003; Gordon, Kane, and Staiger 2006; Kane, Rockoff, and Staiger 2006; Rivkin, Hanushek, and Kain 2005; Rockoff 2004). The magnitude of the variation in teacher effects is quite large, with most estimates of the standard deviation ranging between 0.10 and 0.25 student-level standard deviations in math (with somewhat smaller differences reported for English language arts, see Hanushek and Rivkin 2010).

The size and consistency of these findings—especially when combined with rising anxiety about the lagging performance of U.S. students in international comparisons—has produced a flurry of policy proposals to promote "teacher quality" or "teacher effectiveness." Despite the outpouring of interest, little has changed in the way that teachers are evaluated and compensated, in the content of preservice training, or in the type of professional development offered.

The primary stumbling block has been a lack of consensus on valid measures for recognizing and rewarding effective teaching. On one hand, a handful of districts have begun using student achievement gains (adjusted for prior achievement and other student characteristics) as a direct measure of teacher effectiveness (for example, Hillsborough County Florida, Dallas and Houston in Texas, Denver Colorado, New York City). However, even supporters of such policies recognize their limitations. First, these estimates—sometimes referred to as "value added" estimates—are currently feasible only in a handful of grades and subjects, where there is mandated annual testing. In fact, less than a quarter of K-12 teachers are likely to be in grades and subjects where such measures are possible. Second, in the absence of evidence of effective *teaching practices*, test-based measures by themselves offer little guidance that could direct teacher training or professional development. Test-based measures allow one to identify effective teachers on the job, but not to replicate them. Third, there is the danger that teachers will focus on teaching test-taking skills at the cost of teaching other more valuable skills if they are not provided with clear signals about legitimate ways to improve their practice. Finally, some have questioned whether variation in student achievement growth from teacher to teacher reflects "teacher effects" or something different, such as unmeasured differences in baseline characteristics between different classrooms (Rothstein 2010).

On the other hand, there are, as yet, few alternatives to test-based measures that could provide reliable and valid information on the effectiveness of a teacher's classroom practice. Despite decades of evidence that teachers differ in their impacts on youth, our efforts at evaluating teacher effectiveness through direct observation of teachers in the act of teaching remains a largely perfunctory exercise. In a recent analysis of the teacher evaluation systems in 14 school districts, Weisberg, Sexton, Mulhern, and Keeling (2009) report that most districts have only a binary rating system, with more than 98 percent of teachers rated the highest category (usually labeled "satisfactory"). Based on such findings, many have questioned whether classroom observations are a hopelessly flawed approach to assessing teacher-effectiveness. In this paper, we test whether classroom observations—when done by trained professionals, external to the school, using an elaborated set of standards—can identify teaching practices most likely to raise achievement. Using data from the Cincinnati Public Schools' (CPS) Teacher Evaluation System (TES), we find that evaluations based on well-executed classroom observations *do* identify effective teachers and teaching practices. Teachers' scores on the classroom observation component of

Cincinnati's evaluation system predict achievement gains in both math and reading. Among teachers with similar overall scores, differences in specific practices also predict differences in achievement, though the practices differ for math and reading. These findings support the idea that "teacher effectiveness" need not be measured based on student achievement gains alone, but that it should be possible to build a system that incorporates measures of practice as well.

II. Relating Measures of Practice to Student Achievement

A. Limitations of Existing Evidence

While the quality and availability of student achievement data have increased noticeably over the past decade, the incidence of quality classroom observation data lags far behind. Owing primarily to the latter, there are relatively few studies relating student achievement to discrete, intentionally observed teacher classroom practices. Holtzapple (2003) and Milanowski (2004a and 2004b), using early data from the same school district and observation program we study here, found a positive relationship between teachers' final evaluation scores, which are a function of classroom observations, and student achievement growth. Grossman et al. (2010) compared the observed instructional practices of teachers identified as top-quartile in terms of raising student achievement to those identified as second-quartile; the former group scored higher on all observed practices. Absent data derived directly from classroom observation, other researchers have studied the relationship between student achievement and subjective ratings of teachers by their administrator (Jacob and Lefgren 2008; Rockoff, Staiger, Kane, and Taylor 2010), or a peer or mentor teacher (Rockoff and Speroni 2010). These works also find positive, meaningful correlations between teacher ratings and teachers' ability to raise student achievement.

The relatively consistent results from such studies are encouraging, but since they use either summary scores or subjective teacher ratings on general attributes, little can be said about *which* specific classroom practices employed by teachers are most important in promoting achievement. Data accumulated by Cincinnati Public Schools' Teacher Evaluation System (TES) over the last ten years offer the opportunity to address this gap. The TES data used in our analyses derive from classroom observations conducted by trained evaluators who use a detailed rubric to score the practices they observe. For each of (typically) four observations over the course of the school year (one by a school administrator, the rest by a peer evaluator), teachers are scored on dozens of discrete practices. Our data offer an additional advantage over previous studies in that the longitudinal nature of the data at our disposal allows us to observe student achievement and teacher practices in separate classrooms. The dynamics of a particular class of students may positively (negatively) but independently affect *both* observed teaching practices and student achievement creating the potential for overestimating the strength of their relationship; by using separate classes we can avoid this potential problem.

B. Estimation Challenges

In this paper we present estimates of the relationship between specific classroom practices, as measured by the TES process, and student achievement gains, as measured by end-of-year standardized state tests. There are at least two challenges in estimating such relationships: the nonrandom assignment of students and teachers to each other and the nonrandom assignment of observed classroom practices across teachers.

That schools assign students to teachers nonrandomly, even within subjects and grade-levels, should surprise no one. Student-to-teacher assignment dynamics are a core consideration of any effort to study systematic teacher influences on students. Of late the empirical evidence on the potential for bias from sorting on student characteristics has benefited from widespread efforts to estimate individual teachers' *total* contribution to student achievement growth—often called "teacher value-added" studies. Researchers have made considerable progress in the empirical methods of estimating a teacher's contribution, and a few basic strategies, most notably controlling for students' prior achievement in a flexible way, are now widely practiced (McCaffrey, Lockwood, Koretz, and Hamilton 2003; Hanushek and Rivkin 2010). Nevertheless Rothstein (2010), among others, has demonstrated reasons to remain cautious. The general concern is that systematic assignment of students with high (low) *unobserved* potential for achievement growth, even conditional on prior achievement, will lead to misestimates of a teacher's total contribution to achievement growth, their "value-added" score. Recent evidence suggests that, empirically, sorting on student unobservables does not introduce large bias, particularly when teachers are observed teaching multiple classes of students (Kane and Staiger 2008; Kodel and Betts 2009).

Empirically we find little correlation between teachers' TES scores and observable characteristics of the students they are assigned. However, lacking exogenous variation in the assignment of teachers and students to each other, we respond to this potential threat by examining the sensitivity of our results to different specifications including controlling for observable student characteristics and other observable teacher characteristics, fitting school fixed effects models, and fitting teacher fixed effects models for the subsample of teachers who were evaluated by TES during two (or more) years in our data. Estimates of the relationship between classroom practices and student achievement do appear to be biased upward by nonrandom sorting, but remain positive and significant even in models that control for school or teacher fixed effects.

A second challenge to estimation is that, even if students and teachers were randomly assigned to each other, teachers who systematically engage in certain classroom practices also may have other *un*observed traits that promote (hinder) student achievement growth. Failure to control for these unobserved traits would, of course, lead to biased estimates of the causal impact of TES practices on student achievement growth. This problem of inference confronts virtually all studies that have sought to understand which particular teacher characteristics (for example, classroom experience and credentials) explain the variation in "total teacher effects" on student achievement gains.

We address this challenge in two ways. First, we control for experience, the one observable teacher characteristic that has consistently been found to be associated

with student achievement (Rockoff 2004; Gordon, Kane, and Staiger 2006). Second, the teacher fixed effects models that we estimate control for all time-invariant teacher characteristics that might be correlated with both student achievement growth and observed classroom practices. As we discuss, results from these models do not support a hypothesis that latent teacher characteristics explain our primary estimates. We discuss remaining caveats along with the presentation of our results.

III. Data

A. Data from the TES System

In place since the 2000–2001 school year, Cincinnati Public Schools' Teacher Evaluation System (TES) is a case study in high-quality teacher observation programs. During the year-long TES process, teachers are typically observed and scored four times: three times by an assigned peer evaluator—high-performing experienced teachers external to the school—and once by a local school administrator.[1] Both peer evaluators and administrators complete an intensive TES evaluator training course, and must accurately score videotaped teaching examples to check inter-rater reliability.

All new teachers are required to participate in TES during their first year in the district, must do so again to achieve tenure (typically in the fourth year), and every fifth year thereafter. Teachers tenured before 2000–2001 are being phased into the five-year rotation. Additionally, teachers may volunteer to participate; most volunteers do so to post the higher scores necessary to apply for select positions in the district (for example, lead teacher or TES evaluators).

The value of the TES data for this study is that the TES scoring rubric—based on Charlotte Danielson's *Enhancing Professional Practice: A Framework for Teaching* (1996)—used by the evaluators contains quite specific language describing classroom practices. In this analysis we focus on Domain 2 "Creating an Environment for Learning" and Domain 3 "Teaching for Learning" of the TES evaluation framework. These are the only two TES domains (of four total domains) that address classroom practices, and they are the only two domains that are used in scoring each of the four classroom observations.[2] Domains 2 and 3 include over two dozen specific practices grouped into eight "standards" of teaching. For each individual practice the rubric provides language describing what performance looks like at each scoring level: "Distinguished" (a score of 4), "Proficient" (3), "Basic" (2), and "Unsatisfactory" (1).

Table 1 reproduces the rubric language associated with a rating of "Proficient" on each of the different practices in the standards of Domains 2 and 3. For example, Standard 2.1 includes two practices both related to how a teacher goes about creating "an inclusive and caring environment" for students. A teacher would be rated on both of these practices after each observation. To provide a sense of how an evaluator

1. 97 percent were observed and scored between two and six times.
2. In Domains 1 and 4 teachers are scored on artifacts, such as lesson plans and records of communications with parents, which teachers compile and submit to TES evaluators.

Table 1
Eight TES Standards and Associated "Proficient" Practice Descriptions

Domain 2. Creating an Environment for Learning	
2.1 The teacher creates an inclusive and caring environment in which each individual is respected and valued.	• Teacher interactions with all students demonstrate respect. Interactions are inclusive and appropriate. • Teacher encourages respectful interactions among individuals and appropriately addresses any disrespectful interactions among individuals.
2.2 The teacher establishes effective routines and procedures, maintains a safe and orderly environment, and manages transitions to maximize instructional time.	• Teacher establishes and uses effective routines and procedures for managing student groups, supplies, and/or equipment. • Teacher acts to maintain a safe environment. • Teacher establishes and directs procedures for transitions. No instructional time is lost.
2.3 The teacher manages and monitors student behavior to maximize instructional time.	• Teacher monitors student behavior at all times which promotes individual, group, and/or whole class time on task. • Teacher response to misbehavior is appropriate and consistent.
Domain 3. Teaching for Learning	
3.1 The teacher communicates standards-based instructional objectives, high expectations, instructive directions, procedures, and assessment criteria.	• Teacher writes lesson plans with clear and measurable standards-based instructional objectives. • Teacher selects and designs instructional activities that are aligned to the instructional objective, establish high expectations for student performance, provide opportunities for students to make continuous progress toward meeting the standards, and makes connections within or across disciplines. • Lesson plans are aligned with the lesson observed. • Teacher clearly and accurately communicates standards-based instructional objectives. • Teacher clearly and accurately communicates instructional directions and procedures for the activity. • Teacher communicates high expectations for standards-based student work. • Teacher emphasizes completion of work and encourages students to expend their best effort. • Teacher clearly communicates to students the assessment criteria that are aligned with the standards-based instructional objectives.

(continued)

Table 1 *(continued)*

3.2 The teacher demonstrates content knowledge by using content specific instructional strategies.	• Teacher uses instructional strategies that are effective and appropriate to the content. • Teacher conveys accurate content knowledge, including standards-based content knowledge.
3.3 The teacher uses standards-based instructional activities that promote conceptual understanding, extend student thinking, and monitors/adjusts instruction to meet individual needs.	• Teacher uses challenging standards-based activities at the appropriate cognitive level that promote conceptual understanding. • Teacher creates situations that challenge students to think independently, and creatively or critically about the content being taught. • Teacher monitors and adjusts instruction/activities/ pacing to respond to differences in student needs. • Teacher pursues the active engagement of all students.
3.4 The teacher engages students in discourse and uses thought-provoking questions aligned with the lesson objectives to explore and extend content knowledge.	• Teacher initiates and leads discourse at the evaluative, synthesis, and/or analysis levels to explore and extend the content knowledge. • Teacher asks thought-provoking questions at the evaluative, synthesis, and/or analysis levels that focus on the objectives of the lesson. • Teacher seeks clarification through additional questions. • Teacher provides appropriate wait time.
3.5 The teacher provides timely, constructive feedback to students about their progress toward the learning objectives using a variety of methods, and corrects student errors/ misconceptions.	• Teacher provides accurate, specific and timely feedback to students about their progress toward the learning objectives. • Teacher provides feedback using a variety of methods and facilitates student self-assessment. • Teacher corrects student content errors to individuals, groups, and/or the whole class by offering explanations that clarify the process or concept. • Teacher addresses content-related misconceptions as they arise.

Source: Cincinnati Public Schools Teacher Evaluation System 2005.

might apply the rubric language, compare, for the first practice in Table 1 (the first bulleted item), teaching practice that would lead to a rating of Distinguished, Proficient, Basic, or Unsatisfactory.[3]

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

- Distinguished: "Teacher interactions with all students demonstrate a positive, caring rapport and mutual respect. Interactions are inclusive and appropriate."
- Proficient (reproduced in Table 1): "Teacher interactions with all students demonstrate respect. Interactions are inclusive and appropriate."
- Basic: "Teacher interactions with students are generally appropriate."
- Unsatisfactory: "Teacher interactions with students are negative, demeaning, and/or inappropriate."

Cincinnati provided us records of each classroom observation including the scores evaluators assigned for each practice as a result of that observation, and an identifier for each evaluator. Using these data we calculated eight TES "standard"-level scores for each teacher (teacher-by-year if they were evaluated more than one year) as follows: first, we averaged the scores across evaluator observations for each individual practice; and, second, we averaged the individual practice scores within a standard.[4]

We estimate that differences in evaluator (different peer evaluators and administrators) account for about one-quarter (23 percent) of the variation in TES scores from individual observations, and that the teachers being evaluated account for just under half (45 percent) of the variation. The remaining variation (32 percent) is due to residual variation from observation to observation for the same teacher.[5] In analyses not shown here but available upon request, we find that a teacher's TES scores tend to increase over the course of the evaluation year, with the largest within year gains accruing to inexperienced teachers, particularly those in their first and second years.[6]

B. Student and Class Data in Cincinnati

In addition to TES observation data, we have panel data on Cincinnati students for the 2003–2004 through 2008–2009 school years. The student-by-year observations include information on the student's gender, race/ethnicity, English proficiency status, participation in special education or gifted and talented programs, class and teacher assignments by subject[7], and standardized test scores.

4. In the TES program a teacher's formal evaluation is based on her final, overall score in each of the four domains. Each final domain-level score is the rounded average of the constituent standard-level scores, but the evaluators have latitude in adjusting the final standard-level scores so that they are not perfect averages of the individual observation scores. These four scores are kept in the teacher's personnel file, and are the more commonly cited scores in research on TES.

5. Scores from two sequential observations of the same teacher conducted by her/his assigned peer evaluator are correlated 0.74. Scores from two sequential observations where one was conducted by the administrator and one by the peer evaluator are correlated 0.59. These and other additional results discussed in this paper are available from the authors in an appendix posted online.

6. Evidence on these patterns comes from a teacher fixed effects model that fits a teacher's overall TES score from each classroom observation to the week of the school year in which the observation occurred, teacher experience indicators, and the interactions of those measures.

7. We could not identify a class for 3 percent of students in math, and 4 percent in reading. These students were almost always missing from the class schedule data entirely, or, much less frequently, did not have a class listed in one or the other subject.

Between 2003–2004 and 2008–2009 Cincinnati students, in general, took end of year exams in reading and math in third through eighth grades.[8] Over the course of 2003–2004 to 2005–2006 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. In 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). Our empirical strategy requires both an outcome test (end of year test in school year t) and a baseline test (end of year test in school year t-1). Thus, our analysis sample will exclude some entire grade-by-year cohorts for whom the state of Ohio did not administer a test in school year t or t-1.[9]

IV. Empirical Strategy

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, γ, formalized in Equation 1, between the elements of Λ_{kt} and A_{ikt}, the achievement of student i taught by teacher k during school year t, conditional on student i's prior achievement, $A_{i,t-1}$, and observable characteristics, X, of student i that might also influence achievement or assignment to teachers,

$$(1) \quad A_{ikt} = \alpha + \Lambda_{kt}\gamma + A_{i,t-1}\beta + X_{it}\delta + v_{ikt}$$

While the elements of a teacher's true Λ_{kt} are unobserved, one could sample a teacher's practices by visiting his classroom. Records of such observations, including Cincinnati's extensive TES data, are potentially useful, even if error prone, measures of Λ_{kt}. In Equation 2 we substitute our observed information about teacher k's practices, the vector TES_{kT}, in place of Λ_{kt}, and add the term w_{kT} to capture the error in using TES_{kT} as a measure of Λ_{kt}.

$$(2) \quad A_{ikt} = \alpha + TES_{kT}\gamma + A_{i,t-1}\beta + X_{it}\delta + \varepsilon_{ikt}$$
$$\varepsilon_{ikt} = w_{kT} + v_{ikt}$$

If w_{kT} is simple measurement error, then our estimates suffer attenuation bias associated with traditional measurement error. Lacking instruments for *TES,* there is

8. Third grade students also took a beginning of year exam in reading allowing us to measure reading growth in third grade, but not math growth.

9. Beginning in the fall of 2003–2004, third grade students were administered a beginning of year test in reading which we use as the baseline test for appropriate cohorts. We also have test scores from 2002–2003 to serve as a baseline for the 2003–2004 school year.

little we can do in this situation other than report our estimates as lower bound estimates. Importantly, however, the added error component, w_{kT}, may be related to the specific group of students the teacher is observed teaching when her TES scores are measured. Because of this possibility estimates of γ may be biased when using samples where $t=T$ to estimate Equation 2. The potential bias arises because 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. 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 these two classes have identical values of Λ_{kt}. 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 independent of the students he taught. Additionally, the same unusual social cohesion in Class B also may result in positive peer effects that raise achievement independent of the teacher's contribution.

Thus, while using TES measures and student achievement measures from the same school year ($t=T$) may a priori be the most intuitive approach, we instead estimate Equation 2 using observations of student achievement from teachers' classes in school years near but not contemporaneous with TES measurement ($t \neq T$).[10] Specifically, we focus on the achievement of students a teacher taught within two years (before or after) his TES scores were measured ($|T-t|=1$ or $|T-t|=2$).[11]

We include student achievement observations from school years both before *and* after a teacher's participation in TES in part to address constraints imposed by the structure of the TES program; constraints that potentially create a selected sample. The sample of teachers we observe teaching students *after* participating in TES may be better on average because teachers with poor TES scores may be asked to leave (or choose to leave) the district.[12] However, that positive selection is balanced to some extent by the teachers with particularly good TES scores who also leave the classroom after participating in TES to fill nonteaching roles in the district that require high TES scores. By contrast, the sample of teachers we observe teaching students *before* participating in TES is likely more experienced on average because all first year teachers participate in TES. However, most teachers participate in TES

10. In theory the challenge we describe could be overcome by using two or more different classes taught in the same school year, one where TES scores were measured and the other(s) where achievement was measured. The TES data do not allow us to pursue this approach.

11. In results not reported here but available upon request, we find that our results are robust to using a broader window (adding $|T-t|=3$, $|T-t|=4$, etc.).

12. From 2003–2004 through 2008–2009, 63 Cincinnati teachers participated in TES because they had been identified for "Intervention," where the stakes of the TES evaluation include termination; 12 of those 63 were dismissed formally and another 23 left the district otherwise (The New Teacher Project 2009, p. 33).

again during their fourth year in the district, allowing for observations of many second or third year teachers classrooms in a *before* TES sample.

While some of the differences implied by these dynamics are unobservable, our sense is that a combined sample including observations both before and after TES participation should better represent the district's teachers. That assumption is supported by a comparison of observable characteristics in Table 2. Our combined two-year before-or-after window sample compares favorably with other teachers and students in the district at the same time (compare Columns 1 and 2 in both the math and reading panels). Teachers in both math and reading samples are, on average, less experienced than the rest of the district's teachers (approximately one-half of our sample have less than ten years experience, compared to one-quarter of all other teachers). This is not surprising because veteran teachers participate in TES, according to a phase-in schedule that started with more recently hired teachers and is proceeding backward through the rolls. Additionally, baseline test scores for our sample are somewhat lower in math (0.04 standard deviations, significant at $p < 0.01$), but not in reading. By contrast, when we separate our estimation sample into three mutually exclusive groups—teachers observed teaching students only before TES participation (Column 3), before and after participation (Column 4), and only after participation (Column 5)—we see many more differences in observable characteristics. Later we report our estimates for each of these three subsamples; in general they are similar despite these observable differences.

Our use of noncontemporaneous observations also requires further consideration of teacher experience levels in our model. When using TES scores measured in the past ($T < t$) to proxy for current practices, our proxy will be inaccurate to the extent the teacher's practices have improved with the experience of the intervening school years. Similarly, when using TES scores measured in the future ($T > t$) our proxy will inaccurately include experience the teacher has not yet accumulated. These differences in classroom experience will more acutely affect our estimates for young teachers where the marginal returns to experience are relatively large (see Kane, Rockoff, and Staiger 2006 for a review). In response we include two measures of teacher k's experience in the vector TES_{kT}: experience at the time of the TES observation, T, and experience at the time of the student achievement observation, t. As we will show later, our results are not particularly sensitive to the exclusion of these experience terms.

We estimate Equation 2 using OLS. As suggested above, A_{ikt} is the end of year math (reading) test score for student i taught by teacher k during school year t. The vector $A_{i,t-1}$ 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 (functionally grade-by-year fixed effects). When the baseline score was missing for a student, we imputed $A_{i,t-1}$ with the grade-by-year mean, and included an indicator for missing baseline score. The 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 (iii) retained in grade or participating in (iv) special education, (v) gifted, or (vi) limited English proficient programs.

To this point we have not discussed in detail the nature of the scores included in the TES_{kT} vector. One intuitive approach would be to simply include the eight TES

Table 2
Observable Characteristics of Student and Teacher Samples

	(A) Math					(A) Reading				
			Teachers Observed[a]					Teachers Observed[a]		
	Not In Estimation Sample	Estimation Sample	Only Before TES	Before & After TES	Only After TES	Not In Estimation Sample	Estimation Sample	Only Before TES	Before & After TES	Only After TES
	(1)	(2)	(3)	(4)	(5)	(1)	(2)	(3)	(4)	(5)
Student characteristics										
Baseline test score	0.102	0.061	0.125	0.127	−0.039	0.060	0.056	−0.074	0.065	0.120
Standard deviation	(1.030)	(0.936)	(0.982)	(0.945)	(0.895)	(0.993)	(0.982)	(0.960)	(0.990)	(0.976)
Grade 3						23.3	22.3	21.0	25.2	18.6
Grade 4	25.5	24.1	19.2	27.8	22.3	21.7	22.6	20.0	29.9	12.8
Grade 5	18.0	20.9	17.0	25.6	17.7	18.0	16.9	17.7	19.2	12.7
Grade 6	12.8	15.4	16.5	11.9	18.6	13.4	12.9	16.8	12.8	10.8
Grade 7	23.6	22.4	28.9	19.2	22.8	14.2	12.8	11.2	7.5	22.0
Grade 8	20.1	17.2	18.3	15.5	18.6	9.3	12.4	13.4	5.3	23.0

Racial/ethnic minority	75.9	77.8	79.3	77.3	77.6	76.7	78.2	79.7	79.3	75.9
Special education	23.4	20.6	22.4	20.5	19.9	23.3	20.9	24.7	19.9	20.0
English language learner	3.0	2.7	1.5	2.6	3.4	2.9	2.5	3.9	2.4	1.7
Ever retained	18.9	21.2	21.6	20.2	22.1	18.7	20.1	24.0	18.8	19.6
Number of students	28,607	16,196	2,933	6,867	6,396	42,372	20,125	3,864	9,941	6,320
Teacher characteristics										
First year teachers	0.1	0.3	0.0	0.0	0.8	0.5	0.1	0.0	0.3	0.0
1 year experience	1.1	3.1	5.1	1.9	3.3	1.4	3.3	6.1	1.4	4.8
2 years experience	1.8	3.7	8.9	3.2	0.8	2.0	5.4	8.8	4.8	3.8
3 years experience	2.6	2.0	3.8	1.3	1.7	3.7	3.2	4.7	2.8	2.7
4 years experience	2.3	8.5	3.8	5.2	15.7	2.5	7.3	3.4	5.9	12.9
5–9 years experience	16.4	27.0	19.0	20.0	41.3	14.5	25.8	19.6	21.8	38.2
10+ years experience	75.7	55.5	59.5	68.4	36.4	75.4	54.9	57.4	62.9	37.6
Number of teachers	494	207	64	56	87	813	365	112	117	136

a. "Teachers Observed": Teachers who taught students in a tested (and pretested) subject, grade, and school year where "before" and "after" are relative to the TES participation year.

standards scores from Domains 2 and 3 individually. In practice, however, the scores across these eight standards are highly correlated (pair-wise correlations ranging between 0.619 and 0.813) so that estimates of the effects of individual standards' coefficients (the γs) tend to be unstable and difficult to interpret. In those results (not included here, but available upon request) very few of the coefficients are statistically significant and many are wrong signed (assuming all the TES rubric practices are supposed to positively impact student achievement).

Given the highly correlated nature of the standard scores, we instead include three composite TES scores in the vector TES_{kT}; these composite scores were derived from a principal components analysis of the eight standards in Domains 2 and 3, and are thus constructed to be uncorrelated. The first principal component score, which we call *Overall Classroom Practices*, includes all eight standards almost equally weighted, capturing the general importance of the specific teaching practices measured by Domains 2 and 3. The second, which we call *Classroom Environment Relative to Instructional Practices*, is a contrast between the three Domain 2 standards' scores—a teacher's skill in managing classroom environment—and the five Domain 3 standards' scores—a teacher's skill in *instructional practices*. The third, which we call *Questions & Discussion Approach Relative to Standards & Content Focus*, is a contrast between the standard 3.4 score—a teacher's use of questions and classroom discussion—and the average of three other standard scores, 2.2, 3.1 and 3.2, which together measure a teacher's attention to routines, conveying standards-based instructional objectives to the students, and teaching in a manner that demonstrates content-specific pedagogical knowledge in teaching these objectives. (See Table 1 for detailed descriptions of each standard.) These first three principal components explain 87 percent of the variance of the eight standard scores, and the relative Eigen values suggest retaining at most three components.

To aid interpretation, instead of using the specific component loadings to form linear component scores, we have elected to use simplified scores designed to mirror the component scores. 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. Note that the second and third principal components are relative measures, so that a negative value does not indicate poor performance. Table 3 reports the means and standard deviations for each standard score, and these three constructed composite scores.[13]

V. Results and Discussion

A. Estimating the Relationship Between TES Scores and Achievement

We find that teachers' classroom practices, as measured by TES scores, do predict differences in student achievement growth. The estimated magnitudes of the rela-

13. The correlation between each of the three principal 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 ($\rho 1,2 = 0.110$, $\rho 1,3 = 0.049$, $\rho 2,3 = -0.107$).

Table 3
Teacher TES Score Descriptive Statistics

	Mean	Standard Deviation
Domain 2 "Creating an Environment for Learning"	3.38	(0.480)
Standard 2.1	3.52	(0.472)
Standard 2.1	3.27	(0.498)
Standard 2.1	3.40	(0.591)
Domain 3 "Teaching for Learning"	3.12	(0.431)
Standard 3.1	3.09	(0.467)
Standard 3.2	3.20	(0.464)
Standard 3.3	3.22	(0.455)
Standard 3.4	3.05	(0.495)
Standard 3.5	3.05	(0.495)
1. Overall classroom practices (average of all eight standards)	3.21	(0.433)
2. Classroom environment relative to instructional practices (average of Domain 2 minus average of Domain 3)	0.25	(0.270)
3. Questions & discussion approach relative to standards & content focus (Standard 3.4 minus average of 3.1, 3.2 and 2.2)	−0.14	(0.303)

tionships are, however, sensitive to the nonrandom assignment of students to teachers. We ultimately prefer a model that includes available controls for observable student and teacher characteristics, and school fixed effects; our later discussion of implications for policy and practice will focus on those estimates.

Before turning to that discussion we present our results, including evidence on student-to-teacher sorting that suggests a focus on our preferred model. Table 4, Columns 1–4 (Panel A for math achievement, Panel B for reading) present estimates from variations on Equation 2. Column 1 reports the relationship between our three TES scores and student achievement *ignoring* any potential sorting (that is, no student, teacher, or school level controls are included). The coefficients are large; a one-point increase in *Overall Classroom Practices* (measured by the average of all eight standard-level scores) predicts an increase of one-half a standard deviation in student achievement in both math and reading. This naïve estimate is, however, the implicit estimate used whenever simple correlations of observation scores and test scores are cited to demonstrate a relationship as is sometimes done in the sector.

Students and teachers are, of course, rarely if ever randomly assigned to each other. Cautious readers will, therefore, be wary of making inferences based on Column 1 of Table 1. In Columns 2, 3, and 4 we add student-, teacher-, and school-level controls respectively.

As reported in Column 2, adding controls for observable student characteristics—including, most notably, prior achievement scores—reduces the coefficient on *Overall Classroom Practices* by 60 percent for both math and reading. The predicted

Table 4
Estimates of the Relationship Between Student Test Scores and Teacher TES Score Principal Components

	(A) Math					
	(1)	(2)	(3)	(4)	(5)	(6)
1. Overall classroom practices	0.543**	0.221**	0.202**	0.105**	0.275**	0.362**
	(0.108)	(0.041)	(0.037)	(0.032)	(0.037)	(0.117)
2. Classroom environment relative to instructional practices	0.231^{+}	0.128*	0.121*	0.082*	−0.021	0.023
	(0.122)	(0.051)	(0.051)	(0.040)	(0.088)	(0.329)
3. Questions & discussion approach relative to standards & content focus	0.065	0.001	−0.009)	0.001	−0.031	0.051
	(0.140)	(0.060)	(0.061)	(0.039)	(0.097)	(0.219)
Student controls		Y	Y	Y	Y	Y
Teacher experience controls			Y	Y	Y	Y
School fixed effects				Y		
Teacher fixed effects						Y
Teacher sample	207	207	207	207	49	49
Student sample	16,196	16,196	16,196	16,196	4,109	4,109
Adjusted *R*-squared	0.049	0.527	0.529	0.556	0.545	0.561

	(B) Reading					
	(1)	(2)	(3)	(4)	(5)	(6)
1. Overall classroom practices	0.487**	0.209**	0.214**	0.141**	0.106^{+}	0.294*
	(0.094)	(0.036)	(0.037)	(0.029)	(0.055)	(0.134)
2. Classroom environment relative to instructional practices	0.160	0.059	0.059	0.036)	0.201^{+}	0.212
	(0.145)	(0.049)	(0.049)	(0.039)	(0.106)	(0.157)
3. Questions & discussion approach relative to standards & content focus	0.184	0.109*	0.100*	0.061	0.011	−0.297
	(0.135)	(0.046)	(0.046)	(0.038)	(0.072)	(0.236)
Student controls		Y	Y	Y	Y	Y
Teacher experience controls			Y	Y	Y	Y
School fixed effects				Y		
Teacher fixed effects						Y
Teacher sample	365	365	365	365	81	81
Student sample	20,125	20,125	20,125	20,125	5,251	5,251
Adjusted *R*-squared	0.039	0.528	0.529	0.550	0.550	0.582

Note: Each column represents a separate student-level specification. Clustered (teacher) standard errors in parentheses. ** $p<0.01$, * $p<0.05$, + $p<0.1$.

Table 5
Correlation Between Assigned Students' Incoming Achievement Level and Teacher TES Scores

	Correlation With Students' Baseline Test Scores			
	Math		Reading	
	District Wide	Within Schools	District Wide	Within Schools
	(1)	(2)	(3)	(4)
1. Overall classroom practices score				
Teacher TES scores observed in past school years	0.124	−0.008	0.139	0.021
Teacher TES scores observed in future school years	0.178	−0.017	0.096	0.022

gain in student achievement is now around one-fifth of a standard deviation. The coefficients on the second and third TES scores also shrink noticeably. We take this drop as evidence that students with high (low) achievement potential are systematically assigned to teachers with high (low) TES scores.

In fact, this sorting is evident in the simple correlation between students' prior test scores and their current year teachers' overall TES score reported in Table 5: 0.124 in math and 0.139 in reading. Note, however, that this positive correlation is also true when the teacher's TES scores are from TES observations in *future* school years—that is, the scores themselves could not have been known to the principal (or whoever is responsible for class assignments). This suggests, perhaps not surprisingly, that TES scores are correlated with other teacher characteristics used in student-teacher assignment, characteristics that threaten our estimates via omitted variable bias.

While teacher characteristics used in student assignment are largely unobserved (to researchers), teacher experience level is a significant exception. In Column 3 of Table 4 we add teacher experience controls; both the teachers' current experience (year t) and experience at the time her TES scores were measured (Year T). The added experience terms have little marginal effect on the magnitude of our estimated coefficients.

Finally, in Column 4 of Table 4 we add school fixed effects to control for the important student-teacher sorting that occurs between schools. We point out that the correlations reported in Table 5 are near zero within schools (Columns 2 and 4) suggesting that much of the student to teacher sorting that is related to TES scores occurs across rather than within schools. Another way to read Table 5 is that teaching

practices vary systematically across high and low achieving schools more than they vary across teachers serving high and low performing students within the same school.

This school fixed effects model is our preferred model. The estimates from this model (Column 4) indicate that improving a teacher's *Overall Classroom Practices* by one point is associated with one-seventh of a standard deviation increase in reading achievement, and one-tenth in math. To provide a sense of the magnitude of these estimates consider a student who begins the year at the 50th percentile in reading (math) and is assigned to a top-quartile teacher as measured by *Overall Classroom Practices* score; that student will, on average, score three percentile points higher in reading (two points in math) than a peer who began the year at the same achievement level but was instead assigned to a bottom-quartile teacher. Two or three percentile points is not trivial in comparison to the distribution of overall teacher effect; the difference between being assigned a top-quartile versus bottom-quartile teacher in Cincinnati is associated with a seven percentile gain in reading (six in math).[14]

Additionally, among students assigned to different teachers with similar *Overall Classroom Practices* scores, *math* achievement will grow more for students whose teacher is relatively better than his peers at classroom management; *reading* scores may grow more for students whose teacher is relatively better than her peers at engaging students in questioning and discussion. For math the coefficient on *Classroom Environment Relative to Instructional Practices* is about one-twelfth of a standard deviation and significant. For reading the coefficient on *Questions & Discussion Approach Relative to Standards & Content Focus* is positive and significant at the 0.107 level.

Sorting on *un*observed teacher characteristics that are correlated with both the potential for promoting greater achievement growth and with TES scores remains a concern. To address residual sorting on unobserved teacher differences, at least to the extent that they are time-invariant, we would prefer to estimate these relationships using teacher fixed effects. This is a common empirical strategy when attempting to understand the drivers of teacher differences (for example, see Rockoff 2004 regarding the influence of teacher experience). Columns 5 and 6 of Table 4 report estimates for the subsample of teachers (about one quarter of our full sample) for whom we can estimate a teacher fixed effects specification. In this subsample the variation in TES scores within teachers occurs because each teacher has been evaluated by TES in two (or rarely three) different school years in our data; we thus compare changes over time in an individual teacher's TES scores to changes in the achievement of her students over time.

For comparison across samples, Column 5 replicates the same specification as Column 3 using the teacher fixed effects subsample, and Column 6 adds teacher fixed effects to the specification in Column 5. While less-stable and noisier given

14. The standard deviation in overall teacher effect in Cincinnati is 0.12 student-level standard deviations in math and 0.13 in reading. We obtained these estimates by fitting our preferred model with random effects for teacher and class, but without the TES score or experience in year T covariates. They are similar in magnitude to effects estimated similarly by other researchers in different districts and states (see Hanushek and Rivkin 2010 for a summary).

the smaller sample, the *Overall Classroom Practices* score remains a statistically significant predictor of student achievement growth even within teachers. These results provide some evidence that it is the practices measured by TES themselves that promote student achievement growth and not just the teachers who employ these strategies.

To this point we have modeled the TES-student achievement relationship as linear. Holtzapple's (2003) original analysis of domain-level TES scores found evidence of a nonlinear relationship: students scored noticeably better with "Distinguished" or "Proficient" teachers compared to lower scoring teachers, but differences between the students who had "Distinguished" and "Proficient" teachers were much smaller. In Table 6 we explore the possibility of a nonlinear relationship. In Column 1 we report estimates identical to Table 4 Column 4 except that we replace the linear *Overall Classroom Practices* score with indicators for each quartile. In both math (Panel A) and reading (Panel B) the point estimates increase in a relatively linear fashion among the top three quartiles (though the differences are not statistically significant), but the bottom quartile (empirically a score below about three), especially for reading, is noticeably different.

We do find results similar to Holtzapple when using the same formal evaluator domain-level ratings. Column 2 replaces our three TES measures with a single indicator for teachers who were rated "Distinguished" or "Proficient" for Domains 2 and 3 in their evaluator's final judgment. In both math and reading, these teachers' students score about one-tenth of a standard deviation higher. Finally, in Column 3 we restrict our sample to just those teachers who received a formal domain level rating of "Distinguished" or "Proficient" and estimate our preferred specification. Our objective here is to test whether our results are robust to focusing on only those teachers whom Cincinnati's formal TES process had identified as performing well. In other words, is there additional benefit from differentiating among these high-scoring teachers. Indeed we find results very similar to our estimates using the entire sample (compare Table 6, Column 3 to Table 4, Column 4).

Last, before turning to additional interpretation of our main results, we explore the sensitivity of our estimates to our choice of sample: teachers and their students one or two years before or after the teacher participated in TES. As discussed in section IV, we prefer the before *or* after sample given the selection which inherently accompanies the structure of the TES program; this sample (used in Table 4, Columns 1–4) compares favorably to other Cincinnati teachers and students on observables (Table 2). In Table 7 we present separate estimates of the TES coefficients for three mutually exclusive subsamples that constitute our preferred sample: teachers we observe teaching a tested subject and grade only before participating in TES, those teaching before and after participation, and those teaching only after participation. The separate estimates were obtained by interacting each TES score with an indicator for the sample group but excluding main effects. These three samples are observably (Table 2) and likely unobservably different, and we might expect different results. The coefficients for *Overall Classroom Practices* are very similar across samples while the coefficients for the other two TES scores vary in magnitude and sign.

The largest differences are for *Classroom Environment Relative to Instructional Practices* in math achievement. The "only before" sample includes relatively more

Table 6
Estimates of the Potential Nonlinear Relationships Between Student Test Scores and Teacher TES Score Measures

	(A) Math			(B) Reading		
	Entire Estimation Sample		Formal Prof or Distg	Entire Estimation Sample		Formal Prof or Distg
	(1)	(2)	(3)	(1)	(2)	(3)
1. Overall classroom practices (linear)			0.100*			0.111*
			(0.047)			(0.044)
1. Relative to bottom quartile, overall classroom practices						
Second quartile, overall classroom practices	0.045			0.095**		
	(0.036)			(0.029)		
Third quartile, overall classroom practices	0.055			0.102**		
	(0.038)			(0.026)		
Top quartile, overall classroom practices	0.091**			0.130**		
	(0.033)			(0.032)		
2. Classroom environment relative to instructional practices	0.091*		0.096*	0.027		−0.012
	(0.042)		(0.043)	(0.040)		(0.046)
3. Questions & discussion approach relative to standards & content focus	0.011		0.014	0.059		0.054
	(0.038)		(0.042)	(0.038)		(0.041)
Received formal evaluator rating of proficient or distinguished		0.098*			0.102**	
		(0.041)			(0.030)	
Teacher sample	207	207	184	365	365	325
Student sample	16,196	16,196	15,062	20,125	20,125	18,500
Adjusted *R*-squared	0.555	0.555	0.560	0.550	0.549	0.552

Note: Each column represents a separate student-level specification including student controls, teacher experience controls, and school fixed effects. Clustered (teacher) standard errors in parentheses. ** $p<0.01$, * $p<0.05$, + $p<0.1$.

Table 7
Estimates of the Relationship Between Student Test Scores and Teacher TES Score Measures for Select Subsamples

	(A) Math (1)	(B) Reading (2)
1. Overall classroom practices		
"Only before TES" sample	0.151**	0.146**
	(0.035)	(0.032)
"Before & after TES" sample	0.128**	0.148**
	(0.032)	(0.029)
"Only after TES" sample	0.147**	0.136**
	(0.033)	(0.029)
2. Classroom environment relative to instructional practices		
"Only before TES" sample	−0.277*	−0.040
	(0.138)	(0.072)
"Before & after TES" sample	0.208**	0.086^{+}
	(0.064)	(0.049)
"Only after TES" sample	0.109	−0.036
	(0.078)	(0.071)
3. Questions & discussion approach relative to standards & content focus		
"Only before TES" sample	−0.164	0.095
	(0.105)	(0.065)
"Before & after TES" sample	0.005	0.053
	(0.076)	(0.061)
"Only after TES" sample	0.009	0.053
	(0.063)	(0.059)
Teacher sample	207	365
Student sample	16,196	20,125
Adjusted *R*-squared	0.557	0.551

Note: Each column represents a separate student-level specification including student controls, teacher experience controls, and school fixed effects. Reported coefficients are the interaction of each TES score and an indicator for the appropriate group; main effects omitted. Clustered (teacher) standard errors in parentheses. ** $p<0.01$, * $p<0.05$, + $p<0.1$.

teachers during their second or third years in the classroom (see Table 2); one potential explanation of the differences in the estimates is that improvements to instructional practices are more important during those early years, or that participating in TES influences a teachers' classroom management practices in particular. We caution against reading too much into any differences in Table 7 given the relatively small sample sizes of each of the separate groups that form the interactions, as well as the observed differences in these groups illustrated in Table 2. Rather we

take these results as additional support for focusing on the combined, likely more representative sample.

B. Implications for Policy and Practice

Which classroom practices measured by the TES process are the most effective at promoting student achievement? What responsible actions could a teacher or school administrator take given our estimates? In answering these two questions we focus on estimates from our preferred specification reported in Table 4, Column 4.

First, we find that a one-point difference in *Overall Classroom Practices* score is associated with one-seventh of a standard deviation increase in reading achievement, and one-tenth in math. These predicted differences suggest student achievement would benefit from efforts to improve teacher skill in all practices measured by TES. For example, a teacher might work to improve the quality of questions that are posed to students during instruction; moving from questions that are simply "relevant to the objectives of the lesson" (a "Basic" practice in standard 3.4) to "thought-provoking questions [that encourage] evaluat[ion], synthesis, and/or analysis" by students (a "Proficient" practice). As another example, a teacher might work to not simply reactively "monitor student behavior at all times" (a "Proficient" practice in standard 2.3), but "anticipate and prevent student misbehavior" (a "Distinguished" practice). Similarly concrete suggestions for improvement can be drawn from most of the practices described by the TES rubric.

In TES language, one point is the difference between, for example, "Distinguished" and "Proficient." However, a one point difference is also, empirically, a little more than two standard deviations (the standard deviation of *Overall Classroom Practices* is 0.44). That metric suggests that moving up one point overall would not necessarily be as easy as a first reading of the rubric might imply.

Second, among students assigned to different teachers with similar *Overall Classroom Practices* scores, we estimate that *math* achievement will grow more for students whose teacher is relatively better than his peers at classroom management. Consider two teachers both with an *Overall Classroom Practices* score of 3.2 (roughly the district average). Teacher A has a Domain 2 ("Creating an Environment for Learning") score of 3.3 and a Domain 3 ("Teaching for Learning") score of 3.1 resulting in an *Environment Relative to Instructional Practices* score of 0.2. Teacher B is the opposite, a Domain 2 score of 3.1, a Domain 3 score of 3.3, and a resulting *Environment Relative to Instructional Practices* score of –0.2. The coefficient on our second principal component for math achievement (0.082, Table 4, Panel A, Column 4) suggests that teacher A's students will score higher than teacher B's students by about 0.03 of a standard deviation (0.4*0.082).

Again, improved practice in all areas is preferable. If, however, teachers must choose a smaller number of practices to focus their improvement efforts on (for example, because of limited time or professional development opportunities), the results for *Environment Relative to Instructional Practices* suggest that *math* achievement would likely benefit by teachers improving their classroom management skills (as described in Domain 2) first before turning to their instructional skills (as describe in Domain 3). In other words, between the two earlier examples of improving class-

room questions and monitoring student behavior, the latter would be the preferable place to start.

Third, again among students assigned to different teachers with similar *Overall Classroom Practices* scores, we estimate that *reading* achievement may benefit more from time spent improving the practice of asking thought-provoking questions and engaging students in discussion, versus times spent on planning, standards, and instructional strategies. Though the coefficient in this case, on *Questions & Discussion Approach Relative to Standards & Content Focus* score, is only significant at the 0.107 level (Table 4, Panel B, Column 4) in the preferred school fixed effects specification.

The concern remains that we may have in part identified practices that effective teachers use rather than practices that themselves have a causal impact on student achievement growth. We suggest, however, that the results here are a responsible input to the decisions that principals and teachers must make regarding how to attempt improvement, particularly since an experiment that randomly assigns teaching *practices* to ascertain causality is currently unfeasible. We note, however, that to estimate the causal effects of the *likely policy action* taken as a result of our estimates, one could advocate an experiment that randomly assigns the content of teacher professional development since determining this content is something that administrators and teachers often do. For example, to test our suggestion that teachers focus first on improving classroom management skills, one could imagine an experiment with three randomly assigned samples of teachers: one receiving training on Domain 2, a second receiving training on Domain 3, and a control group that receives no additional training in either area. If it were established that the training did indeed change practice in these two areas, then comparisons of the student achievement gains across the three experimental groups would provide a plausible test of whether teachers should focus first on classroom managerial skills as our estimates suggest.

Before concluding we examine one additional question: Does the intensive TES process (for example, multiple observations, trained peer evaluators) produce teacher evaluations that have greater power for predicting student achievement gains than more subjective evaluations? For a comparison to the TES coefficients, consider methodologically similar estimates using teacher ratings from a survey of principals (Jacob and Lefgren 2008), and teacher ratings by mentor teachers (Rockoff and Speroni 2010). Jacob and Lefgren report that a one standard deviation increase in principals' ratings predict 0.137 student standard deviations higher achievement for math, and 0.058 for reading. Rockoff and Speroni report a coefficient on mentor rating of 0.054 for math achievement. In both cases the ratings and student achievement measures occurred in the same school year. Comparable estimates for a teacher's average TES score, *Overall Classroom Practices*, are 0.087 in math and 0.078 in reading.[15]

15. These coefficients come from estimates of the same specification as Table 5, Column 4, our preferred model, but (i) are estimated using student achievement and TES scores measured in the same year ($t = T$ in Equation 2), and (ii) are normalized by the standard deviation of average TES score (0.44).

The comparison, especially in math, may lead some to question the need for the more detailed TES process. However, the TES program has the advantage of providing teachers' and administrators details about specific practices that contributed to the score—details that the teachers can consider as they work to improve their evaluations. Additionally, scoring individual practices and then creating a composite score allows for study of the heterogeneity in teacher skill among generally "satisfactory" teachers. In our sample, nearly 90 percent received a formal rating equivalent to "satisfactory" (TES's "Distinguished" or "Proficient"), but our results are robust to focusing on this subsample (Table 6, Column 3) and the standard deviation in average TES score is 0.44. A school administrator desiring to differentiate the support she provides to teachers would be benefited from knowing, for example, that one received a score of 3.5 while the other scored an entire standard deviation lower at 3.1.

VI. Conclusion

The results presented here offer some of the strongest evidence to date on the relationship between teachers' observed classroom practices and student achievement gains. A student starting the year at the 50th percentile will gain three percentile points more in reading achievement in the classroom of a teacher with top-quartile practices (at least as measured by Cincinnati's Teacher Evaluation System) than she would have in the classroom of a bottom-quartile teacher. The difference in math would be two percentile points. Compare those TES-score-related differences to the estimated differences in achievement from *all* sources of teacher contribution: seven percentile points in reading achievement and six in math in Cincinnati. Additionally, among teachers with similar overall classroom skills, differences in teachers' relative skills in specific areas also predict variation in student achievement.

The nature of the relationship between practices and achievement, as estimated here, supports teacher evaluation and development systems that make use of multiple measures. Even if one is solely interested in student achievement outcomes, classroom practice-based measures provide critical information to teachers and administrators on what actions they can take to improve their effectiveness at raising achievement. The descriptions of practices, and different levels of each practice, that comprise the TES rubric can help teachers and administrators map out development plans. For example, a teacher wanting to improve his classroom management skills may find that in reviewing his TES scores he has scored relatively low in standard 2.3. Taking that signal, he might decide to focus on "anticipat[ing] and prevent[ing] student miss-behavior" rather than simply reacting when it arises. While specific, the rubric is not a step-by-step guide for change, and the empirical data suggest moving up could be more difficult than a simple reading of the rubric would suggest.

Evaluations based on detailed observations of classroom practice, like Cincinnati's, are valuable even if they do not predict student achievement dramatically better than more subjective ratings. First, scores for differentiated areas of practice should help focus improvement efforts on skills which lag overall performance. Second, asking evaluators to score individual practices can uncover important het-

erogeneity within the vast majority of teachers currently simply labeled "satisfactory." Many teachers' performance may indeed be "satisfactory" in the language of an employment contract, but that does not mean teachers so labeled are all providing the same service to their students.

Readers still may be cautious interpreting our results as causal given the possibility that we have not eliminated all biases due to nonrandom sorting of students to teachers, and unobserved teacher characteristics that are correlated with observed practices. However, strict causality may be difficult to establish in the case of classroom practices, and status quo approaches to identifying effective practices are far less systematic. Many school districts are currently at various stages of establishing observation systems with characteristics like Cincinnati's (among them Washington, D.C., Chicago, and Los Angeles) which will presumably produce data that allow others to conduct similar studies. Additionally, we recommend testing the effect on student achievement of the policy actions suggested here through experimental provision of varying professional development to teachers.

References

Aaronson, Daniel, Lisa Barrow, and William Sander. 2003. "Teachers and Student Achievement in the Chicago Public Schools." Federal Reserve Bank of Chicago Working Paper WP-2002–28.

Danielson, Charlotte. 1996. *Enhancing Professional Practice: A Framework for Teaching*. Alexandria, Va.: Association for Supervision and Curriculum Development.

Gordon, Robert, Thomas J. Kane, and Douglas O. Staiger. 2006. "Identifying Effective Teachers Using Performance on the Job." Hamilton Project Discussion Paper. Washington, D.C.: Brookings Institution.

Grossman, Pam, Susanna Loeb, Julia Cohen, Karen Hammerness, James Wyckoff, Donald Boyd, and Hamilton Lankford. 2010. "Measure for Measure: The Relationships Between Measures of Instructional Practice in Middle School English Language Arts and Teachers' Value-Added Scores." National Bureau of Economic Research Working Paper 16015.

Hanushek, Eric A. 1971. "Teacher Characteristics and Gains in Student Achievement: Estimation Using Micro Data." *American Economic Review* 61(2):280–88.

Hanushek, Eric A., and Steven G. Rivkin. 2010. "Using Value-Added Measures of Teacher Quality." *American Economic Review* 100(2):267–71.

Holtzapple, Elizabeth. 2003. "Criterion-Related Validity Evidence for a Standards-Based Teacher Evaluation System." *Journal of Personnel Evaluation in Education* 17(3):207–19.

Jacob, Brian A., and Lars J. Lefgren. 2008. "Principals as Agents: Subjective Performance Measurement in Education." *Journal of Labor Economics* 26(1):101–36.

Kane, Thomas J., and Douglas O. Staiger. 2008. "Estimating Teacher Impacts on Student Achievement: An Experimental Evaluation." National Bureau of Economic Research Working Paper 14601.

Kane, Thomas J., Jonah E. Rockoff, and Douglas O. Staiger. 2008. "What Does Certification Tell Us About Teacher Effectiveness? Evidence from New York City." *Economics of Education Review* 27(6):615–31.

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 0902.

The New Teacher Project. 2009. "Human Capital Reform in Cincinnati Public Schools: Strengthening Teacher Effectiveness and Support." New York City, N.Y.: The New Teacher Project.

McCaffrey, Daniel, J.R. Lockwood, Daniel Koretz, and Laura Hamilton. 2003. *Evaluating Value-Added Models for Teacher Accountability.* Santa Monica, Calif.: Rand Corporation.

Milanowski, Anthony. 2004a. "The Relationship Between Teacher Performance Evaluation Scores and Student Achievement: Evidence from Cincinnati." *Peabody Journal of Education* 79(4):33–53.

———. 2004b. "Relationships Among Dimension Scores of Standards-Based Teacher Evaluation Systems, and the Stability of Evaluation Score-Student Achievement Relationships Over Time." Consortium for Policy Research in Education, University of Wisconsin Working Paper Series TC-04–02.

Murnane, Richard. J., and Barbara R. Phillips. 1981. "What Do Effective Teachers of Inner-City Children Have in Common?" *Social Science Research* 10(1):83–100.

Rivkin, Steven G., Eric A. Hanushek, and John Kain. 2005. "Teachers, Schools and Academic Achievement." *Econometrica* 73(2):417–58.

Rockoff, Jonah E. 2004. "The Impact of Individual Teachers on Student Achievement: Evidence from Panel Data." *American Economic Review* 94(2):247–52.

Rockoff, Jonah E., and Cecilia Speroni. 2010. "Subjective and Objective Evaluations of Teacher Effectiveness." *American Economic Review* 100(2):261–66.

Rockoff, Jonah E., Douglass O. Staiger, Thomas J. Kane, and Eric S. Taylor. 2010. "Information and Employee Evaluation: Evidence from a Randomized Intervention in Public Schools." National Bureau of Economic Research Working Paper 16240.

Rothstein, Jesse. 2010. "Teacher Quality in Educational Production: Tracking, Decay, and Student Achievement." *Quarterly Journal of Economics* 25(1):175–214.

Weisberg, Daniel, Susan Sexton, Jennifer Mulhern, and David Keeling. 2009. *The Widget Effect: Our National Failure to Acknowledge and Act on Teacher Effectiveness.* New York, N.Y.: The New Teacher Project.