The Effect of Evaluation on Teacher Performance:

Evidence from Longitudinal Student Achievement Data of Mid-career Teachers

Eric S. Taylor*

John H. Tyler**

September 2010

Measuring and increasing teacher effectiveness have become dominant themes in American education reform movements. The magnitude and measured variation of teacher effectiveness—variation that is as large within as across schools—have produced a flurry of policy proposals to increase "teacher quality" or "teacher effectiveness." An important component of almost all efforts in this area is a push to find methods for identifying more and less effective teachers. Value-added modeling using student performance data and classroom observation evaluations are the two current methods at our disposal for determining teacher effectiveness. Each method has its advocates, and each method has demonstrable advantages and disadvantages. Advocates of classroom observation evaluation suggest that such practice-based evaluation can provide feedback and give professional development direction that can help teachers to become better. To date, however, the effect of evaluation itself on teacher effectiveness is an unanswered empirical question. We address this question asking: do experienced teachers who undergo evaluation in a well developed and rigorous practice-based evaluation system become better teachers as a result of going through the evaluation process? Using data from the Cincinnati Public School system we find evidence that going through Cincinnati's Teacher Evaluation System does increase a teacher's ability to promote student achievement growth as measured by test scores. This effect appears to be non-transitory and is stronger for the weakest teachers and for teachers whose evaluation scores improved the most across the four classroom observations that took place during the year.

* Stanford University
erictaylor@stanford.edu

** Brown University and NBER
John_Tyler@brown.edu

Authors listed alphabetically.

Draft – Please Do Not Cite or Distribute Without Contacting the Authors – Thank You

The Effect of Evaluation on Teacher Performance: Evidence from Longitudinal Student Achievement Data of Mid-career Teachers

Introduction

Measuring and increasing teacher effectiveness have become dominant themes in American education reform efforts. This emphasis is not surprising given the large variation in teachers' abilities to promote student achievement growth. Recent studies have consistently found large variation in teacher effects on student test scores; most estimates of the standard deviation in teacher effects range between 0.10 and 0.25 student-level standard deviations in math with somewhat smaller differences reported for English language arts (Aaronson, Borrow and Sander (2003), Gordon, Kane and Staiger (2006), Kane, Rockoff and Staiger (2006), Rivkin, Hanushek and Kain (2005), Rockoff (2004), Hanushek and Rivkin (2010)).[1]

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". Unlike past proposals which focused on inputs for educators (i.e., pre- and in-service training), these new proposals place great emphasis on measuring observed teacher performance in the classroom. Yet while advocates take motivation from test-score based measures (like those cited in the first paragraph), most policies include multiple measures of teacher effectiveness: test-score based measures, classroom observation measures, reviews of professional artifacts, and others. Multiple measures should provide a richer, less-noisy perspective on effectiveness, but are also preferable for at least two practical reasons. First, test-score based measures require annual, standardized tests, excluding as many as two-thirds of the teachers nationwide who do not teach

[1] Pioneering work in this area was done by Hanushek (1971) and Murnane and Phillips (1981).

subjects and grades that generate the type of testing data required.[2] Second, even when possible, using student test score gains as the sole method for evaluating teachers provides no information on how the least effective teachers (in terms of student test score gains) can become better.

Advocates of practice-based teacher evaluation (e.g., classroom observation, review of practice atrifacts) point to the second point in particular when arguing that practice-based evaluation can lead to better teachers. Not only can all teachers be evaluated using classroom observations, but this method of evaluation can provide developmental feedback on which classroom practices need the most attention if a teacher is going to improve. To date, however, the effect of evaluation itself on teacher effectiveness is an unanswered empirical question.[3] In this paper we address that question asking: do experienced teachers who undergo evaluation in a well-developed and rigorous practice-based evaluation system become better teachers as a result of going through the evaluation process?

Cincinnati's Teacher Evaluation System

The data for our study come from the Cincinnati Public School (CPS) system. In the 2000-2001 school year CPS launched their Teacher Evaluation System (TES), a practice-based evaluation system that gathers data from both classroom observations and from artifacts such as teacher lesson plans, evidence of professional development activities, and family contact logs.

[2] These test score requirements mean that, in most districts and states, only math and English teachers in grades four through eight can be evaluated using test-score based measures alson.

[3] In this paper we are interested in the effect of evaluation per se. Others have explored how using measures of teacher effectiveness for selective retention policies might impact average teacher effectiveness (Golhaber (2010), Staiger and Rockoff (2010)).

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. Both peer evaluators and administrators complete an intensive TES evaluator training course, and must accurately score videotaped teaching examples to check inter-rater reliability. Teachers are informed of the week during which the first observation will occur, with all other observations being unannounced.

Teachers are evaluated on dozens of skills and practices divided into four "domains" based on Charlotte Danielson's *Enhancing Professional Practice: A Framework for Teaching* (1996). The evaluation rubric contains quite specific language describing each teaching practice at four levels: "Distinguished", "Proficient", "Basic", and "Unsatisfactory". For example, the first practice listed in domain 2 is related to how a teacher goes about creating "an inclusive and caring environment" for students:

- Distinguished: "Teacher interactions with all students demonstrate a positive, caring rapport and mutual respect. Interactions are inclusive and appropriate."
- Proficient: "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." [4]

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

Peer evaluators and administrators are required to provide written feedback to the teacher within ten days of each classroom observation. In addition, the peer evaluator and administrator must each have a conference with the teacher after their first classroom observation. These conferences generally occur very close to when the teacher receives the written post-observation report. Owing to union-negotiated guidelines, the evaluator is instructed to not offer suggestions for improvement outside the official rubric language in the feedback and during the conference. Thus an evaluator may point out the stated characteristics of higher-level performance in a given area and reference details of the observation, but should not give an example from, for example, a different teachers' evaluation.

At the end of the year a final summative score in each of four domains of practice is calculated and presented to the evaluated teacher.[5] For beginning teachers (those evaluated in their first and their fourth years), the consequences of a poor evaluation could be non-renewal of their contract. For tenured teachers, the consequences of the evaluation include determining eligibility for lead teacher status or additional tenure protection, or if the evaluation was poor, placement in the peer assistance program.

Teachers undergo a comprehensive TES evaluation only at defined intervals: the first year as a new hire, the fourth year, then every five years after that point. However, teachers hired before the TES program began in 2000-01 were "phased in" to the program. That is, the first year these veteran teachers received a TES evaluation was in the middle of their career and determined by a pre-agreed schedule; we return to this schedule and these teachers later.

Data provided by the Cincinnati Public Schools identify the year(s) in which a teacher was evaluated by TES, the dates when each observation occurred, and the scores. We combine these

[5] For more details on the scoring process see Kane et al. (forthcoming).

TES data with additional administrative CPS data that allow us to match teachers to students and student test scores. We use these combined data to explore the evaluation-performance relationship for CPS teachers.

Practical and Theoretical Considerations of the Evaluation-Performance Relationship

For at least two decades the effect of evaluation on teacher performance has been an asked, but unanswered question. A 1987 review of the teacher evaluation literature noted that there was "no definitive answer" to the question of whether evaluation programs were useful for improving teaching (Weber p. 56). Johnson (1990) soon followed reporting that evaluation systems at that time had low utility as a means for improving performance. Almost two decades later Donaldson (2009) concluded that "teacher evaluation has generally failed to influence teacher quality and student learning" (p. 9); the claim of "failure", however, was not based on empirical analysis

While existing evidence on the evaluation of teachers is limited, the general question underlying our present analysis—Does evaluation improve (degrade) performance?—has motivated theoretical and empirical work in many disciplines and sectors. That existing work can heuristically be divided into two categories, though the two do overlap. First, the effects of the evaluation process *per se* on performance. In this first category the (hypothesized) performance response is caused by the process, so these effects are more likely short-run or non-persistent; ending when the process ceases. Second, the effects of signals communicated through evaluation regarding how to improve performance. To the extent an individual acts on these signals these effects should be more lasting.

Effects of the Evaluation Process

The simplest model of performance response to evaluation, drawn from principal-agent theory, posits that absent evaluation or monitoring some subset of individuals (e.g., teachers) would prefer a level of personal performance lower than what would be optimal for the group (e.g., school or community more generally). These "knavish" individuals who are solely self-interested and opportunistic are contrasted with "knightish" individuals who are altruistic and motivated by a desire to promote the welfare of their clients (Levacic 2009 and Le Grand 2000). Knavish individuals should respond rationally to evaluation by increasing effort and performance while they are under evaluation, but return to lower effort levels later. This choice is rational since the supervisor or group will presumably use the evaluation results to infer performance since the last and until the next evaluation. In contrast, knightish individuals' level of effort and performance should be essentially unaffected by evaluation.[6] Thus the average effect of an evaluation program would depend on the distribution of knavish and knightish individuals, but any impact would likely be fleeting.

Since the education enterprise is characterized by ambiguous agent (teacher) motives, moral hazard, information asymmetry, multiple principals, and multiple outputs, writing optimal and enforceable contracts is very difficult (Dixit 2002). The absence of enforceable contracts suggests knavish teachers have room to modulate their effort and performance. Assuming the evaluation program values and can accurately assess performance which leads to increased student achievement—and there is evidence this is true at least in Cincinnati (Kane et al.

[6] Individuals who do not consciously adjust their performance in response to evaluation may nevertheless empirically increase their performance. Such behavior is often called a "Hawthorne" effect.

forthcoming)—we would expect to see achievement increase in the year a teacher was being evaluated, but return to pre-evaluation levels after.

By contrast, work from psychology suggests that short-run changes in performance associated with evaluation may result because evaluation affects a critical moderating factor: individuals' motivations. One line of research suggests that "task focused" evaluation which scores individuals solely on the basis of successful task completion tends to increase intrinsic motivation to perform, independent of any feedback. Meanwhile, evaluation that compares individuals' performance to some norm or to the performance of other individuals tends to decrease intrinsic interest in the tasks at hand (Harackiewicz, Abrahams, and Wageman 1987). A different, but related, line of research suggests that performance on tasks requiring creativity may be adversely impacted when evaluation is seen by the individual being evaluated as a form of extrinsic motivation to perform well on the tasks (Amabile 1979).

It is not immediately clear how these motivation-related models should apply to performance under evaluation in education. Robust teacher evaluation systems like Cincinnati's do measure many discrete tasks and often take pains to avoid comparison of individuals, but the same systems also describe task performance at different levels in normative terms (e.g., Distinguished, Advanced, Proficient). Additionally, while the "creativity" required for any profession is debatable, teachers do often express discomfort with a disconnect they perceive between what they feel successful teaching is and how extant evaluation systems define successful teaching.

Effects of the Signals Sent by Evaluation

An evaluation often sends signals, both intentionally and unintentionally, to individuals regarding how they should change their behavior to improve performance. Indeed, the intentional signals provided in formal feedback and coaching are a frequently cited mechanism through which evaluation can improve teacher performance. However, the signals can be weak, ignored by the individual, or unintentionally wrong.

Research on teaching specifically suggests that the benefits of formal feedback will be largest when (1) post-evaluation feedback is specific and prescriptive, (2) when it is delivered in a timely manner by an evaluator deemed to be credible by the teacher, and when the prescriptive elements of the feedback can be integrated by the teacher into their practice and/or when the teacher has access to professional development suggested by the feedback (Milanowski and Hememan 2001; Kimball 2002; Milanowski 2004). As noted earlier, there is little empirical student achievement evidence on the effect of teacher evaluation generally, let alone these dimensions of feedback individually.

Research on professional evaluation more generally suggests that feedback benefits may be greater when the evaluation is multi-sourced. Often called “360 degree” feedback, in multi-source evaluation individuals receive evaluations from subordinates, peers, and individuals higher on the organization chart. Advocates point to three advantages: (1) employees gain a more comprehensive perspective of their work performance when feedback is provided by people with a different perspectives on it, (2) feedback is less likely to be ignored if the sources include peers and superiors because the raters have more status and power than subordinates, and (3) one may be more motivated to change one’s behavior when ratings are lower than self-ratings (Seifert, Yukl, and McDonald (2003)). The empirical evidence on the effects of multisource feedback is somewhat mixed. A 1996 meta-analysis finds only small improvements in performance associated with multisource feedback and performance actually decreased a third of the time (Kluger and DNisi (1996)). One the other hand, two more recent studies, one experimental, suggest that multisource feedback, when combined with coaching or with a feedback workshop, leads to increased performance levels (Luthans and Peterson (2003) and Smither et al. (2003)).

Existing teacher evaluation programs and proposals differ somewhat on these dimensions of formal feedback. In recent years, many teacher evaluation systems have used poor instruments that lack meaningful measurements and, for a host of reasons, are actually constructed to minimize the differentiation of teachers (Donaldson 2009). In the Cincinnati case, by contrast, the detailed TES rubric and observation process provide substantive inputs for the kind of detailed, prescriptive feedback and planning described by education researchers. However, there are bounds to how prescriptive TES evaluators themselves are allowed to be in their formal feedback. Additionally, TES provides two sources of feedback—one peer and one superior—more than many teacher evaluation systems but not quite the multi-source ideal posited in the 360-degree feedback model. These positive, if not up to the ideal, characteristics of TES suggest we might be more likely to find a positive effect of being evaluated by the TES program.

Still, feedback benefits performance only if acted upon. In addition to not receiving a signal weakened by lack of the characteristics discussed above, evaluated individuals may consciously ignore or dismiss feedback. Attribution theory in psychology posits that actors tend to attribute the causes of their behavior (e.g., the behavior being evaluated) to stimuli inherent in the situation, while observers tend to attribute behavior to stable dispositions of the actors (Heider 1958). To the extent that such actor-observer bias is operating in teacher evaluation, teachers undergoing evaluation might discount the results and recommendations, and mute the effect of evaluation on performance

Even assuming quality feedback and take up of the feedback, any impact on student achievement will only occur after the suggested skills are developed or put into practice by the evaluated teacher. These intermediate steps could delay any observed effect. If a recommended change in practice is easy we might expect to see effects immediately—during the year of evaluation in the TES case—and lasting into the future. However, a recommended change may require formal training and on-the-job practice to implement; in such cases we might not expect to observe the total performance improvement effect for some time.

Finally, feedback need not be formal or individualized to effect performance. Milanowski (2001) has suggested that the descriptive language of a classroom observation rubric can provide specific guidance to teachers—even those who have not been through a formal evaluation—on (hypothesized) best practices. Assuming teachers act on this signal prior to evaluation, we would see improvements in average effectiveness in the period before participation in the evaluation program. The signals may, however, be unintended or counterproductive. Studies by Harkins (2006) suggest that the potential for evaluation leads participants to put greater effort into the prepotent response [a response with higher priority than other responses] and that "...this mere effort alone can account for the typical finding that evaluation improves performance on simple items and debilitates performance on complex ones" (p. 436). Again, whether the effect of such unintentional signals is positive or negative turns on the nature of the tasks under evaluation.

<u>Empirical Strategy</u>

Our objective is to estimate the extent to which a teacher's participation in TES improves (diminishes) her effectiveness in promoting student achievement growth. Using annual district administrative data on students, teachers, and classes in the 2003-04 through 2009-10 school years we estimate the following specification:

$$(1)\ A_{ijt} = \alpha + \delta_1 currentTES_{jt} + \delta_2 pastTES_{jt} + X_{ijt}\beta + exper_{jt}\gamma + \tau_j + \varepsilon_{ijt}$$

where A_{ijt} is the math achievement of student i taught by teacher j in school year t, as measured by the end-of-year state test.[7] The variable $currentTES_{jt}$ is equal to one if teacher j participated in TES during school year t and zero otherwise (i.e., T=t where T represents the year teacher j participated in TES). Similarly $pastTES_{jt}$ is equal to one if teacher j participated in some past school year (i.e., t>T). The coefficients of interest capture differences in the achievement of students taught during, δ_1, or after, δ_2, teacher participation in TES compared to students taught before participation (the implicit left-out category, i.e., t<T).

[7] Between 2002-03 and 2009-10 Cincinnati students, in general, took end of year exams in reading and math in third through eighth grades. 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. 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 (e.g., end of year test in school year t) and a baseline test (e.g., 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.

Our specification also includes a teacher fixed effect, represented in equation 1 by τ_j. We prefer the resulting within-teacher estimates of $\delta_1$and δ_2 for two primary reasons. First, existing evidence suggests that inexperienced and experienced teachers vary greatly in their ability to promote student achievement (see reviews in Hanushek and Rivkin 2010, Gordon, Kane and Staiger 2006). To the extent high (low) ability teachers are more likely to participate in TES (e.g., through differential attrition from the district, or volunteering for the program) simple cross-sectional estimates would be biased. Second, the teacher fixed effect will account for time-invariant, non-random differences in the assignment of students to specific teachers. Some teachers may year after year be asked to teach classes with high (low) potential for achievement gains (e.g., through principal favoritism, or school assignment).[8]

However, not all the dynamics of student-teacher assignment need be time-invariant. To account for variation in students assigned to a given teacher from year to year, we include a vector of observable student characteristics, X_{ijt}. Most notably, X_{ijt} captures the student's prior achievement including the main effect of the prior year math test score, the score interacted with each grade-level, and fixed effects for each test (i.e., grade-by-year fixed effects). When the baseline score was missing for a student, we imputed with the grade-by-year mean, and included an indicator for missing baseline score.[9] Additionally, X_{ijt} includes separate indicators for student gender, racial/ethnic subgroup, special education classification, gifted classification, English proficiency classification, and whether the student was retained in grade.

[8] For theoretical and empirical discussions of the potential bias see Rothstein (2010) and Kodel and Betts (2009).

[9] Our estimates are robust to excluding students with missing baseline test scores.

And while teacher effectiveness varies across careers, there is evidence of returns to experience on average especially early in the career (Gordon, Kane and Staiger 2006). Accordingly we include controls for the teacher's years of experience and years of experience squared, represented by $exper_{jt}$. Our estimates are robust to, alternatively, specifying $exper_{jt}$ with a series of indicator variables for categorized experience-levels.

We estimate equation 1 using the sample of teachers (and their students) who were hired by Cincinnati Public Schools between 1993-94 and 1999-2000—before the implementation of the TES program in 2000-01—but who were eventually required to participate in TES according to a schedule which "phased-in" veteran teachers. The phase-in schedule, determined during the TES program's planning stages and detailed in table 1, delayed the participation of teachers already working in the district as of the 1999-2000 school year. The delay allows us to observe student achievement for this sample of teachers' classes before they participated in TES; as implied above, these *before* years serve as our counterfactual.

Table 1: Timing of First Scheduled TES Participation for Veteran Teachers

Teacher Contract Year	Scheduled First Participation Year	Experience at Time of Scheduled Participation*
1999-2000	2006-07	8 years
1998-99	2007-08	10 years
1997-98	2005-06	9 years
1996-97	2006-07	11 years
1995-96	2007-08	13 years
1994-95	2008-09	15 years
1993-94	2009-10	16 years

*Note: "Experience" is the expected value. Teachers who take a leave of absence, or began employment at CPS with prior experience would have different levels of experience.

Table 2 reports descriptive characteristics for the teachers and students included in our estimation sample (column 2), and for those excluded from our sample (column 1). The excluded group is composed of teachers (and their assigned students): (a) hired in 1992-93 or earlier, (b) hired in 2000-01 or later, and (c) hired between 1993-97 and 1999-2000 who did not remain teaching in the district. The first excluded group, those hired in 1992-93 or earlier, will be phased-into the program during future school years.[10] The second excluded group, teachers hired since 2000-01, are required to participate in TES during their first year working in the district; this requirement holds for both true novices and veterans moving to the district from elsewhere, and prohibits a before-TES counterfactual observation.[11]

Our analysis sample is, as expected given its selection using the phase-in schedule, more likely to be mid-career: as reported in table 2 column 2, 83.4 percent of our observations are teachers in their fifth through 19th year of teaching, compared to just 47.9 percent of the other teachers in the district who are not in our estimation sample. Students taught by our analysis teachers have similar baseline test scores in both math (a mean of 0.072 versus 0.054, a test of the difference yields a p-value of 0.06), and reading (a mean of 0.066 versus 0.072, a test of the difference yields a p-value of 0.53). Additionally, analysis sample teachers were slightly more likely to be teaching earlier grades, but had similar students in terms of demographic and program participation characteristics.

[10] Some teachers hired in 1992-93 or earlier have participated in TES on a voluntary basis.

[11] New hires participate in TES for a second time typically in their fourth year in the district (fifth year if they were veteran new hires). Later we estimate whether a second “dose” of TES participation has an effect on teacher performance.

The pattern of scheduled participation years for our sample creates relatively exogenous variation in the timing of teachers' TES participation. Thus, for example, a teacher hired in 1998-99 would participate in 2007-08 which for the average teacher would be their tenth year teaching; while a teacher hired the prior year in 1997-98 would participate in 2005-06 their ninth year (see table 1). This variation allows us to identify the returns to experience and overall temporal trends, if any, separately from the year of TES participation. Had the TES program designers decided to build the phase-in schedule based on experience level this separation would likely not have been possible. While we can and do control for teacher experience, the existing literature suggests the marginal returns to experience are relatively small beyond year five (see for example Rockoff 2004), and as reported later our estimates are not substantively changed by the exclusion of experience controls.

Most but not all teachers hired between 1993-94 and 1999-2000 participated in TES during the scheduled phase-in year reported in table 1. About 25 percent of teachers in our sample volunteered to participate in TES before their scheduled phase-in year. Many teachers who requested to participate early did so to fulfill the requirements for obtaining "lead teacher status" (eligibility for certain valued positions with greater compensation). We interpret volunteering as a signal of some latent characteristic likely positively correlated with teacher effectiveness. Accordingly we estimate the coefficients on $currentTES_{jt}$ and $pastTES_{jt}$ separately for volunteers and non-volunteers using interactions, and also test whether volunteers were observably better before they elected to participate in TES.

Table 2 also provides descriptive statistics for our analysis sample (column 2) separated into scheduled participants (column 3) and volunteer participants (column 4). On average, the students of volunteer participants begin the school year with higher achievement—a difference of 0.041 standard deviations in math (p=0.02) and 0.069 in reading (p<0.01). This difference may be in partly influenced by slightly fewer special education and English language learner students. Volunteers also are noticeably more likely to be teaching earlier grade levels. Volunteers and scheduled participants, however, have similar experience profiles.

Additionally some teachers hired between 1993-94 and 1999-2000 stopped teaching in the district before their scheduled TES participation year. It is possible the decision to leave was influenced by the prospect of TES participation. Perhaps, for example, a teacher self-aware of his own limited effectiveness may have chosen to leave the district rather than face formal evaluation. If such related attrition occurred, our within teachers strategy will still produce internally valid estimates of the effect on the "treated" teachers, but the attrition would suggest potential general equilibrium effects of the TES program as a whole. Later we present results with these leavers included in the estimation sample as a check of robustness.

Results and Discussion

We find that, on average, participation in the Teacher Evaluation System (TES) improves mid-career teachers' effectiveness in promoting student achievement growth in math. Table 3 reports the coefficients of interest estimated using variations on equation 1, and the sample described in the previous section. Our estimates are relatively robust to the choice of additional covariates, but differ importantly when contrasting teachers who did or did not volunteer to participate in TES.

Table 3 column 1 reports the uncontrolled differences in mean math achievement levels for students assigned to teachers during and after TES participation relative to students assigned to teachers before participation. The differences are small and not significant. However, the essentially descriptive statistics in column 1 ignore any non-random assignment of students across teachers or across years within teachers, and thus inferences based on column 1 risk under (or over) stating the influence of TES participation.

When we estimate differences within teachers (column 2 which adds teacher fixed effects), students assigned to a teacher during the year the teacher participates in TES score 0.072 standard deviations higher in math, on average, than students assigned to the same teacher in years before she participated in TES. And students assigned to a teacher's in years after the teacher participates in TES score 0.111 standard deviations higher in math on average. In other words, conditional on the teacher they are assigned, we would expect students' test scores to be higher if their teacher has participated in TES.

The differences reported in table 3 could, however, be artifacts of changes over time in the type of students assigned to a teacher, or in the teacher's experience level. However, when we add controls for observable student characteristics (column 3) and then controls for teacher experience (column 4), the estimates remain similar. With student and teacher experience controls, math achievement is 0.062 standard deviations higher in the year of TES participation, and 0.113 standard deviations higher in subsequent years.

The stability of estimates across columns 2 and 3 may surprise readers who are aware of the typical variation in average incoming student achievement across teachers and classes. Indeed for the estimation sample, cross-teacher differences account for about one-quarter of the variation in baseline math test scores. Despite the variation *across* teachers, we observe little variation *within* teachers over time. Table 4 reports a series of simple regressions with baseline student characteristics as outcomes and indicators for our within teacher periods of interest,

$currentTES_{jt}$ and $pastTES_{jt}$ in equation 1 terms, as predictors. Most coefficients are small and not statistically significantly different from zero.

While student assignment patterns may not be correlated with TES participation timing, the accumulated experience of any individual teacher will be correlated—each teacher's experience level after TES participation must necessarily be greater than his experience before participation. However, most existing evidence suggests that the marginal returns to experience are small after the first five years teaching (Rockoff 2004, Gordon, Kane and Staiger 2006), and essentially none of the teachers in our sample have fewer than five years experience by 2003-04 when our observations begin.

In column 5 of table 3 we add an indicator for the school year immediately prior to the year the teacher participated in TES (i.e., $t=(T\text{-}1)$); accordingly the reported coefficients are now differences relative to students taught two or more years prior to TES participation (i.e., $t<(T\text{-}1)$). Separating out the year prior to participation allows us to test whether teachers who were about to participate in TES were already on an upward trajectory. As reported in column 5, the coefficient on prior year is positive (0.033 standard deviations) but not statistically significantly. Additionally, the coefficients of interest increase somewhat. These results suggest that teachers—or at least some teachers as discussed later—were on an upward trajectory not captured by the returns to experience, but we cannot rule out that the slight trend we see is the result of chance.

The estimates in table 3 columns 1 through 5 do not make a distinction between teachers who volunteered to participate in TES and those who participated when their scheduled phase-in year came up. Volunteering may signal some latent characteristic positively correlated with teacher effectiveness, and possibly also correlated with growth from TES participation. In column 6 we estimate our coefficients of interest separately for scheduled participants and volunteer participants.[12] The coefficients for scheduled participants—non-volunteers—remain similar to the combined estimates (column 5) though slightly smaller.

By contrast, the point estimates in column 6 for volunteer participants are larger.[13] The results for volunteers suggest two things. First, volunteering is likely a signal of latent ability. Students taught by TES volunteers score one-tenth to one-twelfth of a standard deviation better in math beginning in at least the year of TES participation. Second, even though they are unlikely to be representative, the estimated coefficients in column 6 suggest volunteers may also be improving as a result of TES participation. The coefficients increase from 0.053 to 0.161 and then to 0.226; a positive trajectory though not as steep as for non-volunteers.

The estimates in table 3 column 6, our preferred estimates, suggest that a student taught by a mid-career past-TES-participant will score about one-eighth of a standard deviation higher in math than a similar student taught by the same teacher before the teacher participated in TES. If those two students began their respective years with the teacher at the 50th percentile of math achievement, the first student would score about five percentile points higher at the end of the

[12] Econometrically, we interact the variables of interest both with an indicator for volunteer, *vol*, and also with an indicator for scheduled participant, *sch*. In equation 1 notation:

$$A_{ijt} = \alpha + \delta_{11} currentTES_{jt} * sch_{jt} + \delta_{21} pastTES_{jt} * sch_{jt} + \delta_{12} currentTES_{jt} * vol_{jt}$$
$$+ \delta_{22} pastTES_{jt} * vol_{jt} + X_{ijt}\beta + exper_{jt}\gamma + \tau_j + \varepsilon_{ijt}$$

The inclusion of teacher fixed effects precludes estimating a main effect of *vol*.

[13] In results available upon request, we show that the estimates in table 3 column 6 are very similar if we estimate equation 1 for the volunteer and scheduled samples separately, though the standard errors are somewhat larger.

year. Additionally, a student taught in the year the teacher participates in TES will score one-thirteenth of a standard deviation higher than students taught before participation.

The standard deviation in total teacher effect on student math achievement for our analysis sample is about 0.22 student-level standard deviations.[14] Assume, momentarily, that the general level and distribution of teacher effectiveness in the district was not changing. Under that assumption, participating in TES would move the average non-volunteer teacher up nearly three-fifths of a standard deviation in the teacher distribution (i.e., 0.127 divided by 0.22).[15] In other words, taking that teacher from average (50th percentile) to nearly top-quartile (72nd percentile). However, all the teachers in our sample are required to participate in TES. Thus participation may improve individual and overall district performance, but not dramatically change the relative distribution of teachers within the district. Certainly a positive outcome, but a difficult one to measure given that our data are limited to just the district.

[14] We obtained this estimate by fitting a specification similar to equation 1 except that we omit the teacher-level covariates and fixed effects, and add random effects at the teacher and class (nested within teacher) levels. The magnitude is on the high-end of estimates obtained by other researchers using a similar empirical strategy in other settings (see Hanushek and Rivkin 2010 for a summary of other estimates).

[15] This improvement in teacher effectives may also be picked up by the TES classroom observation scores. In those data (see Kane, Taylor, Tyler and Wooten (forthcoming) for a description) just over 500 teachers have participated in TES two different school years. The average change in overall TES score from first to second participation is also around three-fifths of a standard deviation (0.26 TES points divided the standard deviation in overall TES score of 0.44).

That student achievement is higher even in years following TES participation is consistent with the hypothesis that teachers act on the signals they receive from evaluation, and, at least in the case of math teaching in Cincinnati, those signals lead to improved performance. Put differently, we do not find the kind of transitory boost in performance during the year of TES participation that would support models where teachers only adjust their behavior when actively under evaluation.

Still, much of the causal story remains unclear. First, even if teachers' actions changed post-evaluation we do not know what their expectations were when the evaluation process began. For example, some teachers may have begun the TES participation year planning to adhere to the TES rubric only during evaluation, or not expecting to learn much from the process; but found the feedback helpful and ultimately adjusted their behavior long run. Second, we cannot say what teachers changed about their behavior, nor which changes were most important to student achievement growth. Following the TES rubric's explicit suggestions for best practices is only one possible mechanism. Alternatively, the general peer- and self-scrutiny may have uncovered opportunities for improvement in areas not addressed by the TES rubric.

As shown in table 5, the effects of TES participation are not uniform; the change from before to after participation is larger for both teachers who received relatively low TES scores (panel A) and teachers whose TES scores grew the most during TES participation (panel B). In table 5 panel A, we interacted our indicator for past TES participation, $pastTES_{jt}$, with indicators for quartile of overall TES score received during the year of participation, T. This overall TES score is the average of more than two dozen teaching practices scored in four separate classroom observations.[16] The difference between average math student achievement after TES participation versus before participation was largest for teachers with bottom-quartile TES scores: 0.279 student-level standard deviations. Similarly in table 5 panel B, the estimated difference is largest for teachers in the bottom two quartiles of TES score growth: 0.183 for the quartile of largest growth and 0.229 for the quartile of second largest growth, though we cannot reject that these coefficients are equal. This TES score growth is the change in overall TES score from the first to the last classroom observation during the TES year.

Our focus in this paper has been on math achievement. In similar analyses of reading achievement we do not find significant differences in student achievement associated with TES participation.[17] For reading, the coefficients of interest are close to zero and not statistically significant. Several studies now have found less variation in teachers' effects on reading achievement compared to the variation in teachers' effects on math achievement (see Hanushek and Rivkin 2010 for a summary). These smaller reading teacher differences could arise because students learn reading in many settings at school and at home outside a formal reading class. The smaller differences could also arise because reading instruction practices are more homogeneous than math instructional practices. Scripted reading curriculum is one mechanism that would homogenize reading instruction. Reading instruction may also be more homogenous because teachers are better at finding, sharing, and adopting best practices. In either case classroom

[16] These two dozen teaching practices are collectively known as TES Domains 2 and 3 (see Kane et al. (forthcoming) for more information about the process, rubric, and scores).

[17] Reading results for all tables included in this paper are available from the authors upon request.

observation may turn up little feedback for improvement—especially among mid-career teachers.

Conclusion

The estimates presented here support the hypothesis that teacher evaluation programs can improve teacher performance, as measured by student test score gains, both during the year of the evaluation process and in the years following evaluation. Our estimates suggest that a student taught by a teacher after that teacher participates in the TES evaluation program will score about one-eighth of a standard deviation higher in math than a similar student taught by the same teacher before the teacher participated in TES. If those two students began their respective years with the teacher at the 50th percentile of math achievement, the first student would score about five percentile points higher at the end of the year. Additionally, students taught during the school year a teacher is participating in the evaluation process will also score higher in math.

Advocates of teacher evaluation should, however, take caution when extrapolating these results to other teacher evaluation programs and proposals. First, Cincinnati's investment in the Teacher Evaluation System is substantial: a detailed rubric describing practices shown to correlate positively with student achievement, multiple observations and feedback opportunities over the course of an entire school year, regular evaluator training. Second, we do not find effects of TES evaluation on students' reading achievement. Third, the teachers in our analysis sample were all beyond their fifth year of teaching when they participated in the evaluation. The effects may be larger (smaller) for teachers earlier in (or very late in) their career.

Our results suggest optimism that well-structured teacher evaluation programs can improve the average effectiveness of mid-career teachers at least in mathematics. Thus system-wide gains from evaluation need not only come through selective termination of teachers who score low. However, the dimensions of “well-structured” remain elusive; a critical gap in a time when many new evaluation systems are under development. The entire sector would be well served by K-12 systems willing to experimentally vary the components of their evaluation system (including the timing of teacher participation), and measure any resulting differences in teacher effectiveness.

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.

Amabile, T. M. (1979). "Effects of external evaluation on artistic creativity." *Journal of Personality and Social Psychology* 37(2): 221-233.

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

Dixit, A. (2002). "Incentives and Organizations in the Public Sector: An Interpretative Review." *The Journal of Human Resources* 37(4): 696-727.

Donaldson, M. L. (2009). “So long, Lake Wobegon? Using teacher evaluation to raise teacher quality.” Center for American Progress.

Goldhaber, D. and M. Hansen (2010). “Assessing the Potential of Using Value-Added Estimates of Teacher Job Performance for Making Tenure Decisions,” National Center for Analysis of Longitudinal Data in Education Research, Working Paper 31.

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

Hanushek, E. (1971). "Teacher Characteristics and Gains in Student Achievement: Estimation Using Micro Data." *The American Economic Review* 61(2): 280-288.

Hanushek, Eric A., and Steven G. Rivkin. 2010. “Using Value-Added Measures of Teacher Quality.” *American Economic Review* 100(2):267-271.

Harackiewicz, Judith M., Steven Abrahams, and Ruth Wageman. (1987). "Performance Evaluation and Intrinsic Motivation: The Effects of Evaluative Focus, Rewards, and Achievement Orientation." *Journal of Personality and Social Psychology* 53(6): 1015-1023.

Harkins, S. G. (2006). "Mere Effort as the Mediator of the Evaluation-Performance Relationship." *Journal of Personality and Social Psychology* 91(3): 436-455.

Heider, F. (1958). *The Psychology of Interpersonal Relations*. New York, Wiley.

Johnson, S. M. (1990). *Teachers at Work: Achieving Success in Our Schools*. New York, Basic Books.

Kane, Thomas J., Eric S. Taylor, John H. Tyler, and Amy L. Wooten (forthcoming). "Identifying Effective Classroom Practices Using Student Achievement Data." *Journal of Human Resources*.

Kimball, S. M. (2002). "Analysis of Feedback, Enabling Conditions and Fairness Perceptions of Teachers in Three School Districts with New Standards-Based Evaluation Systems." *Journal of Personnel Evaluation in Education* 16(4): 241-268.

Koedel, C. and J. 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.

Kluger, A. N. and A. DeNisi (1996). "The Effects of Feedback Interventions on Performance: A Historical Review, a Meta-Analysis, and a Preliminary Feedback Intervention Theory." *Psychological Bulletin* 119(2): 254-284.

Le Grand, J. (2000). “From knight to knave? Public policy and market incentives.” *Risk, trust and welfare*. P. Taylor-Gooby (ed.). Basingstoke, UK, Macmillan**:** 21-30.

Levačić, R. (2009). "Teacher Incentives and Performance: An Application of Principal–Agent Theory." *Oxford Development Studies* 37(1): 33 - 46.

Luthans, F. and S. J. Peterson (2003). "360-degree feedback with systematic coaching: Empirical analysis suggests a winning combination." *Human Resource Management* 42(3): 243-256.

Milanowski, A. T. and H. G. Heneman (2001). "Assessment of Teacher Reactions to a Standards-Based Teacher Evaluation System: A Pilot Study*." *Journal of Personnel Evaluation in Education* 15(3): 193-212.

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

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

Murnane, R. J. and B. 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):417-458.

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

Rothstein, J. (2010). "Teacher Quality in Educational Production: Tracking, Decay, and Student Achievement." *Quarterly Journal of Economics* 125(1): 175-214.

Seifert, C. F., G. Yukl, et al. (2003). "Effects of Multisource Feedback and a Feedback Facilitator on the Influence Behavior of Managers Toward Subordinates." *Journal of Applied Psychology* 88(3): 561-569.

Smither, J. W., M. London, et al. (2003). "Can Working With an Executive Coach Improve Multisource Feedback Ratings Over Time? A Quasi-Experimental Field Study." *Personnel Psychology* 56(1): 23-44.

Staiger, D. O. and J. E. Rockoff (2010). "Searching for Effective Teachers with Imperfect Information." *The Journal of Economic Perspectives* 24: 97-117.

Weber, J. R. (1987). "Teacher Evaluation as a Strategy for Improving Instruction. Synthesis of Literature." North Central Regional Educational Laboratory, Elmhurst, IL.

Table 2: Observable Student and Teacher Characteristics of Estimation Sample

	Not In Estimation Sample	Main Estimation Sample		
		Total	Scheduled Participation	Volunteer Participation
	(1)	(2)	(3)	(4)
Student Characteristics				
Baseline Math Score	0.072	0.054	0.043	0.084
Standard Deviation	(1.009)	(0.938)	(0.941)	(0.930)
Baseline Reading Score	0.066	0.072	0.054	0.123
Standard Deviation	(1.000)	(0.937)	(0.950)	(0.899)
Grade 4	22.0%	25.9%	22.3%	36.1%
Grade 5	16.8%	26.0%	22.5%	36.0%
Grade 6	13.0%	18.7%	21.2%	11.6%
Grade 7	25.6%	17.5%	20.0%	10.2%
Grade 8	22.5%	12.0%	14.0%	6.2%
Male	50.4%	48.1%	47.2%	50.4%
Racial/Ethnic Minority	76.1%	79.7%	79.1%	81.1%
White	23.8%	20.4%	20.9%	18.9%
Special Education	19.1%	17.7%	18.2%	16.5%
English Language Learner	2.6%	3.0%	3.6%	1.5%
Gifted & Talented	9.8%	11.4%	11.8%	10.1%
Retained in Grade	1.1%	0.7%	0.8%	0.7%
Number of Students	44,648	14,208	10,503	3,705
Teacher Characteristics				
First Year Teaching	0.9%	0.0%	0.0%	0.0%
1 Year Experience	2.0%	0.0%	0.0%	0.0%
2 Years Experience	2.6%	0.0%	0.0%	0.0%
3 Years Experience	3.0%	0.0%	0.0%	0.0%
4 Years Experience	4.2%	0.9%	0.8%	1.3%
5-9 Years Experience	19.2%	15.6%	14.2%	20.1%
10-19 Years Experience	28.7%	67.8%	68.4%	66.3%
20 or More Years Experience	39.5%	15.6%	16.6%	12.5%
Contract Year 1992 or earlier	36.9%	0.0%	0.0%	0.0%
Contract Year 1993	1.6%	8.6%	10.0%	4.0%
Contract Year 1994	1.8%	18.1%	15.0%	28.0%
Contract Year 1995	0.9%	3.8%	3.7%	4.0%
Contract Year 1996	0.9%	13.3%	16.3%	4.0%
Contract Year 1997	0.5%	21.0%	23.7%	12.0%
Contract Year 1998	0.5%	21.9%	18.8%	32.0%
Contract Year 1999	1.2%	13.3%	12.5%	16.0%
Contract Year 2000 or later	44.9%	0.0%	0.0%	0.0%
Number of Teachers	561	105	80	25

Table 3: Estimated Differences in Math Achievement for Students Taught Before, During, and After TES Participation

	(1)	(2)	(3)	(4)	(5)	(6)
School Year Relative to Year of TES Participation (Years Prior Omitted)						
Year Immediately Prior					0.033	
					(0.042)	
Year Immediately Prior * Scheduled Participant						0.035
						(0.046)
Year Immediately Prior * Volunteer Participant						0.053
						(0.072)
Year of Participation	-0.012	0.072+	0.063+	0.062+	0.086+	
	(0.087)	(0.043)	(0.036)	(0.036)	(0.045)	
Year of Participation * Scheduled Participant						0.077+
						(0.045)
Year of Participation * Volunteer Participant						0.161
						(0.103)
Years After	0.047	0.111*	0.116*	0.113*	0.145*	
	(0.106)	(0.054)	(0.047)	(0.048)	(0.060)	
Years After * Scheduled Participant						0.127*
						(0.059)
Years After * Volunteer Participant						0.226*
						(0.106)
Teacher Fixed Effects		Y	Y	Y	Y	Y
Teacher Experience Controls				Y	Y	Y
Student-level Controls			Y	Y	Y	Y
Teacher Clusters	105	105	105	105	105	105
Student Observations	14,208	14,208	14,208	14,208	14,208	14,208
Adjusted R-squared	0.013	0.245	0.576	0.576	0.576	0.576

Note: Each column represents a separate student-level specification predicting math test score as a function of grade-by-year fixed effects, and the indicated covariates. Student-level controls include prior year achievement (main effect, interaction with grade level, and indicator for missing value) and indicators for gender, race/ethnicity subgroup, special education classification, English language learner classification, gifted and talented classification, and students retained in grade. Clustered (teacher) standard errors in parentheses. ** indicates p<0.01, * p<0.05, and + p<0.10.

Table 4: Within Teacher Variation in Assigned Student Characteristics Relative to TES Participation Timing

	Baseline Math Test Score	Male	Racial/ Ethnic Minority	White	Special Education	English Language Learner	Gifted & Talented	Retained in Grade
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
School Year Relative to Year of TES Participation (Years Prior Omitted)								
Year of Participation * Scheduled Participant	-0.043	0.017	-0.052	0.052	0.016	0.000	0.031	0.001
	(0.046)	(0.016)	(0.039)	(0.039)	(0.017)	(0.004)	(0.028)	(0.004)
Year of Participation * Volunteer Participant	-0.068	-0.023	-0.012	0.012	0.040+	-0.006	-0.013	-0.019
	(0.097)	(0.026)	(0.029)	(0.029)	(0.022)	(0.008)	(0.034)	(0.012)
Years After * Scheduled Participant	-0.023	-0.003	-0.008	0.008	-0.001	0.014	0.011	0.002
	(0.048)	(0.019)	(0.022)	(0.022)	(0.015)	(0.009)	(0.018)	(0.005)
Years After * Volunteer Participant	-0.091	0.002	0.013	-0.013	0.040	0.002	-0.001	-0.017
	(0.107)	(0.025)	(0.039)	(0.039)	(0.025)	(0.006)	(0.022)	(0.012)
Teacher Clusters	105	105	105	105	105	105	105	105
Student Observations	14,208	14,208	14,208	14,208	14,208	14,208	14,208	14,208
Adjusted R-squared	0.185	0.028	0.236	0.236	0.124	0.222	0.123	0.008

Note: Each column represents a separate student-level specification predicting the student characteristic indicated as a function of the indicated covariates, teacher experience, and teacher fixed effects. Clustered (teacher) standard errors in parentheses. ** indicates p<0.01, * p<0.05, and + p<0.10.

Table 5: Heterogeneity in Estimated Difference in Math Achievement for Students Taught After TES Participation

(A) Quartile of Overall TES Score		(B) Quartile of the Change in Overall TES Score from First to Last Observation During the TES Year	
	(1)		(2)
School Year Relative to Year of TES Participation (Years Prior Omitted)		School Year Relative to Year of TES Participation (Years Prior Omitted)	
Year Immediately Prior	0.023	Year Immediately Prior	0.033
	(0.043)		(0.042)
Year of Participation	0.082+	Year of Participation	0.086+
	(0.045)		(0.046)
Years After * Bottom Quartile TES Score	0.279**	Years After * Top Quartile TES Score Change	0.183+
	(0.089)		(0.098)
Years After * 2nd Quartile TES Score	0.116	Years After * 3rd Quartile TES Score Change	0.229**
	(0.088)		(0.084)
Years After * 3rd Quartile TES Score	0.122+	Years After * 2nd Quartile TES Score Change	0.110
	(0.065)		(0.094)
Years After * Top Quartile TES Score	0.090	Years After * Bottom Quartile TES Score Change	0.029
	(0.087)		(0.075)
Teacher Clusters	105	Teacher Clusters	105
Student Observations	14,208	Student Observations	14,208
Adjusted R-squared	0.577	Adjusted R-squared	0.577

Note: Each column represents a separate student-level specification predicting math test score as a function of teacher fixed effects, teacher experience controls, student-level controls, grade-by-year fixed effects, and the indicated covariates. Student-level controls include prior year achievement (main effect, interaction with grade level, and indicator for missing value) and indicators for gender, race/ethnicity subgroup, special education classification, English language learner classification, gifted and talented classification, and students retained in grade. Clustered (teacher) standard errors in parentheses. ** indicates p<0.01, * p<0.05, and + p<0.10.