ANNENBERG
BROWN UNIVERSITY

EdWorkingPaper No. 19-97

Teacher Skill Development: Evidence from Performance Ratings by Principals

Matthew A. Kraft
Brown University

John P. Papay
Brown University

Olivia L. Chi
Harvard University

We examine the dynamic nature of teacher skill development using panel data on principals' subjective performance ratings of teachers. Past research on teacher productivity improvement has focused primarily on one important but narrow measure of performance: teachers' value-added to student achievement on standardized tests. Unlike value-added, subjective performance ratings provide detailed information about specific skill dimensions and are available for the many teachers in non-tested grades and subjects. Using a within-teacher returns to experience framework, we find, on average, large and rapid improvements in teachers' instructional practices throughout their first ten years on the job as well as substantial differences in improvement rates across individual teachers. We also document that subjective performance ratings contain important information about teacher effectiveness. In the district we study, principals appear to differentiate teacher performance throughout the full distribution instead of just in the tails. Furthermore, prior performance ratings and gains in these ratings provide additional information about teachers' ability to improve test scores that is not captured by prior value-added scores. Taken together, our study provides new insights on teacher performance improvement and variation in teacher development across instructional skills and individual teachers.

VERSION: June 2019

Suggested citation: Kraft, Matthew, John Papay, and Olivia Chi. (2019). Teacher Skill Development: Evidence from Performance Ratings by Principals. (EdWorkingPaper: 19-97). Retrieved from Annenberg Institute at Brown University: http://www.edworkingpapers.com/ai19-97

Teacher Skill Development: Evidence from Performance Ratings by Principals

Matthew A. Kraft
Brown University

John P. Papay
Brown University

Olivia L. Chi
Harvard University

Updated: May 2019

Abstract

We examine the dynamic nature of teacher skill development using panel data on principals' subjective performance ratings of teachers. Past research on teacher productivity improvement has focused primarily on one important but narrow measure of performance: teachers' value-added to student achievement on standardized tests. Unlike value-added, subjective performance ratings provide detailed information about specific skill dimensions and are available for the many teachers in non-tested grades and subjects. Using a within-teacher returns to experience framework, we find, on average, large and rapid improvements in teachers' instructional practices throughout their first ten years on the job as well as substantial differences in improvement rates across individual teachers. We also document that subjective performance ratings contain important information about teacher effectiveness. In the district we study, principals appear to differentiate teacher performance throughout the full distribution instead of just in the tails. Furthermore, prior performance ratings and gains in these ratings provide additional information about teachers' ability to improve test scores that is not captured by prior value-added scores. Taken together, our study provides new insights on teacher performance improvement and variation in teacher development across instructional skills and individual teachers.

JEL I21, J24

Correspondence regarding the paper can be sent to Matt Kraft at mkraft@brown.edu. We are grateful for the support of the Charlotte-Mecklenburg Schools. All errors and omissions are our own.

I. Introduction

Over the past two decades, a great deal of policy attention has focused on improving educational outcomes for children by strengthening the quality of the teacher workforce. Alternative certification pathways, recruitment and retention bonuses, teacher selection, and teacher dismissal have all been widely discussed and studied.[1] However, districts still prioritize investments in current teachers' professional development rather than efforts to remake the teacher workforce. For example, Jacob and McGovern (2015) estimate that the 50 largest school districts spend $8 billion annually on teacher development. This dwarfs spending on policy instruments designed to recruit or selectively retain more effective teachers, including prominent federal grants such as Race to the Top, which allocated $4.35 billion over four years.

The scholarly consensus is that current expenditures on formal teacher professional development programming as implemented at-scale across the United States have produced relatively little return on investment (Hill, 2007; Garet et al., 2008; Jacob & Lefgren, 2004; Jacob & McGovern, 2015). However, the limited effectiveness of professional development efforts stands in stark contrast to what we know about teacher improvement. A well-established body of literature on the returns to teacher experience has documented that teachers make large gains in productivity during their initial years on the job, and that these gains can persist well into the mid-career if not beyond.[2] On-the-job experience is among the most reliable and effective drivers of professional improvement for early career teachers. Thus, better

[1] See for example on alternative certification pathways (e.g., Kane, Rockoff, & Staiger, 2008; Boyd, Grossman, Lankford, Loeb, Wyckoff, 2006), on recruitment and retention bonuses (e.g. Steele, Murnane, & Willett, 2010; Feng & Sass, 2018), on teacher selection (Jacob, Rockoff, Taylor, Lindy, & Rosen, 2016; Goldhaber, Grout, & Huntington-Klein, 2017; Rockoff, Jacob, Kane, & Staiger, 2011), and teacher deselection (e.g. Goldhaber & Hansen, 2010; Winters & Cowen, 2013).

[2] See for example Rockoff, 2004; Boyd et al., 2008; Kraft & Papay, 2014; Papay & Kraft, 2015; Atteberry, Loeb, & Wyckoff, 2015; Ladd & Sorensen, 2017.

understanding how and why teachers improve as they gain experience has the potential to provide insights about how to make formal professional development more effective.

Past research on teacher productivity improvement has focused primarily on teachers' value-added to student achievement on standardized tests. These studies have advanced our understanding of teacher productivity dynamics, but face two important limitations. First, they focus on the relatively few teachers for whom value-added estimates are available, only 15 to 20 percent of teachers in most districts. Second, they examine an important but narrow aspect of teachers' jobs – improving student test scores (see Brighouse, Ladd, Loeb, & Swift 2018).

Two recent studies suggest that the returns to experience patterns based on teachers' contributions to student achievement may not fully capture teachers' skill improvement along other important dimensions. Gershenson (2016) and Ladd and Sorensen (2017) estimate returns to teacher experience as measured by student academic practices and behavior in school and find large and sustained improvement on these dimensions of teacher skill.

We build on this literature using panel data on principals' subjective performance ratings of teachers in Charlotte-Mecklenburg Schools (CMS), the 17th largest school district in the country. Over the past decade, such evaluation systems have grown in prominence across the country, such that nearly all teachers are regularly observed in their classrooms (Steinberg & Donaldson, 2016). From 2002 to 2010, CMS principals rated teachers on eight performance areas, providing a detailed window into teachers' skills across a range of pedagogical and professional domains.

These data enable us to make two central contributions to the teacher effectiveness literature. First, we provide more nuanced evidence on the validity of this new generation of subjective performance ratings, documenting that they contain important information about teacher effectiveness despite limited variation in final ratings. Similar to prior studies, we find that principals assign very few performance ratings below satisfactory (Kraft & Gilmour, 2017; Grissom & Loeb, 2017). However, evaluation ratings in CMS are associated with value-added to student achievement across the full performance range, not just in the tails of the distribution as past studies have shown (Jacob & Lefgren, 2008). Furthermore, we document for the first time that gains in performance ratings contain additional information about teachers' ability to improve test scores that are not captured by teachers' prior value-added scores.

Second, we provide new evidence of the dynamic nature of teacher skill development. Performance ratings are available for teachers across grades and subjects allowing for both greater generalizability and the ability to test for previously unexplored differences in returns to experience across teaching assignments. They also offer opportunities to explore mechanisms behind the observed returns to experience captured by test scores. Finally, performance ratings capture a range of teachers' classroom practices and contributions to their school. Understanding teacher performance improvement more broadly is particularly important given evidence that teachers affect students' long-term outcomes through multiple pathways (Jackson, 2018; Petek & Pope, 2016) and that teachers can play a lead role in creating positive school climates that promote student success in school (Johnson et al., 2014).

Using a within-teacher returns to experience framework, we find that, on average, new teachers make large and rapid improvements in their instructional practices throughout their first ten years on the job. This positive, marginally decreasing pattern of instructional improvement over time provides further evidence that challenges popular claims that teachers stop improving on the job after just three years (Gates, 2009). Echoing past work on teacher improvement in raising student test scores, we find that this average profile masks substantial variation in the improvement patterns of individual teachers. We also find suggestive evidence that the degree to which teachers improve differs across schooling levels. Elementary and high school teachers appear to experience faster rates of growth than middle school teachers. We find little evidence of differences in the improvement profiles among teachers in tested vs. non-tested grades or across certification pathways. These findings are consistent across a range of robustness tests examining rater bias, sample bias, and differential attrition.

II. Relevant Literature

A. Principal's Rating of Teacher Performance

Teacher evaluation in the United States has undergone a sea change in the past decade. By the late 2000's, research had clearly documented the large variation in teachers' contributions to student achievement gains (Rockoff, 2004, Hanushek, Rivkin & Kain, 2005), while at the same time showing that nearly all teachers received the same satisfactory performance rating (Weisberg et al., 2009). Emboldened by this evidence, the Obama administration incentivized states to overhaul their evaluation systems with the Race to the Top grant competition and federal waivers from No Child Left Behind (Donaldson & Papay, 2014; Kraft, 2018). By 2016, 44 states had passed legislation that mandated major teacher evaluation reforms (NCTQ, 2016). Beneath the turbulent surface of these contentious reforms, the longstanding practice of administrators evaluating teachers based on classroom observations has remained the dominant component of the evaluation process (Steinberg & Donaldson, 2016). The resilience and pervasiveness of this practice makes understanding what ratings capture and what we can learn from subjective evaluation ratings a continued priority (Cohen & Goldhaber, 2016).

Several prior studies have examined the degree to which performance ratings relate to teachers' value-added scores. These studies cluster in several categories. One line of research examines whether principals can accurately predict which teachers are most effective at raising student test scores, using surveys that ask principals to rate teachers' contributions to students' test-score gains (see below). Another body of work explores ratings of classroom practice by external experts, rather than school-based administrators (e.g,. Kane & Staiger, 2012).

We focus our attention on studies that ask principals or other school-based observers to rate teachers' classroom practices as this is the standard approach in practice.[3] Many such studies rely on low-stakes principal ratings collected by researchers via surveys and interviews (Jacob & Lefgren, 2008; Harris and Sass, 2014; Rockoff et al., 2012), finding unadjusted correlations between principals' ratings of teachers' overall performance and value-added scores that range between 0.19 to 0.29 in math and 0.18 to 0.28 in reading. Jacob and Lefgren (2008) and Harris and Sass (2014) also find that principals are better able to identify the most and least effective teachers in their schools, as judged by value-added, than they are at differentiating teachers in the middle of the performance distribution. Furthermore, Harris and Sass show that principals' overall low-stakes ratings predict future teacher value-added to student achievement in math (but not reading), conditional on prior value-added in the same subject.

A smaller body of work examines the relationship between value-added and high-stakes ratings mandated by districts. Analyzing pilot data from Chicago Public Schools' new evaluation system, Sartain and her colleagues (2011) show that the average value-added scores of teachers rated Unsatisfactory are substantially lower than those rated Distinguished. They also document that principals rate their teachers systematically higher than evaluators from outside the school.

[3] For simplicity, we refer to observers as "principals" throughout the paper because in many settings, including CMS, principals conduct the lion's share of observations.

Kane, Taylor, Tyler, and Wooten (2011) find robust positive relationships between student achievement gains and teachers' evaluation scores based on classroom observation ratings by peer evaluators and administrators in Cincinnati Public Schools. Grissom and Loeb (2017) examine both low-stakes and high-stakes ratings done by principals in Miami-Dade County Public Schools, showing that principals inflate their evaluation ratings on high-stakes measures but preserve a rank order similar to low-stakes ratings.

Taken together, these studies suggest that principals' ratings of practice relate meaningfully to measures of effectiveness derived from test scores, although in most cases they are only able to differentiate teachers at the tails of the distribution.

B. *Returns to Teacher Experience*

Efforts to document the productivity returns to teacher experience have a long history (Murnane & Phillips, 1981). Recent analyses use teacher fixed effects models to isolate within-teacher productivity improvements that accrue as individual teachers gain classroom experience (Rockoff, 2004; Boyd et al., 2008; Harris & Sass, 2011; Kraft & Papay, 2014; Ost, 2014; Papay & Kraft, 2015; Atteberry, Loeb, & Wyckoff, 2015). These studies all find rapid early-career returns to experience, with mixed evidence about the nature of productivity returns after the first 5 to 10 years.

Recent evidence examining teachers' returns to teaching experience on non-test score outcomes suggests teachers also improve substantially in other dimensions over their careers. Ladd and Sorensen (2017) find large and sustained returns to experience in the form of teacher effects on students' behavioral outcomes (attendance & disciplinary referrals) as well as academic behaviors (time spent reading and doing homework outside of school). Gershenson (2016) finds further evidence of teachers' sustained productivity growth as measured by how teachers affect student attendance. Most relevant to our work, Jacob and Walsh (2011) estimate teacher returns to experience based on a single overall subjective performance rating across a four-year panel dataset from Chicago Public Schools (CPS). They find a pattern of rapid positive returns to experience through approximately 8 years of experience that becomes a gradual but steady decline as teachers move into the later years of their career.

We build on and extend the work of Jacob and Walsh (2011) in several important ways. CPS teachers were rated on a single, 4-category scale with 66 percent of teachers earning the top rating. In contrast, teachers in CMS were evaluated on eight different domains using a four-category scale. This allows us to examine differences across specific domains of teacher skills and to construct composite ratings to maximize variation and substantially reduce the proportion of top-coded evaluation ratings to 18 percent. Furthermore, we extend their focus on the average returns to experience profile by documenting large variation in performance improvement across individual teachers and testing for systematic differences in returns to experience estimates based on teachers' schooling levels, certification pathways, and subjects taught.

III. Data & Measures

A. Data

We use a rich administrative dataset from CMS that links students, teachers, and test records across a nine-year panel of data from 2001/02 to 2009/10. Student data include demographic information and annual state test results in math and English language arts (ELA). Human resource data include demographics, teacher experience in the state, and principal evaluation ratings on the Teacher Performance Appraisal Instrument – Revised (TPAI-R).

Our primary outcomes of interest are the subjective performance ratings teachers receive on the TPAI-R. From 2001 to 2010, North Carolina required that public school teachers be evaluated using the TPAI-R. The rubric is based on the North Carolina Professional Teaching Standards and consists of eight overall domains defined by 43 skill indicators (see Appendix B for the full instrument). The eight domains covered are:

- *Management of Instructional Time*: Teacher is prepared at the start of the lesson, gets the class started quickly, and uses time for learning.
- *Management of Student Behavior*: Teacher has an established set of rules that govern classroom behavior, frequently monitors student behavior and addresses inappropriate behavior.
- *Instructional Presentation*: Teacher instruction demonstrates strong disciplinary knowledge and clear learning objectives. Teacher uses questions effectively, checks for understanding, and develops critical thinking and problem solving skills.
- *Instructional Monitoring*: Teacher frequently checks and assesses student learning and adjusts instruction to students learning needs.
- *Instructional Feedback*: Teacher provides regular and prompt feedback to students, probing incorrect responses and encouraging productive interaction among peers.
- *Facilitating Instruction*: Teacher develops instructional materials aligned with district and state goals, integrates diverse instructional resources into the class, and plans appropriate instruction for diverse learners.
- *Communicating within the Education Environment*: Teacher treats all students fairly and engages members of the broader school community to support student learning.
- *Performing Non-Instructional Duties*: Teacher carries out non-instructional duties and adheres to laws, policies, and regulations. Teacher engages in reflection and professional development in an effort to grow professionally.

Evaluators assign a score for each domain using a four-point scale (i.e., Unsatisfactory, Below Standard, At Standard, Above Standard) based on evidence collected through classroom observations, teaching artifacts, and teacher conferences. Probationary teachers are evaluated based on three full observations by administrators and one peer observation. Tenured teachers are evaluated based on one full observation and two short snap-shot observations by administrators. Full observations are followed by post-conferences to provide feedback.

District documents identify a threefold purpose of the evaluation system: to serve as a guide for teacher reflection on their practice, to promote instructional improvement, and to measure performance for accountability. During the years we study, probationary teachers could be non-renewed at the discretion of a principal based on their evaluation. Tenured teachers who received a Below Standard or Unsatisfactory rating were placed on an action plan and evaluated again the following year. If a teacher did not improve to at least At Standard, the superintendent could move towards termination. This process, however, was not automatic and occurred rarely.

Figure 1 illustrates the percentage of teacher-year observations for which we have valid TPAI-R performance ratings by experience. Evaluation scores are available for approximately 65 percent of all teacher-year records for probationary teachers, suggesting that these non-tenured teachers were evaluated frequently but not always annually. Coverage for evaluation scores drops from 44 percent for teachers with 5 years of experience, to 29 percent for teachers with 10 years of experience, to only 19 percent for teachers with 20 years of experience. This accords with requirements that tenured teachers be evaluated only once every five years.

We describe the characteristics of CMS teachers and the schools in which they taught in Table 1. District-wide, 79 percent of CMS teachers are female, while 72 percent are white and 24 percent are African American. Thirty-three percent hold advanced degrees and nine percent are National Board certified. Our descriptive sample consists of 9,821 out of the 17,398 teachers in our data with an evaluation score in at least one year we observe. Teachers for whom we do not have any record of an evaluation score primarily consist of tenured teachers with many years of experience and probationary teachers who taught in CMS for only one year.

Teachers who receive evaluations in a given year have broadly similar demographic characteristics as those who do not, although they have substantially less experience, on average, given that less experienced teachers were evaluated more frequently.[4] To formally test for differences across these samples, we compare probationary and tenured teachers separately. We find small differences across teacher and school characteristics in both groups, although these differences are statistically significant. Among probationary teachers, those who have performance ratings are more likely to be African American and less likely to be white. They also have less experience, on average, and are more likely to teach in schools with a larger share of Hispanic students, English Language Learners (ELL), and special education students. Among tenured teachers, those who are evaluated are more likely to be non-white and less likely to hold an advanced degree or National Board certification. Similarly, tenured teachers with evaluation ratings also have fewer years of experience, on average, and are more likely to teach in schools with lower average test scores and higher shares of non-white and ELL students.

These descriptive patterns suggest that our results may only generalize to teachers who received evaluation ratings. We present additional analyses below that suggest these patterns do not present serious consequences for the internal validity of our results. In particular, we further examine selection patterns into the analysis sample and find no evidence that the probability of being evaluated was related to prior performance.

B. Performance Evaluation Measures

A principal components analysis (PCA) of the ratings on the eight conceptually distinct domains of the TPAI-R provides evidence that the instrument appears to capture one primary

[4] CMS recognizes teaching experience in other North Carolina districts for the purposes of experience credits towards the salary schedule. However, experienced teachers who transfer into CMS become probationary teachers for at least one year. This explains why the average experience of probationary teachers is greater than 5 years.

latent construct with similar weights across items.[5] Thus, we conduct our primary analyses using a single overall performance score, while also describing patterns in the raw scores across the eight individual domains. Similar to examining student performance on sub-domains sampled on standardized math achievement tests, examining teacher performance on individual domains can shed important light on the nature of specific teacher skills.

We construct our preferred measure of overall performance using a graded response model (GRM). This flexible approach allows us to construct theta scores that incorporate information about the difficulty of individual evaluation domains and the degree to which domains successfully differentiate among individual teachers. GRMs are an extension of item response theory (IRT) that are commonly used in psychological and educational measurement to study items (i.e., survey or test questions) and instruments designed to measure latent traits. IRT models estimate the probability that an individual successfully responds to an item as a function of the item's characteristics and the level of the individual's latent trait. GRMs extend IRT models to ordered categorical items (Samejima, 1969) such as the ratings on the TPAI-R.

To estimate teachers' theta scores, we first fit a GRM of the following form:

$$\Pr\left(R_{dj} \geq k \mid a_d, b_{dk}, \theta_j\right) = \frac{1}{1+\exp\left(-a_d(\theta_j - b_{dk})\right)} \tag{1}$$

where R_{dj} is the rating that teacher j received on domain d, and k refers to the categories of ratings (e.g., on a four-point scale, $k = 1, \ldots, 4$). θ_j represents the latent construct of interest (i.e., teaching skills), which is assumed to have a standard normal distribution in the population. b_{dk} is the difficulty parameter for response category k on domain d. Estimates of b_{dk} correspond to locations on the theta scale, such that a teacher with ability equal to b_{dk} has a 50 percent chance

[5] The Eigenvalue of the first principal component is 4.70, with loadings on each domain that range from 0.29 to 0.37. The Eigenvalue of the second principal component is 0.81, and the scree plot of eigenvalues presents a clear kink at the second principal component.

of receiving a rating of k or higher on domain d. a_d is the discrimination parameter for domain d; a domain rating with a larger discrimination parameter is better at distinguishing between different levels of teaching skills. We estimate teacher j's theta score $\widehat{\theta_j}$ using the posterior means (empirical Bayes) estimates of the latent trait. We fit this model separately in each school year to allow for changes in item difficulty and discrimination across our panel. We then standardize our estimated theta scores within-year to have a mean of zero and unit standard deviation.

These theta scores have several advantages over the standard practice of taking a simple average of ratings across domains. First, the theta scores allow for differences in difficulty across domains on the TPAI-R, whereas a simple average of ratings does not. Second, in estimating the difficulty parameters, the GRM is flexible and does not assume an equal interval scale. Third, the theta scores account for differences in how well items differentiate between teachers with similar skill levels, whereas a simple average does not allow for such differences.

By taking into account differences in difficulty and discrimination across domains, theta scores from the GRMs result in far more variation in scores than taking a simple average. For example, average ratings result in only 39 total unique score values and a concentration of 24 percent of teachers with an average score of 3. The variation in theta scores is substantially greater, with over 2,600 discrete values and no single value representing more than 5 percent of the observations. Figure 2 illustrates how for every mean raw score there is a range of corresponding theta scores. Thus, using theta scores increases precision in our analyses and substantially reduces the degree to which performance ratings suffer from ceiling effects.

Recent research suggests that teachers' classroom composition can influence the observation scores they receive (Whitehurst, Chingos, & Lindquist, 2014; Steinberg & Garrett, 2016; Campbell & Ronfeldt, 2018). Motivated by these findings, we also construct a residualized theta score by removing systematic variation explained by classroom- and school-level average student demographic characteristics. In our data, the correlation between unadjusted theta scores and residualized theta scores is 0.97. Given this strong correlation, it is not surprising that the results we report below are nearly identical across measures.

Teacher value-added

We estimate teachers' value-added to test scores in 4th through 8th grade for math and ELA test scores separately using a standard model in the literature (Koedel, Mihaly, & Rockoff, 2015) as follows:

$$A_{isgjt} = \Omega_g\left(f\left(A_{is,t-1}\right)\right) + \gamma\boldsymbol{X}_{igjt} + \delta_{jt} + \vartheta_g + u_{isgjt} \tag{2}$$

where A_{isgjt} is the standardized test score of student i in subject s in grade g taught by teacher j in time t. We control for prior achievement in math and ELA test scores, as well as their squared and cubed values. We also include the interactions between the linear prior year test score terms in subject s and indicators for grade g. $\boldsymbol{X}_{igjt}$ includes sets of student-, classroom-, and cohort-level covariates. Student-level covariates include indicators for gender, race, English Language Learner status, and special education status. Classroom-level covariates include the mean prior math and ELA achievement of all students taught by teacher j in time t in grade g, as well as the means of the student-level covariates for those students. Cohort-level covariates are defined for students attending the same school and in the same grade as student i in time t. ϑ_g are grade-level indicators, and u_{isgjt} represent the mean zero idiosyncratic error term. Our teacher-by-year fixed effects δ_{jt}, represent teacher j's value-added to test scores in time t. In analyses where teacher value-added estimates are treated as independent variables, we use empirical Bayes (i.e., shrunken) estimates to account for attenuation bias (Jacob & Lefgren, 2008).[6]

One potential concern is our teacher value-added estimates are conflated with school effects, which could bias our analyses of the relationship between teachers' value-added scores and subjective performance ratings. We address this concern by estimating alternative value-added estimates following Cohodes, Setren and Walters (2019) in which we replace teacher-by-year fixed effects in equation (2) with teacher-by-year random effects and school random effects. These alternative value-added estimates are highly correlated with our teacher-by-year fixed effect estimates (0.90 in math and 0.87 in ELA), and applying these alternative estimates in the analyses described below results in quite similar findings.

IV. Empirical Strategy

Our central research questions explore the validity of subjective performance ratings and the skill returns to teaching experience. Specifically, we ask:

(1) What is the relationship between subjective performance ratings and value-added estimates of teacher effectiveness?

(2) How do teachers improve their instructional skills, on average, as measured by performance ratings, throughout their careers?

(3) How do these returns to experience vary across teachers and schools?

A. The relationship between performance ratings and value-added

[6] To construct the empirical Bayes estimates, we first obtain the method of moments estimate of v_t^2, the between-teacher variance for teachers in year *t*, by subtracting the average squared standard error of the fixed-effects estimates in year *t* from the sample variance of the fixed effects estimates in year *t*. Then we construct the shrinkage factor $\lambda_{jt} = \frac{v_t^2}{v_t^2+se_{jt}^2}$ for each fixed effect, where se_{jt}^2 is the squared standard error of the fixed effect for teacher *j* in year *t*. Lastly, we construct $\hat{\delta}_{jt}^{EB}$, the empirical Bayes estimate for teacher *j* in year *t* such that:

$$\hat{\delta}_{jt}^{EB} = (1-\lambda_{jt})\overline{\delta_t} + \lambda_{jt}\widehat{\delta_{jt}}$$

where $\overline{\delta_t}$ is the mean of the fixed effects estimates in year *t*.

We explore the relationship between subjective performance ratings and value-added in two ways. First, we estimate correlations between these measures, focusing on differences across the teacher effectiveness distribution. Second, we examine the extent to which performance ratings, and changes in ratings, predict future productivity. Both of these approaches provide important evidence about the validity and utility of these ratings.

We estimate both unadjusted and adjusted correlations. Note that while we would not expect perfect correlations between these measures because they intentionally measure different elements of teaching performance, we would expect them to be related. The unadjusted correlations are simple pairwise Pearson product-moment correlations. While straightforward, these unadjusted correlations are necessarily attenuated, as they do not account for estimation error in the value-added measures or the measurement error in subjective performance score.

To correct for the estimation error in the fixed effect value-added measures, we adopt the strategy of Jacob and Lefgren (2008) and scale our estimates using a measure of the reliability of the observed teacher effect estimates. To begin, we express our observed teacher effect measure $\widehat{\delta^{VAM}}$ as the sum of the true effect δ^{VAM} and estimation error e: $\widehat{\delta^{VAM}} = \delta^{VAM} + e$. Here, the estimated reliability of the observed value-added measure ($\widehat{\alpha_{VAM}}$) can be expressed as $\frac{\mathrm{Var}(\delta^{VAM})}{\mathrm{Var}\left(\widehat{\delta^{VAM}}\right)}$, or the proportion of the observed variance of value-added scores that is accounted for by true teacher effects. To obtain a measure of the variance of the true teacher effects, Var(δ^{VAM}), we estimate the mean error variance by taking the average of the squared standard errors of the value-added fixed effects estimates and then subtracting this estimate of the mean error variance from the variance of the observed value-added scores.

To correct for the measurement error in the teacher's subjective performance ratings, $\widehat{\theta_J}$, we first calculate a measure of the group-level reliability of the theta scores ($\widehat{\alpha_{EVAL}}$), which can be expressed as $\frac{\mathrm{Var}(\theta)}{\mathrm{Var}(\hat{\theta})}$ (Raju, Price, Oshima, & Nering, 2007). To estimate the variance of the true theta scores $\mathrm{Var}(\theta)$, we subtract the mean of the squared conditional standard errors of measurement (CSEMs) from the variance of the observed theta scores $\mathrm{Var}(\hat{\theta})$. We then disattenuate the correlation by multiplying by the inverse of the product of the square root of the reliability of these two measures following Spearman (1904). Our adjusted correlations can be expressed as:

$$\mathrm{Corr}\left(\widehat{\delta^{EVAL}}, \widehat{\delta^{VAM}}\right) * \frac{1}{\sqrt{(\widehat{\alpha_{EVAL}} * \widehat{\alpha_{VAM}}}} \tag{3}$$

This disattenuation approach is still likely to be a conservative correction given that we do not account for all sources of error such as the rater (principals) and the number of classes observed. We calculate bootstrapped standard errors for the unadjusted and adjusted correlations using 1,000 bootstrap replications.

Following Jacob and Lefgren (2008), we also examine whether the correlation between evaluation ratings and value-added changes when we exclude the teacher-years with evaluation ratings that are in the top and bottom quintiles of the ratings distribution. This exercise helps to reveal the extent to which principals are able to distinguish between teachers' performance as judged by value-added among teachers in the middle of the performance distribution.

We further examine whether prior performance ratings provide additional information about teacher effectiveness in raising student test scores not captured by past value-added estimates and whether gains in these ratings relate to gains in student achievement. Specifically, we modify equation (2) above by removing the teacher-year fixed effects and adding combinations of three additional predictors: teacher j's evaluation rating in year $t-1$, $(EVAL_{j,t-1})$; the change in teacher j's evaluation ratings from year $t-1$ to year t,

($\mathit{\Delta EVAL}_{j,(t-(t-1))}$), and teacher j's empirical Bayes estimate of value-added in year $t-1$, (VA_{t-1}). The coefficients associated with $EVAL_{j,t-1}$ and $\mathit{\Delta EVAL}_{j,(t-(t-1))}$ test whether prior evaluation scores and gains in evaluation scores (from the prior year to the current year) have additional predictive power for how effective teachers are in raising student achievement above and beyond prior-year value-added estimates. In our analysis, we fit this model separately for math and ELA teachers and cluster standard errors at the teacher level.

B. Estimating the returns to teacher experience

We estimate the within-teacher returns to experience using a modified version of the "indicator model" (see Papay & Kraft, 2015 for a detailed discussion), a common approach in the literature (e.g., Harris & Sass, 2011; Ladd & Sorenson, 2017). We estimate the average improvement in performance ratings for individual teachers, using teacher fixed effects models to compare teachers to themselves and focus on within-teacher improvement.

A central consideration with these models is how to parameterize teacher experience. We model the first ten years as a completely flexible set of indicator variables and include bins for higher ranges of experience as follows: 11-15 years of experience, 16-20 years of experience, and 21 to 25 years of experience. The inclusion of these binned experience ranges in the later years allows us to estimate simultaneously both experience and year effects without relying exclusively on a small fraction of teachers with discontinuous career histories. This specification relies on the assumption that individual teachers do not improve, on average, within the binned experience ranges. If teachers do improve within these ranges, our estimates will understate the returns to experience (Papay & Kraft, 2015). Our baseline model is a parsimonious specification that takes the following form:

$$y_{jt} = \varphi * f\left(Exper_{jt}\right) + \kappa \boldsymbol{C}_{jt} + \varsigma \boldsymbol{S}_{jt} + \omega_j + \phi_{gt} + \epsilon_{jt} \tag{4}$$

We restrict the sample to observations in which teachers have less than or equal to 25 years of experience, but focus our attention on estimates from the first ten years of a teacher's career given the rapidly decreasing sample of teachers with evaluation scores who have more than 10 years of experience (see Figure 1). In all models, the omitted experience category is zero years of experience (novices). ω_j represents the critical teacher fixed effects, which allow us to examine within-teacher returns to experience. ϕ_{gt} represents grade by year fixed effects and ϵ_{jt} is the error term.[7]

In our preferred models, we also include a vector of classroom-level demographic characteristics ($\boldsymbol{C}_{jt}$) of the students linked to teacher j at time t, including gender, race, ELL status, and special education status, a vector of school-level demographic characteristics, $\boldsymbol{S}_{jt}$**,** similar to those at the classroom-level, as well as the percent of the school's students that are eligible for free or reduced price lunch.[8] Across all returns to experience models, we cluster standard errors at the teacher level.

C. Examining heterogeneity in the returns to experience

We complement estimates of the average returns to experience by examining the degree to which individual teachers and subgroups vary in their improvement rates over time. We first conduct a variance decomposition of evaluation scores by fitting an unconditional multilevel model with school and teacher random intercepts. Then we model differences in the returns to experience across individual teachers by estimating equation (4) within a multilevel model framework with teacher-specific random intercepts and random-slopes (Kraft & Papay, 2014).

$$y_{jt} = \varphi * f\left(Exper_{jt}\right) + \kappa \boldsymbol{C}_{jt} + \varsigma \boldsymbol{S}_{jt} + \phi_{gt} + \left[\alpha_j + \beta_j Exper^{*}_{jt} + \epsilon_{jt}\right] \quad (5)$$

[7] We use the modal grade level among all students taught by a given teacher *j* in year *t* when teachers teach across multiple grade levels.

[8] Information on individual students' eligibility for free or reduced-price lunch is not available in our data.

$$\text{where } \begin{bmatrix} \alpha_j \\ \beta_j \end{bmatrix} \sim N\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} \sigma^2_{\alpha_j} & \sigma_{\alpha_j\beta_j} \\ \sigma_{\beta_j\alpha_j} & \sigma^2_{\beta_j} \end{bmatrix}\right)$$

Here, the fixed portion of the model mirrors our baseline specification. That is, we include indicator variables for the first ten years of experience, indicators for bins of experience in the ranges 11-15, 16-20, and 21-25 years of experience, classroom and school demographic characteristics, and year fixed effects. We replace teacher fixed effects with teacher random intercepts α_j to model individual teacher effects. We also include random slopes β_j on a linear term for experience, *Exper**, that is censored at 11 years to allow each teacher's returns to experience to deviate from the average profile. We then examine the estimated variance $\widehat{\sigma^2_{\beta_j}}$ of the random slopes β_j to quantify the variation in returns to experience across individual teachers.

Equation (5) provides a computationally parsimonious approach to testing for heterogeneity in teachers' returns to experience. However, it imposes the assumption that teachers' individual deviations from the average non-parametric returns to experience profile are linear. We test the sensitivity of our estimates to this assumption by allowing teachers' individual random-slopes to take on a quadratic functional form. Specifically, we replace the linear term $\beta_j Exper^*_{jt}$ in Equation (5) with $\beta_{j1} Exper^*_{jt} + \beta_{j2}\left(Exper^*_{jt}\right)^2$. Comparing results between our linear and quadratic random-slopes models helps to inform our understanding of how the assumption of linear individual deviations affects the magnitude of our estimates.

We extend these heterogeneity estimates by examining whether improvement rates differ systematically among subgroups in which prior research has documented meaningful differences in the labor market. Specifically, we examine differences across 1) teachers in tested versus non-tested grades and subjects (Cohen-Vogel, 2011; Grissom, Kalogrides, & Loeb, 2017); 2) teachers in elementary, middle, and high schools (Boyd, Lankford, Loeb, Ronfeldt, & Wycoff, 2011); and

3) teachers who entered the profession through traditional or alternative certification pathways (Kane, Rockoff, & Staiger, 2008; Papay, West, Fullerton & Kane, 2012).

V. Findings

A. Teacher skills as rated by principals

In Table 2, we present the distribution of raw evaluation ratings for all teachers and novice teachers separately. Similar to prior studies, we find that principals assign very few performance ratings below the satisfactory rating of At Standard on the four-category performance scale (Kraft & Gilmour, 2017; Grissom & Loeb, 2017). Less than 3 percent of teachers' performance ratings included at least one domain scored either Unsatisfactory or Below Standard. As shown in Figure 3, novice teachers in CMS are most likely to be rated Unsatisfactory or Below Standard on managing student behavior. This suggests that teacher preparation and induction programs should ensure teachers receive dedicated instruction in classroom management, ample opportunities to practice management techniques, and individualized feedback. Interestingly, novice teachers were most likely to excel at professional responsibilities outside the classroom. Novice teachers are most likely to be rated Above Standard on communicating within the education environment and performing non-instructional duties. [9]

B. Validity evidence for principals' assessments of teacher performance

We examine the predictive validity of evaluation ratings on the TPAI-R by analyzing their relationship with teachers' value-added to student achievement. Given our focus on the

[9] Paired t-tests confirm that the differences in the percentage of teachers rated below At Standard are significantly different between nearly all of these domains, including those discussed in the text.

dynamic nature of teacher performance, we examine unadjusted and adjusted correlations between evaluation ratings and value-added scores from the same year.

As shown in Table 3 Panel A, we find unadjusted correlations of 0.23 using math value-added scores and 0.13 using ELA value-added scores. Disattenuating these correlations for measurement and estimation error increases them to 0.29 and 0.19, respectively. These estimates are quite similar to the correlations Grissom and Loeb (2017) found between high-stakes performance ratings and drift-adjusted value-added (0.31 in math and 0.22 in reading). They are slightly larger than the adjusted correlations between math value-added and ratings of teachers' digitally recorded lessons by trained independent raters on four observational instruments in the MET study, which ranged between 0.16 to 0.26 (Kane & Staiger, 2012). It is possible, however, that our estimates are biased upwards due to the influence of the same set of students on a teachers' performance rating and contribution to test score growth. We test the robustness of these findings to the potential threat of correlated errors by using value-added estimates from the prior year and find nearly identical results (see Table 3 Panel B). Thus, performance ratings do reflect, in part, differences in teachers' contributions to student test scores.

We further explore the degree to which these correlations are driven by principals' abilities to identify teachers at the tails of the distribution. Following Jacob and Lefgren (2008), we restrict the sample to only those teachers who were rated in the middle three quintiles of the evaluation rating distribution.[10] As seen in Table 3 Panel A, the estimated correlations are somewhat attenuated, from 0.29 to 0.20 in math and 0.19 to 0.15 in ELA, but the broader conclusions remain the same. These results suggest that principals are able to differentiate teacher performance meaningfully, even among teachers in the middle of the distribution.

[10] Jacob and Lefgren (2008) exclude teachers with top and bottom ratings on an ordinal scale.

A second set of validity evidence includes the degree to which performance ratings capture additional information about teacher productivity that is relevant to student achievement, above and beyond what is reflected in prior value-added estimates. As shown in Table 4 column 1, a one standard deviation (SD) higher prior-year performance rating is associated with approximately a 0.031 SD increase in math and 0.010 SD increase in ELA test score gains. In comparison, a one teacher-level SD higher value-added score from the prior year is associated with a 0.094 SD increase in math and 0.025 SD increase in ELA test score gains (column 2). As shown in column 3, performance ratings contain additional predictive power for test score gains even when we condition on teachers' performance in the prior year as measured by value-added.

The degree to which teachers are improving, as judged by principals on subjective performance ratings, also predicts student achievement gains in math. A one SD increase in teachers' performance score gain is associated with a 0.013 SD gain in math achievement (column 4). Controlling for prior value-added scores in math only attenuates the coefficient on performance gains by a third from 0.013 to 0.009 SD. We find positive, but small and statistically insignificant associations between performance score gains and gains in ELA. Overall, both correlational and conditional associations suggest that principals' subjective ratings reflect teachers' ability to raise student performance on standardized tests and capture additional information about teacher performance beyond test-score based measures of productivity.

C. Teacher skill development

First, we describe how teacher performance in each of the eight evaluation domains changes over time among a balanced sample of 1,296 teachers. This exercise compares raw scores among the same group of teachers in their first and fourth years of teaching and helps to shed light on trends in teacher performance across specific skills. Novice teachers have a mean raw score of 3.27, approximately equivalent to receiving At Standard ratings on six domains and Above Standard ratings on two domains. Thus, there exists meaningful room for growth among novices despite the relatively high overall ratings.

As shown in Figure 4 and reported in Appendix Table A1, we see meaningful improvements, on average, in domain-specific raw performance ratings for teachers across these four years in all domains. Average gains on the four-point rating scales are relatively consistent across domains, ranging between 0.21 and 0.29 score points, which equate to 0.40 and 0.56 SD. On average, improvements are steepest in these early years for Facilitating Instruction (0.56 SD) and shallowest for Communicating within the Educational Environment (0.40 SD), while the average gains across all other dimensions are statistically indistinguishable from each other.

Among this balanced sample, we also see initial evidence of heterogeneity in teachers' improvement trajectories. From their first to their fourth years, approximately 10 percent of teachers receive lower ratings on a given individual performance domain, 55 percent receive the same rating, and 35 percent receive a higher rating. Taking a simple average across these eight domains reveals that 19 percent of teachers experience a decline in their overall mean raw score, 16 percent remain unchanged, and 65 percent improved their overall performance. Thus, ratings do not simply increase monotonically for all teachers, although most teachers do indeed improve.

In Figure 5, we depict our model-based estimates of the average within-teacher returns to experience based on teachers' theta scores from our preferred specification that includes class- and school-level controls. This profile demonstrates that, on average, teachers make rapid improvements in their early careers and suggests more gradual but sustained improvement through year ten. We estimate that after ten years, teachers have improved, on average, by 0.82 SD relative to their performance as novices (see Appendix Table A2 column 2). These estimates are on par with average returns to experience on teacher value-added; in general, past research suggests that teachers improve their performance by about 0.10 to 0.20 student-level standard deviations by year 10 (Rice, 2013), or roughly between 0.6 and 1.2 teacher-level standard deviations (Hanushek & Rivkin, 2010; Jackson, Rockoff, & Staiger, 2014).

We replicate these returns to experience analyses using mean raw scores and standardized mean raw scores to provide further intuition about the magnitude of our findings. After ten years, we find that teachers improve, on average, by 0.33 score points on the raw scale (which ranges from 1 to 4). This equates to a teacher moving from a rating of At Standard to a rating of Above Standard on three of the eight TPAI-R domains. The total estimated returns to experience after 10 years from a model that uses standardized mean raw scores is 0.82 SD, nearly identical to our findings using theta scores, which are also standardized.

D. Heterogeneity in skill development

While understanding the teacher improvement trajectory on average is important, better understanding heterogeneity in the returns to teaching experience is even more directly relevant to policy. An initial way to explore the degree to which individual teachers' performance changes over time is to examine the proportion of variation in performance scores that is within individual teachers across time as compared to between teachers or between schools. We decompose the variation in performance ratings using an unconditional multilevel model with school and teacher random intercepts and find significant variation in scores across all three levels. As shown in Table 5, the majority of the variation in performance ratings, 54 percent, is within teachers over time. In contrast, just over 31 percent of the variation is within schools across individual teachers and less than 15 percent is explained by average differences across schools. This simple variance decomposition illustrates that teacher effectiveness is not fixed, but instead changes across time. In fact, changes within individual teachers over time explains substantially more variation than average differences in ratings across individual teachers. Our results are also consistent with the large body of literature that documents how teacher quality varies more within schools than across schools.

D.1. Individual Variation

We formally test for individual differences in teacher improvement by fitting the multi-level returns to experience model described in equation (5). Our returns to experience analyses to this point have focused on average profiles across all teachers. Prior research has found that these average profiles mask considerable heterogeneity across individual teachers' productivity improvement as measured by contributions to test scores (Kraft & Papay, 2014; Atteberry, Loeb, & Wyckoff, 2015). Our results, reported in Table 6, provide further evidence of differential returns to experience across individual teachers.

We find that the standard deviation of individual linear deviations from the average curvilinear returns to experience profile reported in column 1 is 0.06 evaluation score SDs. This suggests that a teacher who is at the 75th percentile of returns to experience rates is improving her performance by 8 percent of a SD more annually than a teacher whose improvement is at the 25th percentile of the distribution of improvement rates. As shown in Figure 6 Panel A, a prototypical teacher at the 75th percentile of growth rates is improving steadily throughout their first ten years on the job while the performance of a teacher the 25th percentile plateaus after only a few years. Consistent with prior research, we also find a negative correlation between teachers' initial performance rating and their rate of improvement.

In further analyses, we find that our quadratic specification of individual deviations reported in column 2 is statistically significant. We plot the same corresponding improvement profiles for teachers at the 25th and 75th percentiles in order to compare the magnitude of our estimates across models.[11] As shown in Figure 6 Panel B, allowing for quadratic deviations results in an even larger divergence between these returns to experience profiles. If anything, our linear random-slopes model may understate the degree of heterogeneity across individual teachers.

Identifying individual heterogeneity suggests that there is potential for improved teacher policy and practice to promote more rapid development. It also highlights the importance of examining differences by key subgroups of teachers. We further leverage our evaluation data for the full set of K-12 classroom teachers in CMS to test for potential differences in returns to experience across important subgroups: teachers in tested and non-tested subjects, those in different school levels, and those from alternative and traditional preparation pathways. We present these results in Figures 7 through 9 and report the estimates in Appendix Table A3.

D.2. High-Stakes vs. Low-Stakes Classrooms

Almost all prior estimates of returns to experience have been limited to samples of upper elementary and middle school teachers given these are the grades in which annual standardized tests are most often available. Our data allows us to examine whether elementary and middle school teachers assigned to tested grades and subjects receive higher ratings, on average, and are improving at greater rates than their peers in grades and subjects that are not part of school accountability systems.[12] We exclude high school teachers from this analysis because the tested

[11] We apply the formula for the variance of the sum of random variables to calculate the estimated standard deviation of the sum of the linear and quadratic terms in the random part of our model: $\sqrt{Var\left(\beta_{j1}Exper_{jt}^{*} + \beta_{j2}\left(Exper_{jt}^{*}\right)^{2}\right)} = \sqrt{\left(Exper_{jt}^{*}\right)^{2}Var(\beta_{j1}) + \left(Exper_{jt}^{*}\right)^{4}Var(\beta_{j2}) + 2\left(Exper_{jt}^{*}\right)^{3}Cov(\beta_{j1}, \beta_{j2})}$. We then plug in our sample estimates of the variance and covariance terms to construct these profiles.

[12] In CMS, tested grades and subjects included in school accountability measures include 3rd through 8th grade math and ELA and 5th and 8th grade science.

content on the high school portions of the North Carolina Testing Program may span multiple courses and grades, resulting in a lack of clarity as to which high school teaching assignments are high- or low-stakes.

We find that, on average, principals rate teachers in high-stakes classrooms 0.079 SD (*p*<.001) higher than their peers in low-stakes classrooms. This difference is even slightly larger (0.101 SD, p<.001) when we control for teacher experience and restrict comparisons to teachers in the same school by including school fixed effects. Several different human capital practices might explain the higher performance ratings of teachers in high-stakes classroom: greater on-the-job improvement, selective retention patterns, systematic reassignment patterns, and systematic hiring patterns. We do not find evidence that teachers in tested and non-tested grades and subjects differ in the rate at which they are improving on the job. The two returns to experience profiles shown in Figure 7 track each other relatively closely and are not statistically different from each other.

We explore differential retention, reassignment, and hiring patterns across teachers in high- and low-stakes classrooms in Table 7. The table includes results from several linear probability models examining the relationship between teachers' performance scores and the probability they return to their school and teach in a tested or non-tested grade and subject, controlling for experience. Consistent with prior evidence, we find that teachers who are judged to be higher performing by principals are more likely to return to their schools. A one SD increase in performance ratings is associated with a 2.9 percentage point increase in retention.

Columns 2 through 6 of Table 7 suggest that the higher performance of teachers in high-stakes classrooms is at least partially due to the systematic retention of higher-performing teachers in these classrooms as well as the strategic reassignment of lower-performing teachers in these classrooms to low-stakes classrooms. Among teachers in high-stakes classrooms, a one SD higher performance rating is associated with a 4.4 percentage point increase in the probability that they remain at their school in a high-stakes classroom next year (column 2). At the same time, a one SD *lower* performance rating is associated with a 1.4 percentage point increase in the probability that a teacher in a high-stakes classroom is reassigned to a low-stakes classroom the following year (column 3). We find no evidence that higher-performing teachers in low-stakes classroom are systematically assigned to high-stakes classrooms (column 4) or that more effective newly-hired transfer teachers are systematically assigned to high-stakes classrooms (column 6).

D.3. School Levels

Figure 8 depicts the growth profiles of elementary, middle and high-school teachers. The three profiles track each other relatively closely through the first three years but diverge after year three when middle school teachers' improvement slows, on average, while elementary and high school teachers continue to improve, resulting in a gap of approximately 0.25 SD. A joint significance test of the differences between the coefficients on the experience indicators in years 1 through 10 for elementary and middle school teachers is marginally significant (p = 0.09), while that for high school and middle school teachers is not statistically significantly different (p = 0.38).

Our relatively limited sample size restricts our ability to explore the mechanisms that underlie this difference in improvement rates. However, the slower rate of professional growth among middle school teachers is at least consistent with prior evidence that middle schools are particularly challenging work environments (Moore, 2012; Marinell & Coca, 2013). We also find evidence that CMS middle schools are particularly unstable work environments characterized by high turnover rates. As shown in Table 8, over 29% of middle schools teachers leave their schools each year, compared to less than 24% of elementary and high school teachers. Three-year turnover rates among middle schools are almost 60%, a full 10 percentage points higher than elementary and high schools. These higher rates of turnover likely cause organizational disruptions that affect teacher effectiveness and may reduce their on-the-job learning (Ronfeldt, Loeb, & Wyckoff, 2013). Other possible explanations include differential teacher sorting and attrition patterns by performance improvement rates across school levels. We explore these explanations in supplemental analyses and find little empirical support although we lack the precision to rule them out.

D.4. Licensure Pathways

As shown in Figure 9, we find nearly identical returns to experience profiles when comparing teachers who enter the profession through alternative and traditional pathways. Thus, it does not appear that teachers from alternative routes improve at greater rates than those from traditional teacher education programs in CMS as suggested by past studies.

Taken together, these heterogeneity analyses document substantial variation in individual rates of improvement, but more limited evidence that these differences are a product of teaching assignments or certification pathway.

E. Robustness Tests

We test the sensitivity of our main returns to experience estimates using a range of alternative model specifications and report the results in Appendix Table A2. One possible threat to validity involves the potential dynamic sorting of students to teachers over time. For example, if the types of students that teachers were assigned changed systematically as they gained experience, and if student characteristics were related to teacher performance ratings, our estimates could be biased.[13] We explore this threat by examining the sensitivity of a simple baseline returns-to-experience model that excludes the class and school demographics of our preferred model. If the practice of sorting students to teachers biases our results, our estimates without these demographics should differ from our preferred model. Comparing results from this baseline model (column 1) to our preferred model (column 2) suggests that additional demographic controls have little impact on our estimates.

It is also possible that teachers' evaluation ratings reflect, in part, rater effects, or the harshness or leniency of individual principals. Although we do not have detailed data on evaluators, we can account for potential differences in rating norms across schools by including school fixed effects. This approach estimates the average returns to experience for individual teachers relative to all the other teachers who taught in the same school across our panel. As shown in columns 3, our results are quite consistent with the inclusion of school fixed effects.

We also refit our models in a restricted sample of teachers whom we can link to students with prior-year standardized test scores. Here, we examine results from models that include controls for the mean prior math and ELA achievement at the class and school level in year *t*. We find that our estimates are almost unchanged when we include additional controls for average measures of prior academic performance (columns 5 – 7).

Another potential concern is the subjectivity surrounding whether a teacher was formally evaluated in a given year. As we show in Table 1, the teacher and school characteristics of evaluated teachers differ systematically from those who are not evaluated, particularly among tenured teachers. If, for example, teachers who received lower ratings in the prior year are more likely to be rated in the following year, then our estimates could be biased upwards by the over-

[13] As first outlined by Murnane and Phillips (1981), the inclusion of teacher fixed effects removes any threats posed by student sorting across teachers.

representation of teachers with initially lower ratings, but also the potential for larger improvement on the TPAI-R rating system. At the same time, if teachers who are performing poorly in the current year are more likely to be evaluated, then our estimates could be biased downwards by the over-representation of teachers that are not improving.

We formally test whether the probability teachers were evaluated in a given year was related to their prior or current year performance. We do this by regressing a binary indicator for having an evaluation rating, $Evaluated_{jt}$, in a teacher-year dataset on several measures of teacher performance from either the current year *t* or prior year *t*-1, controlling for teacher experience and whether a teacher has tenure:[14]

$$Evaluated_{jt} = \beta_0 + \gamma(Performance_{jt}) + \ \xi * f\left(Exper_{jt}\right) + \alpha(Tenure_{jt}) + \epsilon_{jt} \qquad (6)$$

We fit a series of separate models where $Performance$ is operationalized as evaluation scores from the prior year or empirical Bayes value-added scores from either the prior or current year.

Across all models reported in Appendix Table A4, we find qualitatively small and statistically insignificant coefficients associated with measures of prior and current performance. The largest coefficient we find suggests that having a one standard deviation higher evaluation score in the prior year is associated with a 0.5 percentage point difference in the probability of being evaluated the following year (*p*=0.06). These findings suggest that although the probability of being rated was related to observable teacher characteristics, it was not driven by prior or current performance and is unlikely to bias the internal validity of our estimates substantially.

One final concern is the degree to which potential differential teacher attrition from the district – or the profession – limits the external validity of our findings. Like prior estimates of

[14] We use the same functional form for experience as in equation (4). Teacher tenure and experience are not collinear because experienced teachers who transfer districts become probationary teachers.

the returns to experience, our findings are generalizable to the sample of teachers that remain in the district. From a policy perspective, this group of actual teachers is the relevant target of inference. At the same time, if the probability that teachers remain in the district is related to their rate of improvement then our results may under- or over-state the "true" returns to experience profile for all teachers. We explore this possibility empirically by predicting the probability a teacher leaves the district at the end of each year based on their individual rate of improvement, conditional on experience, and report the results in Appendix Table A5.[15] We find small and statistically insignificant coefficients with inconsistent signs for the relationship between teachers' rates of growth and the probability they leave the district in seven of the eight years. These findings suggest that our results are unlikely to be biased substantially by any consistent dynamic attrition patterns and, thus, are generalizable beyond just our sample.

VI. Conclusion

Subjective performance ratings by principals provide a unique window into the specific skills and productivity growth of teachers across all grade levels and subject areas, not just those in tested grades and subjects. Analyzing these performance scores, we find further evidence of the important role that on-the-job learning plays in improving the productivity of the teacher workforce. Consistent with prior evidence based on test scores, we find, on average, rapid improvement in teacher performance early in the career and suggestive evidence of continued

[15] Specifically, we model a binary indicator for leaving the district, $Attrit_j^t$, as a function of teacher performance growth captured by empirical Bayes predictions of individual teachers' linear deviation from the average returns to experience profile from our multilevel random slopes model (equation 5). We model this relationship, conditional on experience, using the same functional form as in equation 4, as follows:

$$Attrit_j^t = \alpha_0 + \gamma(Growth_j) + \varpi * f\left(Exper_j^t\right) + \varepsilon_j^t$$

We fit this model separately for each year, t, to guard against any mechanical attenuation of our estimates in a teacher-year dataset given that teachers' individual growth rates, $Growth_j$, are time invariant.

growth through at least the first ten years on the job. The magnitude of these gains is large, eight-tenths of a standard deviation after ten years, suggesting that teachers' instructional practice and contributions to the school as a whole improve substantially as they gain experience. In contrast to prior theories about the sequential stages of skill development among teachers (e.g. Fuller & Brown, 1975), gains in the early stage of the career appear to be relatively consistent across the eight skills evaluated by principals in CMS. At the same time, we find substantial variation in productivity improvement rates across individual teachers and suggestive evidence of lower rates of average improvement among teachers working in middle schools.

From a policy perspective, these findings underscore the potential of human capital investments in the teacher labor force, and the perils of relying on a revolving door of inexperienced teachers to staff schools. For example, novice teachers are, on average, at the 31st percentile of the distribution of subjective performance ratings. We find that the typical novice teacher improves to the 62nd percentile after ten years on the job. Districts and schools that struggle to retain teachers into their mid-career fail to capitalize on the large improvements teachers make, on average, as they gain experience. At the same time, the large variation in the rates at which individual teachers are improving also points to the need for teacher preparation, induction, and professional development to provide effective supports to all teachers. Better understanding why some teachers improve more than others, and what programs and professional environments best promote learning on-the-job should be an important component of any efforts to improve the quality of the teacher workforce at scale.

Tables

Table 1: Teacher Sample Characteristics Across Samples

				Probationary Teachers			Tenured Teachers		
	Full Sample	Descriptive Sample	Analytic Sample	Has Performance Rating	Does Not Have Performance Rating	Difference	Has Performance Rating	Does Not Have Performance Rating	Difference
Female	0.79	0.78	0.78	0.78	0.77	0.01*	0.79	0.81	-0.02***
African American	0.24	0.27	0.28	0.27	0.22	0.05***	0.29	0.22	0.06***
Asian	0.01	0.01	0.01	0.01	0.02	-0.00	0.01	0.00	0.01***
Hispanic	0.03	0.04	0.03	0.04	0.04	0.00	0.03	0.01	0.02***
White	0.72	0.67	0.67	0.67	0.70	-0.02***	0.67	0.76	-0.09***
Teacher Experience	10.58	6.63	5.28	5.39	5.83	-0.44***	10.75	15.71	-4.96***
Holds Advanced Degree	0.33	0.27	0.25	0.25	0.24	0.01	0.35	0.40	-0.05***
National Board Certified	0.09	0.02	0.02	0.01	0.01	0.00	0.07	0.17	-0.11***
Average School Math Achievement (SDs)	-0.05	-0.10	-0.09	-0.11	-0.12	0.01	-0.10	0.01	-0.11***
Average School ELA Achievement (SDs)	-0.06	-0.11	-0.10	-0.11	-0.12	0.01*	-0.11	0.01	-0.12***
Share of African American Students in School	44.06	47.14	47.07	46.80	47.94	-1.14***	48.28	39.99	8.29***
Share of Hispanic Students in School	12.75	13.62	13.42	13.79	11.27	2.52***	13.04	12.51	0.53***
Share of ELL Students in School	11.48	12.39	12.17	12.55	9.94	2.61***	11.84	11.23	0.61***
Share of SPED Students in School	10.39	10.64	10.40	10.72	10.23	0.49***	10.39	10.23	0.16
Number of Students in School	1227.01	1206.94	1231.94	1201.98	1222.00	-20.02*	1223.41	1246.29	-22.88*
Mean Raw Performance Score		3.43	3.43	3.39			3.54		
Theta Performance Score		0.00	0.00	-0.08			0.26		
n (Teacher-years)	68,956	26,974	20,952	20,726	11,023	31,749	6,248	30,959	37,207
n (Unique Teachers)	17,398	9,821	6,558	8,502	7,011	13,441	5,046	7,841	8,981

Notes: p<0.001***, p<0.01**, p<0.05*. Average school math and English Language Arts (ELA) achievement are in student test score standard deviation units. Standardized evaluation scores are standardized within-year in a teacher-year-level data set. ELL = English Language Learner, SPED = Special Education.

Table 2: The Distribution of Raw Performance Scores by Domain

	Standard	Mean Score	Unsatisfactory	Below Standard	At Standard	Above Standard	n (teacher-years)
				Panel A: All Teachers			
1	Management of Instructional Time	3.39	0.06%	1.27%	58.24%	40.43%	26,958
2	Management of Student Behavior	3.42	0.16%	1.50%	54.70%	43.65%	26,942
3	Instructional Presentation	3.43	0.10%	1.40%	54.22%	44.28%	26,940
4	Instructional Monitoring of Student Performance	3.39	0.05%	0.76%	59.77%	39.41%	26,951
5	Instructional Feedback	3.39	0.03%	0.56%	59.71%	39.70%	26,933
6	Facilitating Instruction	3.41	0.09%	1.03%	56.45%	42.44%	26,926
7	Communicating within the Educational Environment	3.50	0.05%	0.65%	49.00%	50.30%	26,901
8	Performing Non-Instructional Duties	3.49	0.07%	0.59%	49.38%	49.96%	26,827
				Panel B: Novices			
1	Management of Instructional Time	3.19	0.12%	2.06%	76.29%	21.53%	3,400
2	Management of Student Behavior	3.19	0.26%	3.24%	73.68%	22.82%	3,400
3	Instructional Presentation	3.20	0.15%	2.18%	75.41%	22.26%	3,400
4	Instructional Monitoring of Student Performance	3.18	0.12%	1.29%	79.21%	19.38%	3,400
5	Instructional Feedback	3.19	0.06%	0.82%	79.41%	19.71%	3,399
6	Facilitating Instruction	3.20	0.06%	1.65%	76.32%	21.98%	3,399
7	Communicating within the Educational Environment	3.34	0.06%	0.91%	64.45%	34.58%	3,395
8	Performing Non-Instructional Duties	3.32	0.15%	0.47%	67.06%	32.32%	3,382

Notes: These analyses include 26,974 teacher-years and 3,401 novice teacher-years in 2001-02 to 2009-10. For novice teacher-years, paired t-tests indicate that the probability of scoring either Unsatisfactory or Below Standard in each domain is statistically significantly different at the 5 percent level from the probability of scoring Unsatisfactory or Below Standard in all other domains except for the following combinations: Instructional Monitoring of Student Performance and Communicating within the Educational Environment (p=0.05), Instructional Feedback and Communicating within the Educational Environment (p=0.07), Instructional Feedback and Performing Non-instructional Duties (p=0.26), Communicating within the Educational Environment and Performing Non-Instructional Duties (p=0.49). The probability of scoring Above Standard in each domain is statistically significantly different at the 5 percent level from the probability of scoring Above Standard in all other domains except for the following combinations: Instructional Feedback and Instructional Monitoring of Student Performance (p=0.18), Performing Non-instructional Duties and Communicating within the Educational Environment (p=0.20).

Table 3: Correlations between Theta Performance Scores and Value-added Measures

Math		English Language Arts	
Unadjusted	Adjusted	Unadjusted	Adjusted
Panel A. Current Year Value-Added Scores			
Full Sample			
0.227*	0.287*	0.125*	0.193*
(0.015)	(0.019)	(0.016)	(0.024)
n = 4,236		n = 4,127	
Excluding the Top and Bottom Quintile of Standardized Evaluation Scores			
0.160*	0.203*	0.096*	0.149*
(0.019)	(0.024)	(0.020)	(0.031)
n = 2,566		n = 2,565	
Panel B. Prior-Year Value-Added Scores			
Full Sample			
0.230*	0.292*	0.121*	0.187*
(0.019)	(0.024)	(0.021)	(0.032)
n = 2,509		n = 2,356	
Excluding the Top and Bottom Quintile of Standardized Evaluation Scores			
0.118*	0.150*	0.107*	0.166*
(0.026)	(0.033)	(0.026)	(0.041)
n = 1,501		n = 1,442	

Notes: Significant at 5% level. Standard errors are calculated using a bootstrap method with 1,000 iterations. We adjusted for estimation error in value-added measure following Jacob & Lefgren (2008), where reliability of math and ELA value-added measures are 0.766 and 0.513, respectively. We adjusted for measurement error in theta performance scores, where the reliability of theta scores is 0.813.

Table 4: The Relationship between Prior Theta Performance Scores and Student Achievement Gains

	Panel A: Math Test Scores at time *t*				
	(1)	(2)	(3)	(4)	(5)
Theta Performance Score at time *t-1*	0.031***		0.011**		
	(0.005)		(0.004)		
Math Value-Added at time *t-1*		0.094***	0.092***		0.094***
		(0.004)	(0.004)		(0.004)
Theta Performance Gain Score (*t-1* to *t*)				0.013**	0.009*
				(0.004)	(0.004)
n (Student-years)	74,422	74,422	74,422	74,422	74,422
n (Teacher-years)	2,100	2,100	2,100	2,100	2,100
	Panel B: English Languag Arts Test Scores at time *t*				
Theta Performance Score at time *t-1*	0.010**		0.007*		
	(0.003)		(0.003)		
ELA Value-Added at time *t-1*		0.025***	0.025***		0.025***
		(0.003)	(0.003)		(0.003)
Theta Performance Gain Score (*t-1* to *t*)				0.004	0.003
				(0.003)	(0.003)
n (Student-years)	64,021	64,021	64,021	64,021	64,021
n (Teacher-years)	1,981	1,981	1,981	1,981	1,981

Notes: p<0.001***, p<0.01**, p<0.05*, +p<0.10. Robust standard errors, clustered at the teacher-level, reported in parentheses. Value-added measures are empirical Bayes shrunken estimates that we standardize within-year in a teacher-year-level data set so that a standard deviation differences is comparable across both theta performances scores and value-added scores. Theta performance gain scores are calculated by subtracting the theta score in time *t-1* from that in time *t* .

Table 5: Variance Decomposition of Theta Performance Scores

	Variance	Proportion of Total Variance
Between School (school random effects)	0.152***	0.148
	(0.018)	
Between Teachers (teacher random effects)	0.324***	0.315
	(0.009)	
Within Teachers (residual)	0.553***	0.537
	(0.006)	
n (Teacher-years)	26,974	

Notes: ***p<.001.

Table 6: Individual Heterogeneity in Returns to Experience from a Multilevel Model with Random Intercepts and Slopes

	(1)	(2)
Teacher intercepts (SD)	0.547***	0.526***
	(0.015)	(0.019)
Teacher slopes on linear term (SD)	0.060***	0.193***
	(0.005)	(0.015)
Teacher slopes on quad. term (SD)		0.017***
		(0.002)
Residual (SD)	0.714***	0.693***
	(0.004)	(0.005)
Correlation (intercepts, slope on linear term)	-0.224**	-0.091
	(0.079)	(0.083)
Correlation (intercepts, slope on quad. term)		-0.322*
		(0.132)
Correlation (slopes on linear and quad. terms)		-0.866***
		(0.026)
n(Teacher-years)	20,952	20,952

Notes: ***p<.001, **p<0.01, *p<.05.

Table 7: Models Examining Hiring and Retention Patterns by Performance Ratings Across Tested and Non-tested Grades and Subjects

Sample	Elementary and middle school teachers†	Teaches in **tested** grade/subject in *t*		Teaches in **non-tested** grade/subject in *t*		Teachers who transfer schools between *t* and *t+1*
Outcome	Return to same school in *t+1*	Return to same school & teach in **tested** grade/subject in *t+1*	Return to same school & teach in **non-tested** grade/subject in *t+1*	Return to same school and teach in **tested** grade/subject in *t+1*	Return to same school and teach in **non-tested** grade/subject in *t+1*	Teach in **tested** grade/subject in *t+1*
	(1)	(2)	(3)	(4)	(5)	(6)
Theta Score	0.029***	0.044***	-0.014***	-0.001	0.030***	0.011
	(0.004)	(0.006)	(0.003)	(0.003)	(0.006)	(0.016)
n (Teacher-years)	11,693	6,100	6,100	5,593	5,593	1,112

Notes: p<0.001***, p<0.01**, p<0.05*, +p<0.10. Robust standard errors, clustered at the teacher-level, reported in parentheses. All linear probability models include fixed effects for teacher experience.

†This analysis includes elementary and middle school teachers in 2001-02 to 2008-09 that contribute to the returns to experience profiles depicted in Figure 7.

Table 8: Teacher Turnover Rates Across School Levels

	Annual	3-year
Elementary	23.21%	49.55%
n (Teacher-years)	29,779	21,279
Middle	29.10%	59.64%
n (Teacher-years)	12,506	9,301
High	23.56%	49.56%
n (Teacher-years)	16,567	11,914

Notes: Annual (3-year) turnover is defined as not returning to teach at the same school in the following year (3 years later). Both annual and 3-year turnover rates for middle school teachers are statistically significantly different than those of elementary and high school.

Figures

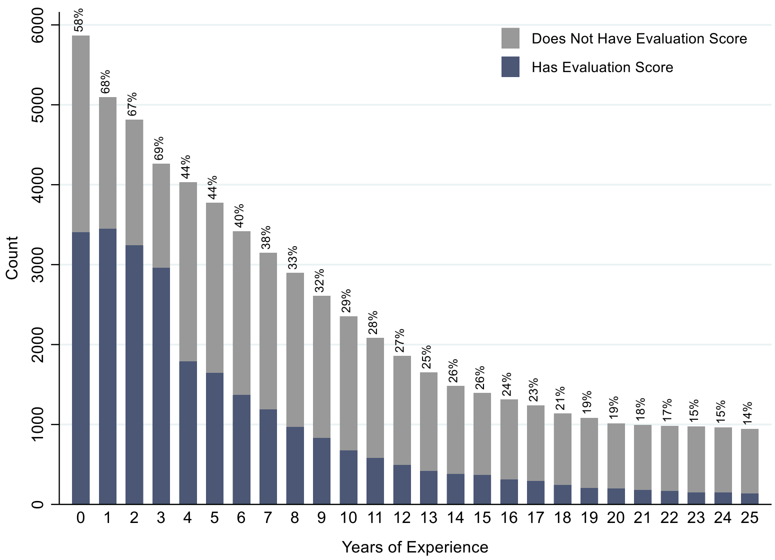


Figure 1. Count of teachers by years of experience

Notes: This figure includes 61,331 teacher-years with 25 or fewer years of teaching experience in 2001-02 through 2009-10. The percent above each bar indicates the percent of teacher-years with evaluation scores for the indicated level of experience.

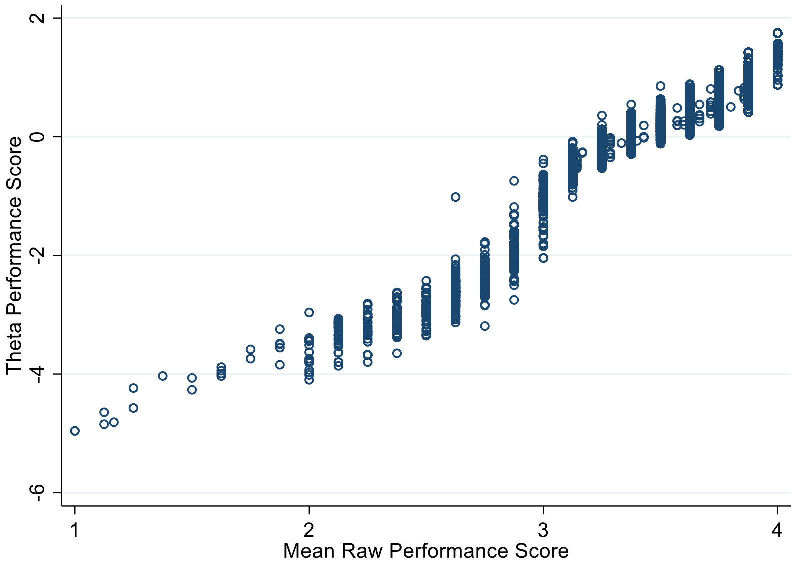


Figure 2. Scatterplot of the relationship between teachers' mean raw performance score and theta performance scores.

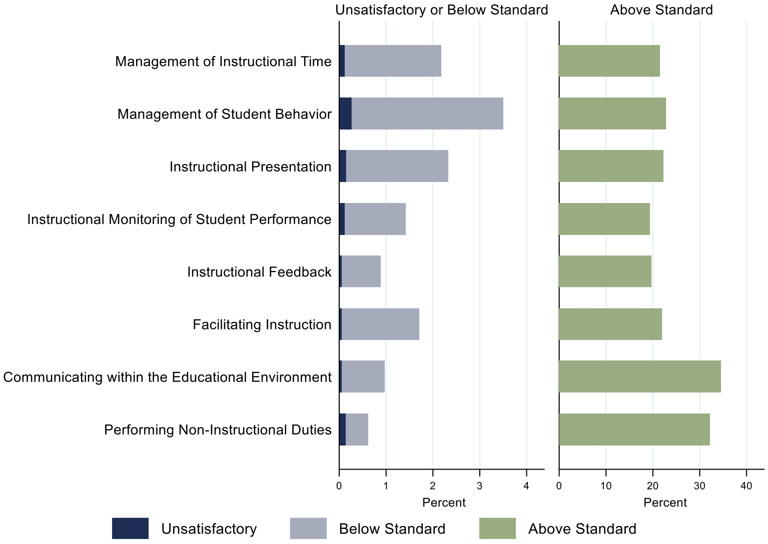


Figure 3: Distribution of raw performance ratings for novice teachers by domain
Notes: This analysis includes 3,401 novice teacher-years in 2001-02 to 2009-10.

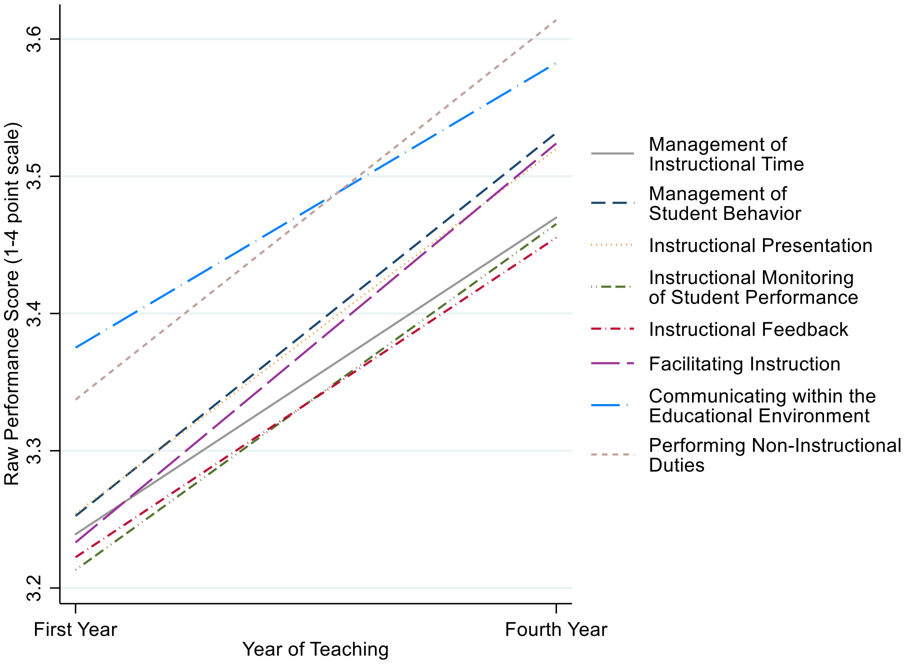


Figure 4. Changes in mean performance scores across a balanced sample of teachers in their first and fourth years of the job.
Notes: This analysis includes 1,296 teachers who have evaluation scores in their first year of teaching and their fourth year of teaching.

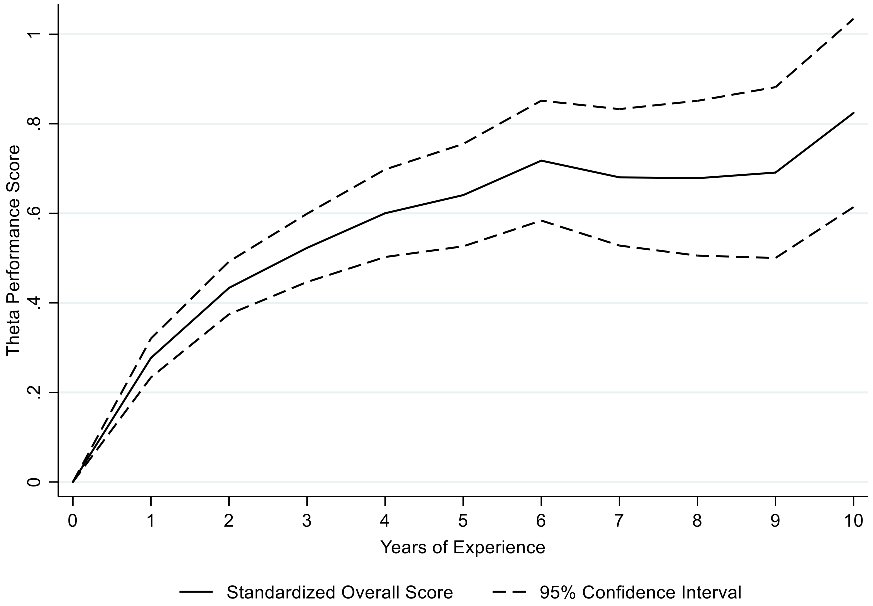
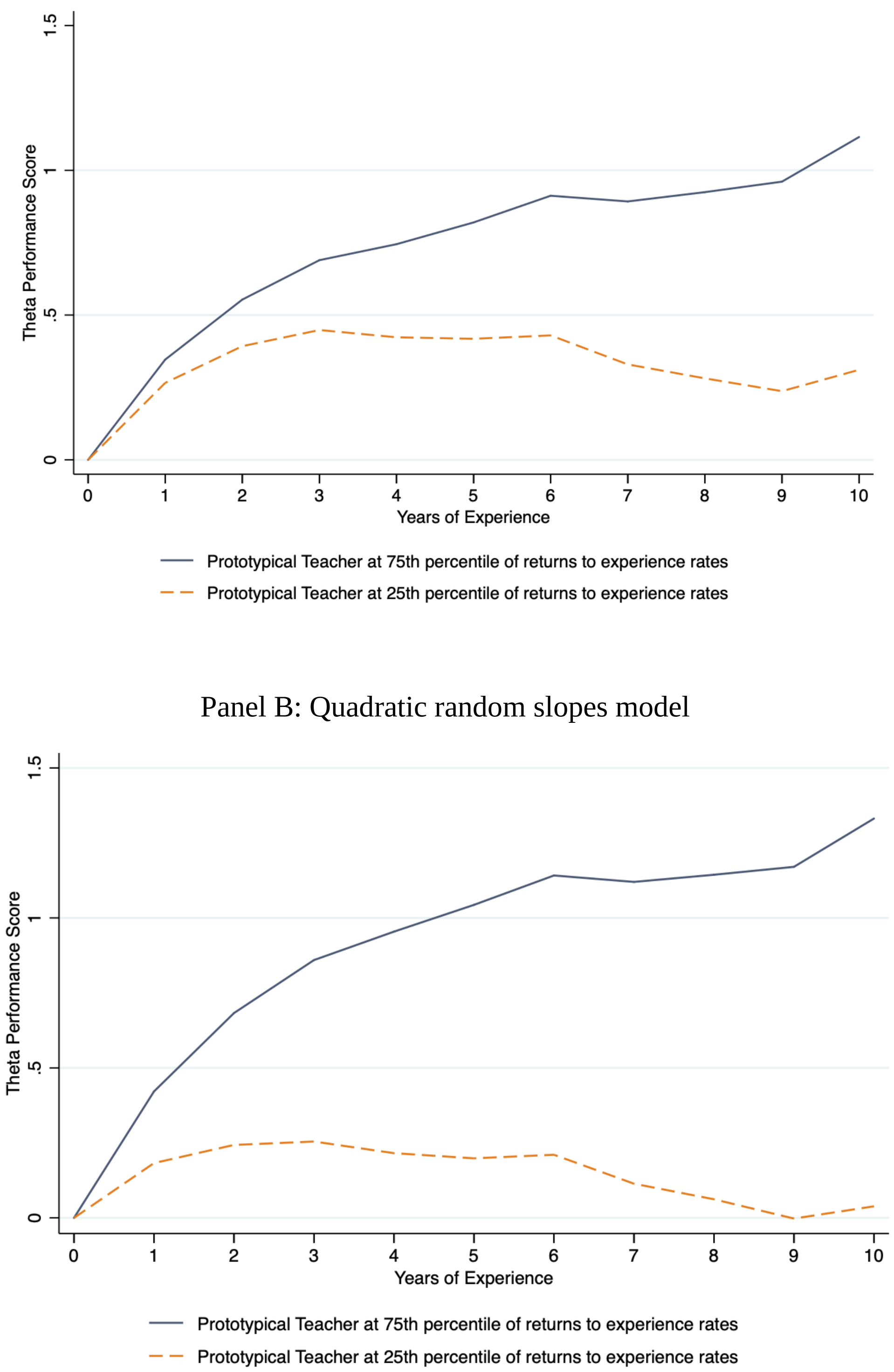


Figure 5. Within-teacher returns to experience

Notes: This analysis includes 20,952 teacher-years from 6,558 unique teachers in 2001-02 to 2009-10. Figure depicts results from estimates reported in Table 5 column 2.



Panel A: Linear random slopes model

Panel B: Quadratic random slopes model

Figure 6. Within-teacher returns to experience for prototypical teachers at the 75th and 25th percentile of returns to experience rates.

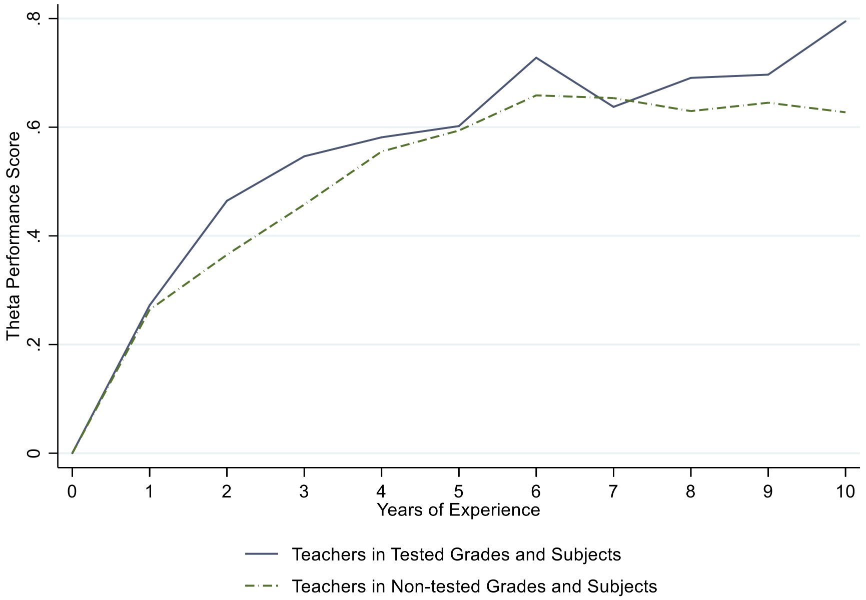


Figure 7. Within-teacher returns to experience for teachers in tested and non-tested grades and subjects.

Notes: This analysis includes 7,030 teacher-years from teachers in tested grades and subjects and 6,469 teacher-years from teachers in non-tested grades and subjects. A joint significance test indicates that the differences between the coefficients on the experience indicators in years 1 through 10 for teachers in tested and non-tested subjects is not statistically significant (p = 0.63).

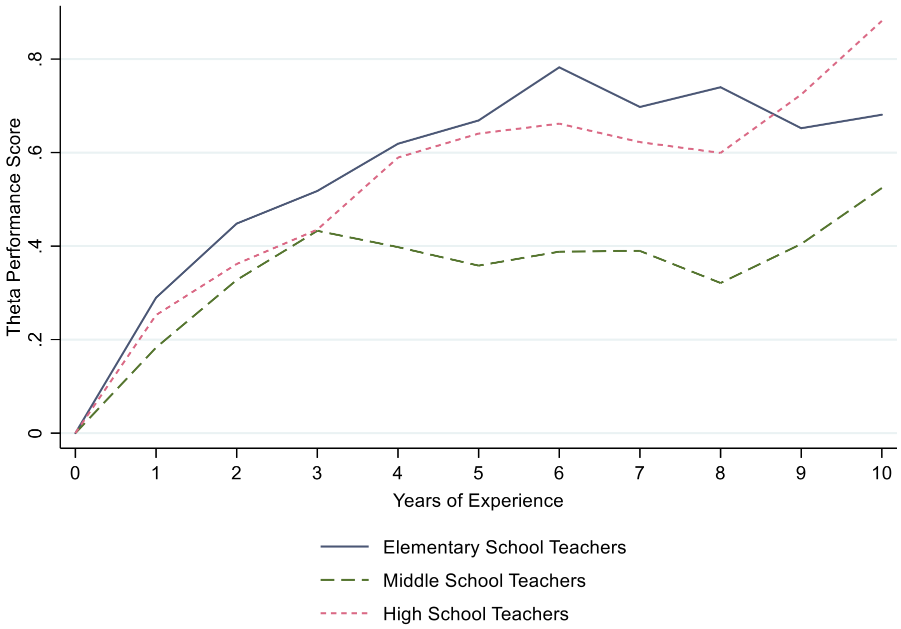


Figure 8. Within-teacher returns to experience for elementary, middle, and high school teachers
Notes: This analysis includes 9,391 elementary school teacher-years, 4,775 middle school teacher-years, and 5,721 high school teacher-years.

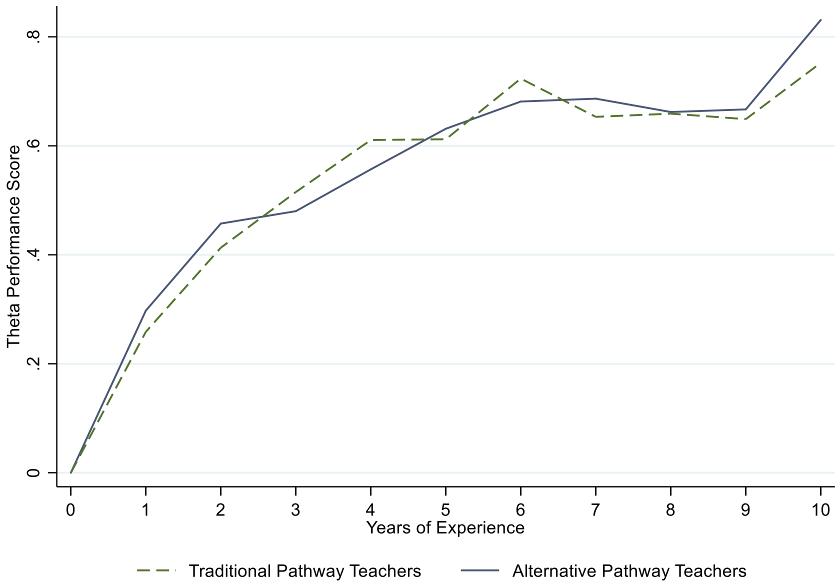


Figure 9. Within-teacher returns to experience for traditional and alternative pathway teachers
Notes: This analysis includes 13,698 teacher-years from traditional pathway teachers and 6,963 teacher-years from alternative pathway teachers.

References

Atteberry, A., Loeb, S., & Wyckoff, J. (2015). Do First Impressions Matter? Predicting Early Career Teacher Effectiveness. AERA Open, 1(4).

Boyd, D., Grossman, P., Lankford, H., Loeb, S., & Wyckoff, J. (2006). How Changes in Entry Requirements Alter the Teacher Workforce and Affect Student Achievement.

Boyd, D., Lankford, H., Loeb, S., Rockoff, J., Wyckoff, J. (2008). The Narrowing Gap in New York City Teacher Qualifications and its Implications for Student Achievement in High-Poverty Schools. Journal of Policy Analysis and Management, 27(4), 793–818.

Boyd, D., Lankford, H., Loeb, S., Ronfeldt, M., & Wyckoff, J. (2011). The Role of Teacher Quality in Retention and Hiring: Using Applications to Transfer to Uncover Preferences Of Teachers And Schools. Journal of Policy Analysis and Management, 30(1), 88-110.

Brighouse, H., Ladd, H. F., Loeb, S., & Swift, A. (2018). Educational Goods: Values, Evidence, and Decision-Making. University of Chicago Press.

Campbell, S. L., & Ronfeldt, M. (2018). Observational Evaluation of Teachers: Measuring More Than We Bargained for?. American Educational Research Journal

Cohen, J., & Goldhaber, D. (2016). Building a More Complete Understanding of Teacher Evaluation Using Classroom Observations. Educational Researcher, 45(6), 378-387.

Cohen-Vogel, L. (2011). “Staffing to The Test”: Are Today’s School Personnel Practices Evidence Based?. Educational Evaluation and Policy Analysis, 33(4), 483-505.

Cohodes, S., Setren, E., & Walters, C. R. (2019). Can Successful Schools Replicate? Scaling Up Boston’s Charter School Sector. *National Bureau of Economic Research Working Paper No. 25796.*

Donaldson, M. L., & Papay, J. P. (2014). Teacher Evaluation for Accountability And Development. In Handbook of Research in Education Finance and Policy (pp. 190-209). Routledge.

Feng, L., & Sass, T. R. (2018). The Impact of Incentives to Recruit and Retain Teachers in “Hard-to-Staff” Subjects. Journal of Policy Analysis and Management, 37(1), 112-135.

Fuller, F.F. (85). Brown. O.H. (1975). Becoming a Teacher. Teacher Education. 74th. Yearbook of the National Society for the Study of Education. Part II, 25-52. Chicago, IL: University of Chicago Press.

Gates, B., 2009. Mosquitos, Malaria, and Education. TED Talk. Available at https://www.ted.com/talks/bill_gates_unplugged/transcript.

Garet, M. S., Cronen, S., Eaton, M., Kurki, A., Ludwig, M., Jones, W., ... & Zhu, P. (2008). The

Impact of Two Professional Development Interventions on Early Reading Instruction and Achievement. NCEE 2008-4030. National Center for Education Evaluation and Regional Assistance.

Gershenson, S. (2016). Linking Teacher Quality, Student Attendance, and Student Achievement. Education Finance and Policy, *11*(2), 125-149.

Goldhaber, D., Grout, C., & Huntington-Klein, N. (2017). Screen Twice, Cut Once: Assessing the Predictive Validity of Applicant Selection Tools. Education Finance and Policy, *12*(2), 197-223.

Goldhaber, D., & Hansen, M. (2010). Using Performance on the Job to Inform Teacher Tenure Decisions. American Economic Review Papers & Proceedings, *100*(2), 250-55.

Grissom, J. A., & Loeb, S. (2017). Assessing Principals' Assessments: Subjective Evaluations of Teacher Effectiveness in Low-and High-Stakes Environments. Education Finance and Policy, *12*(3), 369-395.

Grissom, J. A., Kalogrides, D., & Loeb, S. (2017). Strategic Staffing? How Performance Pressures Affect the Distribution of Teachers Within Schools and Resulting Student Achievement. American Educational Research Journal, *54*(6), 1079-1116.

Harris, D. N., & Sass, T. R. (2011). Teacher Training, Teacher Quality and Student Achievement. Journal of Public Economics, *95*(7-8), 798-812.

Harris, D. N., & Sass, T. R. (2014). Skills, Productivity and the Evaluation of Teacher Performance. Economics of Education Review, *40*, 183-204.

Hanushek, E. A., & Rivkin, S. G. (2010). Generalizations About Using Value-Added Measures of Teacher Quality. American Economic Review Papers and Proceedings, 100(2), 267-71.

Hill, H. C. (2007). Learning in the Teaching Workforce. The Future of Children, 111-127.

Jackson, C. K. (2018). What Do Test Scores Miss? The Importance of Teacher Effects on Non-Test Score Outcomes. Journal of Political Economy, *126*(5), 2072-2107

Jackson, C. K., Rockoff, J. E., & Staiger, D. O. (2014). Teacher Effects and Teacher-Related Policies. Annual Review of Economics, *6*(1), 801-825.

Jacob, A., & McGovern, K. (2015). The Mirage: Confronting the Hard Truth about Our Quest for Teacher Development. TNTP.

Jacob, B. A., & Lefgren, L. (2004). The Impact of Teacher Training on Student Achievement: Quasi-experimental Evidence from School Reform Efforts in Chicago. Journal of Human Resources, 39(1), 50-79.

Jacob, B. A., & Lefgren, L. (2008). Can Principals Identify Effective Teachers? Evidence on Subjective Performance Evaluation in Education. Journal of Labor Economics, *26*(1), 101-136.

Jacob, B., Rockoff, J. E., Taylor, E. S., Lindy, B., & Rosen, R. (2016). Teacher Applicant Hiring and Teacher Performance: Evidence from DC Public Schools (No. w22054). National Bureau of Economic Research.

Jacob, B. A., & Walsh, E. (2011). What's in a Rating? Economics of Education Review, 30(3), 434-448.

Johnson, S. M., Reinhorn, S. K., Charner-Laird, M., Kraft, M. A., Ng, M., & Papay, J. P. (2014). Ready to Lead, but How? Teachers' Experiences in High-Poverty Urban Schools. Teachers College Record, 116(10).

Kane, T. J., Rockoff, J. E., & Staiger, D. O. (2008). What Does Certification Tell Us About Teacher Effectiveness? Evidence from New York City. Economics of Education Review, *27*(6), 615-631.

Kane, T. J., & Staiger, D. O. (2012). Gathering Feedback for Teaching: Combining High-Quality Observations with Student Surveys and Achievement Gains. Research Paper. MET Project. Bill & Melinda Gates Foundation.

Kane, T. J., Taylor, E. S., Tyler, J. H., & Wooten, A. L. (2011). Identifying Effective Classroom Practices Using Student Achievement Data. Journal of Human Resources, 46(3), 587-613.

Koedel, C., Mihaly, K., & Rockoff, J. E. (2015). Value-Added Modeling: A Review. Economics of Education Review, *47*, 180-195.

Kraft. M.A. (2018). Federal Efforts to Improve Teacher Quality. In Hess R. & McShane, M. (Editors). Bush-Obama School Reform: Lessons Learned. Harvard Education Press.

Kraft, M.A. & Gilmour, A. (2017). Revisiting the Widget Effect: Teacher Evaluation Reforms and Distribution of Teacher Effectiveness Ratings. Educational Researcher, 46(5), 234-249.

Kraft, M.A. & Papay, J.P. (2014). Can Professional Environments in Schools Promote Teacher Development? Explaining Heterogeneity in Returns to Teaching Experience. Educational Evaluation and Policy Analysis. 36(4), 476-500.

Ladd, H. F., & Sorensen, L. C. (2017). Returns to Teacher Experience: Student Achievement and Motivation in Middle School. Education Finance and Policy, *12*(2), 241-279.

Marinell, W. H., & Coca, V. M. (2013). " Who Stays and Who Leaves?" Findings from a Three-Part Study of Teacher Turnover in NYC Middle Schools. *The Research Alliance for New York City Schools*.

Moore, C. M. (2012). The Role of School Environment in Teacher Dissatisfaction Among US Public School Teachers. Sage Open, 2(1),

Murnane, R. J., & Phillips, B. R. (1981). Learning by Doing, Vintage, and Selection: Three Pieces of the Puzzle Relating Teaching Experience and Teaching Performance. Economics of Education Review, 1(4), 453-465.

National Council on Teacher Quality (NCTQ). April 2016. State-by-State Evaluation Timeline Briefs. Washington, DC: NCTQ.

Ost, B. (2014). How do Teachers Improve? The Relative Importance of Specific and General Human Capital. American Economic Journal: Applied Economics, 6(2), 127-51.

Papay, J. P., West, M. R., Fullerton, J. B., & Kane, T. J. (2012). Does an Urban Teacher Residency Increase Student Achievement? Early Evidence from Boston. Educational Evaluation and Policy Analysis, 34(4), 413-434.

Papay, J. P., & Kraft, M. A. (2015). Productivity Returns to Experience the Teacher Labor Market: Methodological Challenges and New Evidence on Long-Term Career Improvement. Journal of Public Economics, 130, 105-119.

Petek, N., & Pope, N. (2016). The Multidimensional Impact of Teachers on Students. University of Chicago Working Paper.

Raju, N. S., Price, L. R., Oshima, T. C., & Nering, M. L. (2007). Standardized Conditional SEM: A Case for Conditional Reliability. Applied Psychological Measurement, 31(3), 169-180.

Rice, J. K. (2013). Learning from Experience? Evidence on The Impact and Distribution of Teacher Experience and the Implications for Teacher Policy. Education Finance and Policy, 8(3), 332-348.

Rivkin, S. G., Hanushek, E. A., & Kain, J. F. (2005). Teachers, Schools, and Academic Achievement. Econometrica, *73*(2), 417-458.

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

Rockoff, J. E., Jacob, B. A., Kane, T. J., & Staiger, D. O. (2011). Can You Recognize an Effective Teacher When You Recruit One? Education Finance and Policy, 6(1), 43-74.

Rockoff, J. E., Staiger, D. O., Kane, T. J., & Taylor, E. S. (2012). Information and Employee Evaluation: Evidence from a Randomized Intervention in Public Schools. American Economic Review, 102(7), 3184-3213.

Ronfeldt, M., Loeb, S., & Wyckoff, J. (2013). How teacher turnover harms student achievement. *American Educational Research Journal*, *50*(1), 4-36.

Samejima, F. (1969). Estimation of Latent Ability Using a Response Pattern of Graded Scores (Psychometric Monograph No. 17). Richmond, VA: Psychometric Society. Retrieved from http://www.psychometrika.org/journal/online/MN17.pdf

Sartain, L., Stoelinga, S. R., & Brown, E. R. (2011). Rethinking Teacher Evaluation in Chicago: Lessons Learned from Classroom Observations, Principal-Teacher Conferences, and District Implementation. Research Report. Consortium on Chicago School Research. 1313 East 60th Street, Chicago, IL 60637

Steele, J. L., Murnane, R. J., & Willett, J. B. (2010). Do Financial Incentives Help Low-Performing Schools Attract and Keep Academically Talented Teachers? Evidence From California. Journal of Policy Analysis and Management, 29(3), 451-478.

Steinberg, M. P., & Donaldson, M. L. (2016). The New Educational Accountability: Understanding the Landscape of Teacher Evaluation in the Post-NCLB Era. Education Finance and Policy, 11(3), 340-359.

Steinberg, M. P., & Garrett, R. (2016). Classroom Composition and Measured Teacher Performance: What Do Teacher Observation Scores Really Measure? Educational Evaluation and Policy Analysis, 38(2), 293-317.

Weisberg, D., Sexton, S., Mulhern, J., Keeling, D., Schunck, J., Palcisco, A., & Morgan, K. (2009). The Widget Effect: Our National Failure to Acknowledge and Act on Differences in Teacher Effectiveness. New Teacher Project.

Whitehurst, G., Chingos, M. M., & Lindquist, K. M. (2014). Evaluating Teachers with Classroom Observations. Brown Center on Education Policy: Brookings Institute.

Winters, M. A., and Cowen, J. M. (2013). Would a Value-Added System of Retention Improve the Distribution of Teacher Quality? A Simulation of Alternative Policies. Journal of Policy Analysis and Management 32(3), 634-654.

Appendix A

Table A1: Changes in Raw Performance Scores among a Balanced Sample of Teachers in their First and Fourth Years of Teaching

	Raw Performance Score (1-4 point scale)				Percent of Teachers		
	First year of teaching	Fourth year of teaching	Gain	(standard deviation units)	With Decreasing Scores	Receiving Same Score	With Increasing Scores
Management of Instructional Time	3.24	3.47	0.23	0.45	9.57	58.10	32.33
Management of Student Behavior	3.25	3.53	0.28	0.52	8.87	54.78	36.34
Instructional Presentation	3.25	3.52	0.27	0.50	10.11	53.63	36.27
Instructional Monitoring of Student Performance	3.21	3.47	0.25	0.50	9.03	56.79	34.18
Instructional Feedback	3.22	3.46	0.23	0.46	9.95	57.10	32.95
Facilitating Instruction	3.23	3.52	0.29	0.56	9.95	51.31	38.73
Communicating within the Educational Environment	3.38	3.58	0.21	0.40	12.96	53.86	33.18
Performing Non-Instructional Duties	3.34	3.61	0.28	0.54	12.11	49.31	38.58
Mean Raw Performance Score	3.27	3.52	0.25	0.64	19.44	15.90	64.66

Notes: Gains reported in standard deviation units are based on performance scores standardized across the full panel. Significance tests from seemingly unrelated regressions indicate that the gain in Facilitating Instruction is statistically significantly different from the gains in Management of Instructional Time, Instructional Monitoring of Student Performance, Instructional Feedback, and Communicating within the Educational Environment at the 5 percent level. The gain in Communicating with the Educational Environment is also statistically significantly different from the gains in Management of Student Behavior, Instructional Presentation, Instructional Monitoring of Student Performance, and Performing Non-Instructional Duties at the 5 percent level. n=1,296

Table A2: Within-Teacher Returns to Experience for Theta Performance Scores

	Theta Performance Score						Mean Raw Score	Mean Raw Score (Std.)
	Full Sample			Sample with Mean Prior Achievement			Full Sample	Full Sample
	(1)	(2)	(3)	(5)	(6)	(7)	(8)	(9)
Experience = 1 Year	0.278***	0.277***	0.268***	0.289***	0.287***	0.284***	0.114***	0.289***
	(0.022)	(0.022)	(0.022)	(0.029)	(0.029)	(0.029)	(0.009)	(0.022)
Experience = 2 Years	0.437***	0.434***	0.428***	0.445***	0.441***	0.439***	0.177***	0.447***
	(0.030)	(0.030)	(0.029)	(0.038)	(0.038)	(0.037)	(0.012)	(0.030)
Experience = 3 Years	0.527***	0.523***	0.514***	0.540***	0.535***	0.533***	0.212***	0.536***
	(0.039)	(0.039)	(0.038)	(0.048)	(0.048)	(0.047)	(0.015)	(0.038)
Experience = 4 Years	0.605***	0.600***	0.585***	0.600***	0.593***	0.584***	0.242***	0.611***
	(0.050)	(0.050)	(0.049)	(0.060)	(0.060)	(0.059)	(0.019)	(0.049)
Experience = 5 Years	0.645***	0.641***	0.624***	0.610***	0.602***	0.600***	0.259***	0.655***
	(0.059)	(0.058)	(0.057)	(0.070)	(0.070)	(0.069)	(0.023)	(0.057)
Experience = 6 Years	0.726***	0.718***	0.688***	0.688***	0.682***	0.666***	0.293***	0.741***
	(0.068)	(0.068)	(0.066)	(0.081)	(0.081)	(0.080)	(0.027)	(0.067)
Experience = 7 Years	0.688***	0.680***	0.661***	0.665***	0.661***	0.649***	0.272***	0.687***
	(0.078)	(0.078)	(0.075)	(0.092)	(0.091)	(0.090)	(0.030)	(0.076)
Experience = 8 Years	0.690***	0.678***	0.656***	0.674***	0.671***	0.657***	0.270***	0.682***
	(0.088)	(0.088)	(0.085)	(0.104)	(0.103)	(0.101)	(0.034)	(0.086)
Experience = 9 Years	0.693***	0.691***	0.661***	0.704***	0.701***	0.698***	0.277***	0.701***
	(0.097)	(0.097)	(0.093)	(0.114)	(0.113)	(0.111)	(0.038)	(0.095)
Experience = 10 Years	0.830***	0.824***	0.794***	0.834***	0.832***	0.832***	0.326***	0.823***
	(0.107)	(0.107)	(0.103)	(0.126)	(0.125)	(0.122)	(0.041)	(0.105)
School Demographics		Y	Y	Y	Y	Y	Y	Y
Class Demographics		Y	Y	Y	Y	Y	Y	Y
Mean Prior Achievement					Y	Y		
School Fixed Effects			Y			Y		
n (Teacher-years)	20,952	20,952	20,948	12,883	12,883	12,877	20,952	20,952

Notes: p<0.001***, p<0.01**, p<0.05*, p<0.10+. These estimates include teachers across all grades and subjects from Kindergarten to 12th grade who are linked to at least five students and have 25 or fewer years of teaching experience. Robust standard errors clustered at the teacher-level reported in parentheses. Each model includes teacher fixed effects, modal grade-by-year fixed effects and additional indicators for having 11-15 years of experience, 16-20 years of experience, and 21-25 years of experience. The omitted experience category in all models is zero years of experience.

Table A3: Within-Teacher Returns to Experience for Theta Performance Scores across Subgroups of Teachers

	Full Sample	Tested Grades and Subjects	Non-tested Grades and Subjects	Elementary Teachers	Middle School Teachers	High School Teachers	Traditional Pathway Teachers	Alternative Pathway Teachers
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Experience = 1 Year	0.277***	0.272***	0.264***	0.290***	0.183***	0.253***	0.259***	0.298***
	(0.022)	(0.036)	(0.039)	(0.032)	(0.046)	(0.049)	(0.027)	(0.039)
Experience = 2 Years	0.434***	0.465***	0.365***	0.448***	0.328***	0.362***	0.413***	0.457***
	(0.030)	(0.051)	(0.056)	(0.048)	(0.058)	(0.061)	(0.039)	(0.050)
Experience = 3 Years	0.523***	0.546***	0.458***	0.518***	0.433***	0.435***	0.515***	0.480***
	(0.039)	(0.066)	(0.074)	(0.064)	(0.073)	(0.077)	(0.052)	(0.064)
Experience = 4 Years	0.600***	0.581***	0.555***	0.619***	0.398***	0.589***	0.611***	0.557***
	(0.050)	(0.086)	(0.097)	(0.084)	(0.094)	(0.092)	(0.069)	(0.075)
Experience = 5 Years	0.641***	0.602***	0.594***	0.669***	0.358***	0.641***	0.612***	0.631***
	(0.058)	(0.100)	(0.115)	(0.099)	(0.107)	(0.109)	(0.082)	(0.087)
Experience = 6 Years	0.718***	0.728***	0.659***	0.782***	0.388**	0.662***	0.723***	0.681***
	(0.068)	(0.118)	(0.134)	(0.119)	(0.125)	(0.123)	(0.097)	(0.101)
Experience = 7 Years	0.680***	0.637***	0.654***	0.698***	0.390**	0.622***	0.653***	0.687***
	(0.078)	(0.136)	(0.153)	(0.139)	(0.139)	(0.138)	(0.111)	(0.115)
Experience = 8 Years	0.678***	0.691***	0.630***	0.740***	0.321*	0.599***	0.659***	0.662***
	(0.088)	(0.152)	(0.178)	(0.157)	(0.159)	(0.157)	(0.125)	(0.130)
Experience = 9 Years	0.691***	0.697***	0.645***	0.652***	0.405*	0.724***	0.649***	0.667***
	(0.097)	(0.168)	(0.194)	(0.175)	(0.175)	(0.169)	(0.139)	(0.144)
Experience = 10 Years	0.824***	0.795***	0.627**	0.681***	0.524**	0.882***	0.752***	0.831***
	(0.107)	(0.187)	(0.216)	(0.195)	(0.192)	(0.186)	(0.155)	(0.155)
n (Teacher-years)	20,952	7,030	6,469	9,391	4,775	5,721	13,698	6,963

Notes: p<0.001***, p<0.01**, p<0.05*, p<0.10+. Robust standard errors clustered at the teacher-level reported in parentheses. Each model includes teacher fixed-effects, modal-grade-by-year fixed effects, school and student demographic characteristics, and additional indicators for having 11-15 years of experience, 16-20 years of experience, and 21-25 years of experience. The omitted experience category in all models is zero years of experience. The full sample includes teacher-years in which teachers have 25 or fewer years of teaching experience.

Table A4: The Relationship Between Receiving a Performance Evaluation and Prior and Current Measures of Performance

	(1)	(2)	(3)	(4)	(5)
Prior Year Theta Performance Score	-0.005+				
	(0.003)				
Prior Year Math Value-Added		-0.001			
		(0.006)			
Prior Year ELA Value-Added			0.001		
			(0.006)		
Current Year Math Value-Added				0.004	
				(0.005)	
Current Year ELA Value-Added					0.000
					(0.005)
n (Teacher-years)	20,138	6,607	6,107	9,693	9,336

Notes: p<0.001***, p<0.01**, p<0.05*, p<0.10+. Robust standard errors, clustered at the teacher-level, reported in parentheses. All models include an indicator for tenure as well as an indicator for each year of experience from 1 year to 10 years, having 11-15 years of experience, 16-20 years of experience, and 21-25 years of experience. Value-added measures are empirical Bayes shrunken estimates that we standardize within-year in a teacher-year-level data set so that a standard deviation differences is comparable across both theta performances scores and value-added scores.

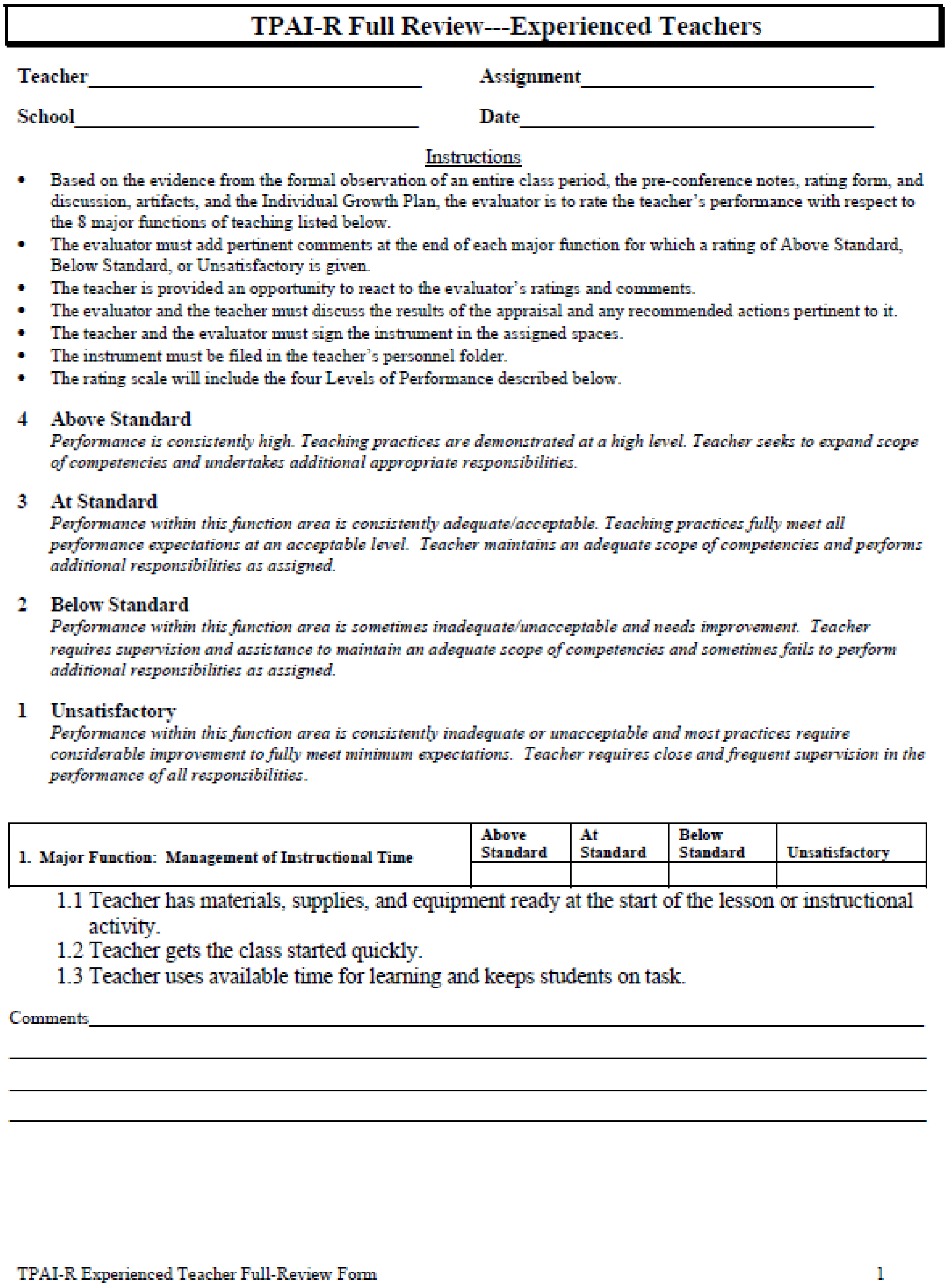
Table A5: The Relationship Between Attrition and Growth in Performance

	Teacher is Not Teaching in *t+1*							
	2002	2003	2004	2005	2006	2007	2008	2009
Teacher Random Slope	-0.005	0.001	-0.000	-0.000	0.003	0.000	0.001	-0.031***
	(0.004)	(0.004)	(0.004)	(0.004)	(0.003)	(0.003)	(0.002)	(0.004)
n (Teacher-years)	1,191	1,633	1,688	2,173	2,434	2,823	2,471	3,677

Notes: p<0.001***. Robust standard errors, clustered at the teacher-level, reported in parentheses. All models include an indicator for each year of experience from 1 year to 10 years, having 11-15 years of experience, 16-20 years of experience, and 21-25 years of experience. The teacher random slope is standardized in a teacher-level data set to have a mean of zero and unit standard deviation.

Appendix B

TPAI-R – Full Review



TPAI-R Full Review---Experienced Teachers

Teacher________________________ Assignment________________________

School________________________ Date________________________

Instructions

- Based on the evidence from the formal observation of an entire class period, the pre-conference notes, rating form, and discussion, artifacts, and the Individual Growth Plan, the evaluator is to rate the teacher's performance with respect to the 8 major functions of teaching listed below.
- The evaluator must add pertinent comments at the end of each major function for which a rating of Above Standard, Below Standard, or Unsatisfactory is given.
- The teacher is provided an opportunity to react to the evaluator's ratings and comments.
- The evaluator and the teacher must discuss the results of the appraisal and any recommended actions pertinent to it.
- The teacher and the evaluator must sign the instrument in the assigned spaces.
- The instrument must be filed in the teacher's personnel folder.
- The rating scale will include the four Levels of Performance described below.

4 **Above Standard**
Performance is consistently high. Teaching practices are demonstrated at a high level. Teacher seeks to expand scope of competencies and undertakes additional appropriate responsibilities.

3 **At Standard**
Performance within this function area is consistently adequate/acceptable. Teaching practices fully meet all performance expectations at an acceptable level. Teacher maintains an adequate scope of competencies and performs additional responsibilities as assigned.

2 **Below Standard**
Performance within this function area is sometimes inadequate/unacceptable and needs improvement. Teacher requires supervision and assistance to maintain an adequate scope of competencies and sometimes fails to perform additional responsibilities as assigned.

1 **Unsatisfactory**
Performance within this function area is consistently inadequate or unacceptable and most practices require considerable improvement to fully meet minimum expectations. Teacher requires close and frequent supervision in the performance of all responsibilities.

1. Major Function: Management of Instructional Time	Above Standard	At Standard	Below Standard	Unsatisfactory

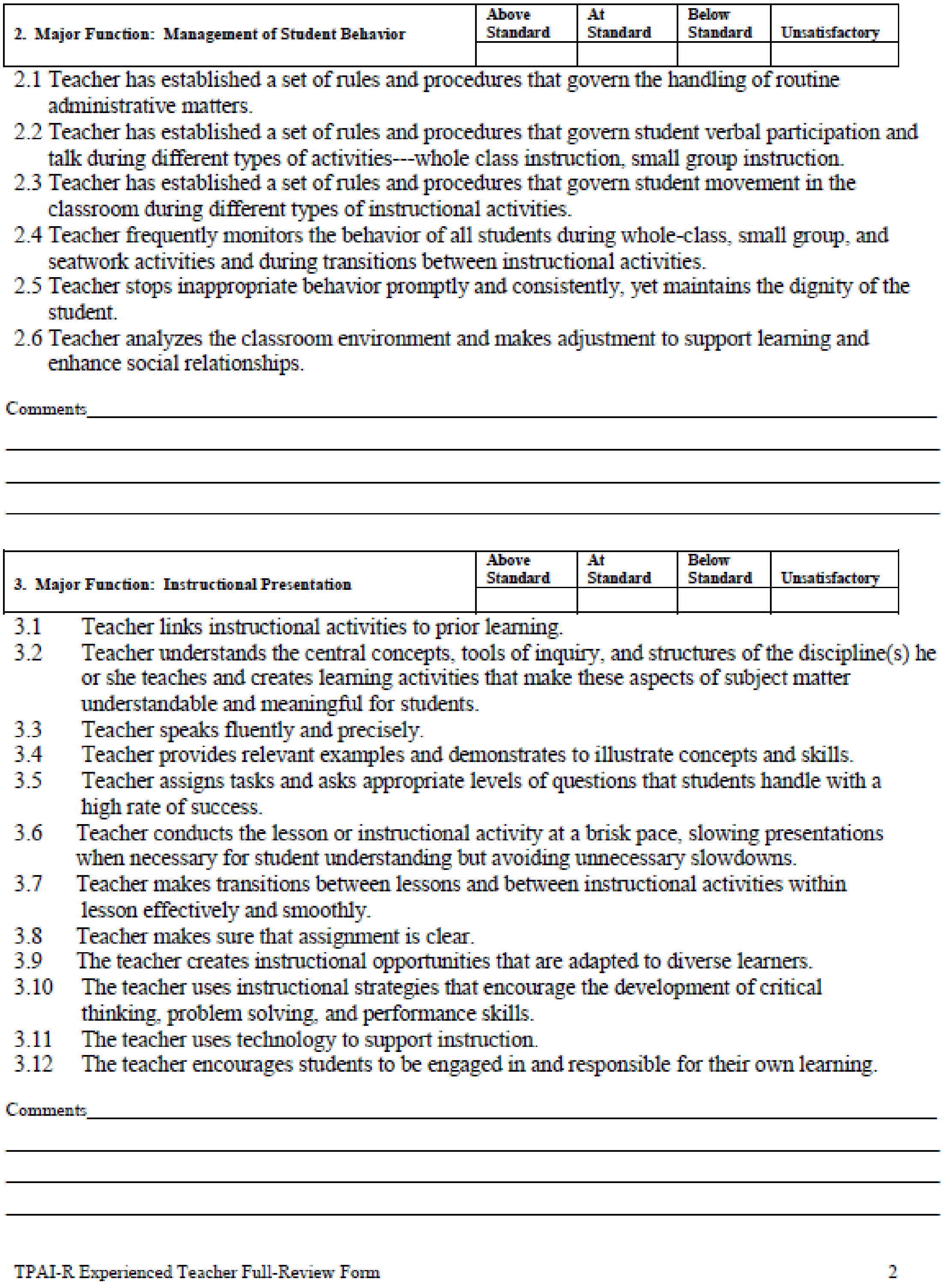
1.1 Teacher has materials, supplies, and equipment ready at the start of the lesson or instructional activity.
1.2 Teacher gets the class started quickly.
1.3 Teacher uses available time for learning and keeps students on task.

Comments____

____

____

____



2. Major Function: Management of Student Behavior	Above Standard	At Standard	Below Standard	Unsatisfactory

2.1 Teacher has established a set of rules and procedures that govern the handling of routine administrative matters.

2.2 Teacher has established a set of rules and procedures that govern student verbal participation and talk during different types of activities---whole class instruction, small group instruction.

2.3 Teacher has established a set of rules and procedures that govern student movement in the classroom during different types of instructional activities.

2.4 Teacher frequently monitors the behavior of all students during whole-class, small group, and seatwork activities and during transitions between instructional activities.

2.5 Teacher stops inappropriate behavior promptly and consistently, yet maintains the dignity of the student.

2.6 Teacher analyzes the classroom environment and makes adjustment to support learning and enhance social relationships.

Comments____

____

____

____

3. Major Function: Instructional Presentation	Above Standard	At Standard	Below Standard	Unsatisfactory

3.1 Teacher links instructional activities to prior learning.

3.2 Teacher understands the central concepts, tools of inquiry, and structures of the discipline(s) he or she teaches and creates learning activities that make these aspects of subject matter understandable and meaningful for students.

3.3 Teacher speaks fluently and precisely.

3.4 Teacher provides relevant examples and demonstrates to illustrate concepts and skills.

3.5 Teacher assigns tasks and asks appropriate levels of questions that students handle with a high rate of success.

3.6 Teacher conducts the lesson or instructional activity at a brisk pace, slowing presentations when necessary for student understanding but avoiding unnecessary slowdowns.

3.7 Teacher makes transitions between lessons and between instructional activities within lesson effectively and smoothly.

3.8 Teacher makes sure that assignment is clear.

3.9 The teacher creates instructional opportunities that are adapted to diverse learners.

3.10 The teacher uses instructional strategies that encourage the development of critical thinking, problem solving, and performance skills.

3.11 The teacher uses technology to support instruction.

3.12 The teacher encourages students to be engaged in and responsible for their own learning.

Comments____

____

____

____

4. Major Function: Instructional Monitoring	Above Standard	At Standard	Below Standard	Unsatisfactory

4.1 Teacher maintains clear, firm, and reasonable work standards and due dates.
4.2 Teacher circulates to check all students' performances.
4.3 Teacher routinely uses oral, written, and other work products to evaluate the effects of instructional activities and to check student progress.
4.4 Teacher poses questions clearly and one at a time.
4.5 Teacher uses student responses to adjust teaching as necessary.

Comments

5. Major Function: Instructional Feedback	Above Standard	At Standard	Below Standard	Unsatisfactory

5.1 Teacher provides feedback on the correctness or incorrectness of in-class work to encourage student growth.
5.2 Teacher regularly provides prompt feedback on out-of-class work.
5.3 Teacher affirms a correct oral response appropriately and moves on.
5.4 Teacher provides sustaining feedback after an incorrect response by probing, repeating the question, giving a clue, or allowing more time.
5.5 The teacher uses knowledge of effective verbal and non-verbal communication techniques to foster active inquiry, collaboration, and supportive interaction in the classroom.

Comments

6. Major Function: Facilitating Instruction	Above Standard	At Standard	Below Standard	Unsatisfactory

6.1 Teacher has long- and short-term instructional plans that are compatible with school and district curricular goals, the school improvement plan, the NC Standard Course of Study, and the diverse needs of students and the community.
6.2 Teacher uses diagnostic information obtained from tests and other formal and informal assessment procedures to evaluate and ensure the continuous intellectual, social, and physical development of the learner.
6.3 Teacher maintains accurate records to document student performance.
6.4 Teacher understands how students learn and develop and plans appropriate instructional activities for diverse student needs and different levels of difficulty.
6.5 Teacher uses available human and material resources to support the instructional program.

Comments

7. Major Function: Communicating within the Educational Environment	Above Standard	At Standard	Below Standard	Unsatisfactory

7.1 Teacher treats all students in a fair and equitable manner.

7.2 Teacher participates in the development of a broad vision of the school.

7.3 Teacher fosters relationships with school colleagues, parents, and community agencies to support students' learning and well being.

Comments____

____

____

____

8. Major Function: Performing Non-Instructional Duties	Above Standard	At Standard	Below Standard	Unsatisfactory

8.1 Teacher carries out non-instructional duties as assigned and/or as need is perceived to ensure student safety outside the classroom.

8.2 Teacher adheres to established laws, policies, rules, and regulations.

8.3 Teacher follows a plan for professional development and actively seeks out opportunities to grow professionally.

8.4 Teacher is a reflective practitioner who continually evaluates the effects of his or her decisions and actions on students, parents, and other professionals in the learning community.

Comments____

____

____

____

Evaluator's Summary:Comments

____

____

____

____

____

Teacher's Reactions to Evaluation

____

____

____

____

____

______________________________ Evaluator's Signature and Date

______________________________ *Teacher's Signature and Date

**Signature indicates that the written evaluation has been seen and discussed and does not necessarily* indicate agreement.