NBER WORKING PAPER SERIES

DO FIRST IMPRESSIONS MATTER? IMPROVEMENT IN EARLY CAREER TEACHER EFFECTIVENESS

Allison Atteberry
Susanna Loeb
James Wyckoff

Working Paper 19096
http://www.nber.org/papers/w19096

NATIONAL BUREAU OF ECONOMIC RESEARCH
1050 Massachusetts Avenue
Cambridge, MA 02138
June 2013

We appreciate helpful comments from Matt Kraft, Eric Taylor, and Tim Sass on previous versions of the paper. We are grateful to the New York City Department of Education and the New York State Education Department for the data employed in this paper. We appreciate financial support from the National Center for the Analysis of Longitudinal Data in Education Research (CALDER). CALDER is supported by IES Grant R305A060018. Support has also been provided by IES Grant R305B100009 to the University of Virginia and by a grant from the National Science Foundation. The views expressed in the paper are solely those of the authors and may not reflect those of the funders or the National Bureau of Economic Research. Any errors are attributable to the authors.

Do First Impressions Matter? Improvement in Early Career Teacher Effectiveness
Allison Atteberry, Susanna Loeb, and James Wyckoff
NBER Working Paper No. 19096
June 2013
JEL No. I21

ABSTRACT

Educational policymakers struggle to find ways to improve the quality of the teacher workforce. The early career period represents a unique opportunity to identify struggling teachers, examine the likelihood of future improvement, and make strategic pre-tenure investments in improvement as well as dismissals to increase teaching quality. To date, only a little is known about the dynamics of teacher performance in the first five years. This paper asks how much teachers vary in performance improvement during their first five years of teaching and to what extent initial job performance predicts later performance. We find that, on average, initial performance is quite predictive of future performance, far more so than typically measured teacher characteristics. Predictions are particularly powerful at the extremes. We employ these predictions to explore the likelihood of personnel actions that inappropriately distinguish performance when such predictions are mistaken as well as the much less discussed costs of failure to distinguish performance when meaningful differences exist. The results have important consequences for improving the quality of the teacher workforce.

Allison Atteberry
Curry School of Education
University of Virginia
P.O. Box 400277
Charlottesville, VA 22904-4277
acma@virginia.edu

Susanna Loeb
524 CERAS, 520 Galvez Mall
Stanford University
Stanford, CA 94305
and NBER
sloeb@stanford.edu

James Wyckoff
Curry School of Education
University of Virginia
P.O. Box 400277
Charlottesville, VA 22904-4277
wyckoff@virginia.edu

Introduction

Teachers vary widely in their ability to improve student achievement, and the difference between effective and ineffective teachers has substantial effects on standardized test outcomes (Rivkin et al., 2005; Rockoff, 2004) as well as later life outcomes (Chetty, Friedman, & Rockoff, 2011). Given the research on the differential impact of teachers and the vast expansion of student achievement testing, policy-makers are increasingly interested in how measures of teacher effectiveness, such as value-added, might be useful for improving the overall quality of the teacher workforce. Some of these efforts focus on identifying high-quality teachers for rewards, to take on more challenging assignments, or as models of expert practice (see for example, teacher effectiveness policies in the District of Columbia Public Schools). Others attempt to identify struggling teachers in need of mentoring or professional development to improve skills (Taylor & Tyler, 2011; Yoon, 2007). Finally, because some teachers may never become effective, some researchers and policymakers are exploring meaningful increases in dismissals of ineffective teaches as a mechanism for improving the overall quality of teachers. One common feature of all of these efforts is the need to establish a system to identify teachers' effectiveness as early as possible in a way that accurately predicts how well these inexperienced teachers might serve students in the long run.

To date, only a little is known about the dynamics of teacher performance in the first five years. As in other occupations, the early career period represents a unique opportunity to identify struggling teachers, examine the likelihood of future improvement, and make strategic pre-tenure investments in improvement as well as dismissals to increase teaching quality. While there are several possible measures of teacher performance, this paper examines value-added estimates in particular. Value-added scores are illustrative of teacher performance more broadly, and their use herein is not intended to suggest that value-added scores should be used in isolation, without regard to classroom practice, or in place of a principal's judgment. The research community acknowledges the limitations of value-added scores as measures of teacher quality, though existing research also suggests that these measures capture something meaningful about how teachers influence student's math and reading skills, as well as longer term outcomes. This paper relies on value-added measures only due to the lack of an alternative measure of teacher effectiveness that covers the first five years of teachers' careers. Similar analyses could use alternative measures as they become available.

This paper explores how teacher performance in the first two years as measured by value-added predicts future teacher performance. In service of this larger goal, we lay out a set of questions designed to provide policy makers with concrete insight into how well teacher value-added scores from the first two years of a teacher's career would perform as an early signal of how that teacher would develop over the next five years. The analyses are based on panel data from the New York City Department of Education that follows all new teachers who began teaching between the 1999-00 and 2006-07 school years through 2011-12 to pursue the following research questions:

- How much do teachers vary in performance improvement during their first five years of teaching?
- To what extent does initial job performance relate to later performance improvement?
- How accurately do measures of initial performance predict future performance?
- Extending the third question, we ask: When predictions are not accurate, what are the tradeoffs associated with making errors?

The following section provides background for the relevance of the research questions, as well as a review of existing literature that helps frame the issue. We then describe the data from

New York City used in the analysis, as well as the analytic approach used to answer these three research questions. The Results section follows, and is organized by research question.

Background and Prior Literature

Research documents substantial impact of assignment to a high-quality teacher on student achievement, as well as the fact that teachers are not uniformly effective (Aaronson, Barrow, & Sander, 2007; Boyd, Lankford, Loeb, Ronfeldt, & Wyckoff, 2011; Clotfelter et al., 2007; Hanushek, 1971; Hanushek, Kain, O'Brien, & Rivkin, 2005; Harris & Sass, 2011; Murnane & Phillips, 1981; Rockoff, 2004). The difference between effective and ineffective teachers affects short term outcomes like standardized test scores, as well as longer term outcomes such as college attendance, wages, housing quality, family planning, and retirement savings (Chetty et al., 2011).

Despite the variation in teacher effectiveness, teacher workforce policies generally ignore variation in quality. In the *Widget Effect,* Weisberg, Sexton, Mulhern, & Keeling, (2009) surveyed twelve large districts across four states and found that performance measures were not considered in recruitment, hiring or placement, professional development, compensation, granting tenure, retention, or layoffs except in three isolated cases (Weisberg, Sexton, Mulhern, & Keeling, 2009). While evaluation and compensation reform is currently popular, the vast majority of districts in the U.S. still primarily use teacher educational attainment, additional credentialing, and experience to determine compensation. In addition, while principal observations of teachers is common practice, there is very little variation in principals' evaluations of teachers (Weisberg et al., 2009).

Given the growing recognition of the differential impacts of teachers, policy-makers are increasingly interested in how measures of teacher effectiveness such as value-added or structured observational measures might be useful for improving the overall quality of the teacher workforce. The Measures of Effective Teaching (MET Project), Ohio's Teacher Evaluation System (TES), and D.C.'s IMPACT policy are all examples where value-added scores are considered in conjunction with other evidence from the classroom, such as observational protocols or principal assessments.

The utility of teacher effectiveness measures for policy use depends on properties of the measures themselves, such as validity and reliability. Measurement work on the reliability of teacher value-added scores has typically characterized reliability using a perspective based on the logic of test-retest reliability, in which a test administered twice within a short time period is judged based on the equivalence of the results over time. Researchers have thus examined the stability of value-added scores from one year to the next, reasoning that a reliable measure should be consistent with itself from one year to the next (e.g., Aaronson et al., 2007; D Goldhaber & Hansen, 2010; Kane & Staiger, 2002; Koedel & Betts, 2007; McCaffrey, Sass, Lockwood, & Mihaly, 2009). When value-added scores fluctuate dramatically in adjacent years, this presents a policy challenge—the measures may reflect statistical imprecision more than true teacher performance. In this sense, stability is a highly desirable property in a measure of effectiveness, because the conclusions one would draw based on value-added in one year are more likely to be consistent with conclusions made in another year.

Year to year variation in value-added measures may be due to errors in measurement but it may also be due to true differences in performance from one year to the next. These true differences over time may be particularly pronounced for new teachers. Researchers have documented substantial increases in value-added over the first years of teacher with a leveling off of returns to experience after five to seven years (Clotfelter, Ladd, & Vigdor, 2006; Clotfelter et al., 2007; Rivkin et al., 2005; Rockoff, 2004).[1] Given that teachers exhibit the largest returns to experience during their early phase, one might expect teacher quality measures to be less stable during this time even if they reliably measure latent true quality as it develops. In theory performance measures early in a teacher's career may be just as predictive of future scores as later measures despite their instability.

That said, there are reasons to be skeptical about our ability to make fair and accurate judgments about teachers based on their first one or two years in the classroom. Anecdotally, one often hears that the first two years of teaching are a “blur,” and that virtually every teacher is overwhelmed and ineffective. If, in fact, first-year teachers’ effectiveness is more subject to random influences and less a reflection of their true abilities, their early evaluations would be less predictive of future performance than evaluations later in their career, with important implications for targeted professional development, tenure and other personnel policies. This paper explores the how actual value-added scores from new teachers’ first two years might be used by policy makers to anticipate the future effectiveness of their teaching force and to identify teachers early in their career for particular human capital responses.

[1] There are clearly higher average student outcomes for students when exposed to teachers with more experience, though there has been more debate about which years are most formative and whether there are no additional returns to experience after a certain point (Papay & Kraft, 2011).

Data

The backbone of the data used for this analysis is administrative records from a range of sources including the New York City Department of Education (NYCDOE), the New York State Education Department (NYSED). The combination of sources provides the student achievement data and the link between teachers and students needed to create measures of teacher effectiveness and growth over time.

New York City students take achievement exams in math and English Language Arts (ELA) in grades three through eight; however, for the current analysis, we restrict the sample to elementary school teachers (grades four and five), because of the relative uniformity of elementary school teaching jobs compared with middle school teaching where teachers specialize. All the exams are aligned to the New York State learning standards and each set of tests is scaled to reflect item difficulty and are equated across grades and over time. Tests are given to all registered students with limited accommodations and exclusions. Thus, for nearly all students the tests provide a consistent assessment of achievement from grade three through grade eight. For most years, the data include scores for 65,000 to 80,000 students in each grade. We normalize all student achievement scores by subject, grade and year to have a mean of zero and a unit standard deviation. Using these data, we construct a set of records with a student's current exam score and lagged exam score(s). The student data also include measures of gender, ethnicity, language spoken at home, free-lunch status, special-education status, number of absences in the prior year, and number of suspensions in the prior year for each student who was active in any of grades three through eight in a given year. For a rich description of teachers, we match data on teachers from the NYCDOE Human Resources database to data from the NYSED databases. The NYCDOE data include information on teacher race, ethnicity, experience, and school assignment as well as a link to the classroom(s) in which that teach taught each year.

Analytic Sample and Attrition

The paper explores how measures of teacher effectiveness—value-added scores—change during the early career. To do this, we rely on the student-level data linked to elementary school teachers to estimate teacher value-added. Value-added scores can only be generated for the subset of teachers assigned to tested grades and subjects. In addition, because we herein analyze patterns in value-added scores over the course of the first five years of a teacher's career, we can only include teachers who do not leave teaching before their later performance can be observed. Not only is limiting the sample to teachers with a complete vector of value-added central to the research question, it also addresses a possible attrition problem. The attrition of teachers from the sample threatens the validity of the estimates because one cannot observe how these teachers would have performed had they remained in the profession, and there is some reason to believe that early attriters may have different returns to experience (Boyd et al., 2007; Dan Goldhaber, Gross, & Player, 2011; Hanushek et al., 2005). As a result, the primary analyses focus on the set of New York City elementary teachers who began between 2000 and 2007 who have value-added scores in all of their first five years.

Despite the advantages to limiting the sample in this way, the restriction introduces a different problem having to do with external validity. If teachers who are less effective leave teaching earlier or are removed from tested subjects or grades, the estimates of mean value-added across the first five years would be biased upward because the sample is limited at the outset to a more effective subset of teachers. That is, teachers who are consistently assigned to tested subjects and grades for five consecutive years may be quite different from those who are not. Given this tradeoff, we conduct sensitivity analyses and present results also for a less restrictive subsample that requires a less complete history of value-added scores.

Table 1 gives a summary of sample sizes by subject and additional requirements based on minimum value-added scores required. There are 7,656 math teachers (7,611 ELA) who are tied to students in NYC, began teaching during the time period in which they could possibly have at least five years of value-added scores, and teach primarily elementary grades during this time. At a very minimum, teachers must possess a value-added score in the first year, which in itself limits the math sample to 4,170 teachers (4,180 for ELA). Our primary analytic sample for the paper is the subset of 842 math teachers who possess a value-added score in at least each of her first five years (859 ELA). The sample sizes decrease dramatically as one increases the number of required value-added scores, which demonstrates our limited ability to look much beyond the first five years. The notable decrease in sample size reveals that teachers generally do not receive value-added scores in every school year, and in research presented elsewhere we examine why so few teachers receive value-added over a consecutive panel (Atteberry, Loeb, & Wyckoff, 2013). Because the requirement of having five consecutive years of value-added scores is somewhat restrictive, we also examine results for the somewhat larger subsample of teachers who remain in the New York City teacher workforce for at least the first five years but have value-added scores in their first year and two of the following four years (n=2,068 for math, 2,073 for ELA).

Methods

The overarching analytic approach in this paper is to follow a panel of new teachers as they go through their first five years and retrospectively examine how performance in the first two years predicts performance thereafter. In order to do so, we first estimate yearly value-added scores for all teachers in New York City. We then use these value-added scores to characterize teachers' developing effectiveness over the first five years to answer the research questions outlined above. We begin by describing the methods used to estimated teacher-by-year value-added scores, and then we lay out how these scores are used in the analysis.

Estimation of Value Added

Although there is no consensus about how best to measure teacher quality, this paper defines teacher effectiveness using a value-added framework in which teachers are judged by their ability to stimulate student standardized test score gains. While imperfect, these measures have the benefit of directly measuring student learning and they have been found to be predictive of other measures of teacher effectiveness such as principals' assessments and observational measures of teaching practice (Atteberry, 2011; Grossman et al., 2010; Jacob & Lefgren, 2008; Kane & Staiger, 2012; Kane, Taylor, Tyler, & Wooten, 2011; Milanowski, 2004), as well as long term student outcomes (Chetty et al., 2011). Our methods for estimating teacher value-added are consistent with the prior literature. Equation 1 describes our approach.[2]

$$A_{itgsy} = \beta_0 + A_{itgs,y-1}\beta_1 + A^{other}_{itgs,y-1}\beta_2 + X_{itgsy}\beta_3 + C_{tgsy}\beta_4 + S_{sy}\beta_5 + \pi_g + \theta_{yt} + \varepsilon_{jitgsy} \qquad (1)$$

[2] To execute the model described in equation (1), we use a modified version of the method proposed by the Value-Added Research Center (VARC). This approach involves a two-stage estimation process, which is intended to allow the researcher to account for classroom characteristics, which are collinear with the teacher-by-experience fixed effects that serve as the value-added models themselves. This group of researchers is currently involved in producing value-added scores for districts such as New York City, Chicago, Atlanta, and Milwaukee (among others). For more information, see http://varc.wceruw.org/methodology.php

The outcome A_{itgsy} is the achievement of student i, with teacher t, in grade g, in school s, at time y, and it is modeled as a function of a vector $A_{itgs,y-1}$of that student's prior achievement in the prior year in the same subject and $A^{other}_{itgs,y-1}$in the other subject (math or ELA); the students' characteristics, X_{itgsy}; classroom characteristics, C_{tgsy}, which are the aggregate of student characteristics as well as the average and standard deviation of student prior achievement; $S_{sy}\beta_5$, school time-varying controls, grade fixed effects, π_g; teacher-by-experience fixed effects (θ_{yt}); as well as a random error term, ε_{jitgsy}.[3] The teacher-by-experience fixed effects become the value-added measures which serve as the outcome variable in our later analyses. They capture the average achievement of teacher t's students in year y, conditional on prior skill and student characteristics, relative to the average teacher in the same subject and grade. Finally, we apply an Empirical Bayes shrinkage adjustment to the resulting teacher-by-year fixed effect estimates to adjust for measurement error.

In the model presented above for the estimation of teacher-by-year value-added scores, we make several important analytic choices about model specification. Our preferred model uses a lagged achievement approach wherein a student's score in a given year serves as the outcome, with the prior year score on the right-hand side (as opposed to modeling gain scores as the outcome).[4] The model attends to student sorting issues through the inclusion of all available student covariates rather than using student fixed effects, in part because the latter restricts the

[3] The effects of classroom characteristics are identified from teachers who teach multiple classrooms per year. The value-added models are run on all teachers linked to classrooms from 2000 on, however the analytic sample for this paper is limited to elementary grade teachers.

[4] Some argue that the gain score model is preferred because one does not place any prior achievement scores which are measured with error on the right-hand side, which introduces potential bias. On the other hand, the gain score model has been criticized because there is less variance in a gain score outcome and a general loss of information and heavier reliance on the assumption of interval scaling. In addition, others have pointed out that the gain score model implies that the impacts of interest persist undiminished rather than directly estimating the relationship between prior and current year achievement (McCaffrey, Lockwood, Koretz, Louis, & Hamilton, 2004; McCaffrey et al., 2009).

analysis to comparisons only between teachers who have taught at least some students in common.[5] At the school level we also opt to control for all observed school-level covariates that might influence the outcome of interest rather than including school fixed effects, since this would also only allow valid comparisons within the same school. In an appendix, we examine results across a variety of value-added models, including models with combinations of gain score outcomes, student, and school fixed effects.

RQ 1. How Much Do Teachers Vary in Performance Improvement during their First Five Years of Teaching?

We first estimate the mean returns to experience for teachers during their first five years in order to establish that findings from this dataset are consistent with prior literature. Importantly, however, we also consider whether teachers vary around that overall pattern. That is, we look for evidence of variability in the developmental trajectories of teacher in terms of effectiveness in the early career.

Annual student-level test score data provide the base for estimating returns to experience. In creating measures of growth, we tackle common problems researchers face when estimating returns to experience, particularly isolating the impact of experience on student achievement. We estimate teachers' improvement with experience using a standard education production function quite similar to Equation 1 in that both include the same set of lagged test scores, student, classroom, and school covariates, as well as grade fixed effects. We remove teacher-by-

[5] A student fixed effects approach has the advantage of controlling for all observed and unobserved time-invariant student factors, thus perhaps strengthening protections against bias. However, the inclusion of student-level fixed effects entails a dramatic decrease in degrees of freedom, and thus a great deal of precision is lost (see discussion in McCaffrey et al., 2009). In addition, experimental research by Kane and Staiger (2008) suggests that student fixed effects estimates may be *more* biased than similar models using a limited number of student covariates.

experience fixed effects and replace them with experience level and year fixed effects. The coefficients of interest are those on the set of experience variables. If the experience measures are indicator variables for each year of experience, the coefficient on the binary variable that indicates an observation occurred in a teacher's fifth year represents the expected difference in outcomes between students who have a teacher in her first versus fifth year, controlling for all other variables in the model. We plot these estimated coefficients alongside estimates from other research projects since the mean trend has been the focus of considerable prior work.

We are primarily interested in the extent to which teachers vary around this mean trend. In order to explore this, we randomly sample 50 teachers from our analytic sample and plot their observed value-added scores during their first five years. We also present the standard deviation of estimated value-added scores across teachers at each year of experience to examine whether the variance in teacher effectiveness appears to be widening or narrowing during the early career. If we observe a narrowing in the range of effectiveness during the early career, one might assume that teachers converge to some extent in terms of performance. If, on the other hand, the standard deviation remains the same or widens, it suggests that existing differences in performance may be sustained over time.

RQ 2. To What Extent Does Initial Job Performance Relate to Later Performance Improvement?

To build off the analyses exploring variability around mean returns to experience, we explore whether one possible source of that variability is differences in teachers' initial effectiveness. We therefore begin by estimating mean value-added score trajectories during the first five years separately by quintiles of teachers' initial performance. Policy makers often translate raw evaluation scores into multiple performance groups in order to facilitate direct action for top and bottom performers. We also adopt this general approach for characterizing early career performance for a given teacher for many of our analyses. (The creation of such quintiles, however, requires analytic decisions that we delineate in Appendix A.) In addition, we estimate the proportion of variability in future performance that can be accounted for using performance measures in the first and second year.

In order to examine how the development of teacher effectiveness during the early career varies by quintile of initial performance, we model the teacher-by-year value-added measures generated by Equation (1) as outcomes using a non-parametric function of experience with interactions for initial quintile. We plot the coefficients on the interactions of experience and quintile dummy variables to illustrate separate mean value-added trajectories by initial quintile.

Quintile groupings may obscure differences between teachers at either extreme within the same quintile, or it may exaggerate the differences between teachers just on either side of one of these cut points. For this reason, we present analyses that move away from reliance on quintiles in order to characterize the relationship between continuous measures of initial and future performance among new teachers. We estimate regression models that predict a teacher's continuous value-added score in a future period as a function of a set of her value-added scores in the first two years of teaching.

We use Equation (3) to predict each teacher's value-added score in a given "future" year (e.g., value-added score in years three, four, five, or the mean of these) as a function of value-added scores observed in the first and second year. We present results across a number of value-added outcomes and sets of early career value-added scores, however Equation (3) describes the fullest specification which includes a cubic polynomial function of all available value-added data in both subjects from teachers' first two years:

$$E\left[VA_{m,y=3,4,5}\right] = \beta_0 + f^3\left(VA_{m,y=1}\right) + f^3\left(VA_{m,y=2}\right) + f^3\left(VA_{e,y=1}\right) + f^3\left(VA_{e,y=2}\right) \qquad (3)$$

We summarize results from forty different permutations of Equation (3)—by subject and by various combinations of value-added scores used—by presenting the adjusted R-squared values from each model. This comparison illustrates the proportion of variance in future performance that can be accounted for using early value-added scores, and to easily consider the comparative improvements of using more scores or different scores in combination with one another.

RQ 3. How Accurately do Measures of Initial Performance Predict Future Performance?

We characterize the predictive power of early career performance measures from the first two years in order to provide guidance to policy-makers and district leaders seeking to anticipate the longer-run performance of their developing workforce. First, we are interested in whether *any* initially high-performing teachers are later among the lowest-performing teachers and whether *any* initially low-performing teachers are later among the highest-performing teachers. For this we present a quintile transition matrix that tabulates the number of teachers in each initial quintile (rows) by the number of teachers in each quintile of the mean of their following three years (columns), along with row percentages.

We next examine residuals and confidence intervals around forecasted future scores from the most promising specifications of Equation (4) above. We conclude the section by presenting the distribution of future performance scores separately by quintiles of initial performance. This allows one to visually examine the extent to which initial teacher groupings based on initial performance quintiles overlap in estimated skill in future years. To the extent that these distributions are distinct from one another, it suggests that the initial performance quintiles accurately predict future performance, and the extent to which the distributions overlap indicates potential errors in predictions.

RQ 4: When Predictions are Not Accurate, What Are the Tradeoffs Associated with Making Errors?

Because we know that errors in prediction are inevitable, we present evidence on the nature of the miscategorizations one might make based on value-added scores from a teacher's first two years. We present a framework for thinking about the kinds of mistakes likely to be made and for whom those mistakes are costly. We base this framework loosely on the statistical concept of Type I and Type II errors, and we then apply this framework to historical data from New York City. We propose a hypothetical policy mechanism in which value-added scores from the early career are used to rank teachers and identify the strongest or weakest for any given human capital response (be it merit pay, professional development, probation, dismissal, etc.). We then follow teachers into their third through fifth years and calculate the proportion of the initially identified teachers who actually turn out to be high- or low- effective teachers in the long run. In addition, we present some evidence on how teachers of different race/ ethnicity might be differentially affected by policies which attempt to predict future performance based on initial performance measures.

Results

RQ 1. How Much Do Teachers Vary in Performance Improvement during their First Five Years of Teaching?

Figure 1 depicts returns to experience from eight studies, as well as our own estimates using data from New York City.[6] Each study shows increases in student achievement as teachers accumulate experience such that by a teacher's fifth year her or his students are performing, on average, from five to 15 percent of a standard deviation of student achievement higher than when he or she was a first year teacher. This effect is substantial, given that a one standard deviation increase in teacher effectiveness is typically 15 to 20 percent of standard deviation of student achievement. Thus, the average development over the first few years of teaching is from one-third to a full standard deviation in overall teacher effectiveness.[7]

Figure 1 demonstrates that early career teacher experience is associated with large student achievement gains, on average. However, average early career improvement may obscure the substantial variation across teachers around this mean trajectory—that is, some teachers may improve a lot over time while others do not. Indeed, we find evidence of substantial variance in value-added to student achievement across teachers. Figure 2 plots the observed value-added score trajectories for 50 teachers who were randomly sampled from the set of New York City elementary teachers that have value-added scores in their first five years (our analytic sample),

[6] Results are not directly comparable due to differences in grade level, population, and model specification, however Figure 1 is intended to provide some context for estimated returns to experience across studies for our preliminary results.

[7] See Hanushek, Rivkin, Figlio, & Jacob (2010) for a summary of studies that estimate the standard deviation of teacher effectiveness measures in terms of student achievement. The estimates for Reading are between 0.11 and 0.26 standard deviations across studies, while the estimates for math are larger and also exhibit somewhat more variability (0.11 to 0.36, but with the average around 0.18 standard deviations (Aaronson et al., 2007; Hanushek & Rivkin, 2010; Jacob & Lefgren, 2008; Kane, Rockoff, & Staiger, 2008; Thomas J. Kane & D.O. Staiger, 2008; Koedel & Betts, 2011; Nye, Konstantopoulos, & Hedges, 2004; Rivkin et al., 2005; Rockoff, 2004; Rothstein, 2010).

alongside the mean value-added scores (red) in the same period. This graph illustrates notable variability around the mean growth during this time period, which suggests that the mean returns to experience may not characterize individual teachers well.

To further explore variation in returns to experience, we calculate the standard deviation of teacher value-added scores across teachers within each year of experience for both the complete analytic sample and the teachers randomly selected for Figure 2. For English Language Arts (ELA) the standard deviations in teacher value-added is 0.18 across teachers in their first year (experience = 0). For math, the standard deviation of first-year teacher value-added is approximately 0.21.[8] As Figure 2 shows, the variance in both ELA and math value-added scores increases yearly. The standard deviation in math value added is 0.24 by the fifth year of teaching, representing an increase of 15 to 30 percent from the first year. The trends suggest that the processes associated with teacher development create greater differences in teaching effectiveness over these early years of teaching.

RQ 2. To What Extent Does Initial Job Performance Relate to Later Performance Improvement?

One way to make sense of the substantial variability observed above is to examine mean value-added scores over years of experience separately by quintiles of initial performance. If initial performance provides insight into future performance, we should see that the highest quintile of initial performance continues to be the highest performing quintile over time (and vice versa for the initially lowest quintile). We group teachers by initial performance quintiles of the mean of their first *two* years. Figure 3 plots mean value-added scores by experience for each

[8] The standard deviations reported here are calculated as the standard deviation of estimated value-added scores, and recall that the primary value-added scores used throughout the paper are shrunk. These standard deviations are not intended to estimate the true variance of teacher effectiveness by experience year, but rather to show a trend over time. The subject of estimating the true variance is taken up in a separate paper.

quintile of performance in the first two years among teachers with value-added scores in at least the first five years. (See Appendix for a series of checks using different samples of teachers based on minimum years of value-added scores required, definitions of initial performance quintiles, and specifications of the value-added model.)

Figure 3 provides evidence of consistent differences in value-added across quintiles of initial performance. On average, the initially lowest-performing teachers are consistently the lowest-performing, the highest are consistently the highest. While the lowest quintile does exhibit the most improvement, this set of teachers does not, on average, "catch up" with other quintiles, nor are they typically as strong as the median first year teacher even after five years.

The results in Figures 1-3 begin to provide a picture of how teachers improve over the first five years. First, consistent with prior findings this is a period of growth overall. Second, in the face of this overall trend, we also observe considerable variability in the patterns of development during this time frame, as evidenced by the plots of individual teachers in Figure 2 and the depiction of quintile-based trajectories in Figure 3.

In Table 3, we present adjusted R-squared values from various specifications of Equation (4) above, and we present results across five possible sets of early career value-added scores to explore the additional returns to using more value-added scores. One evident pattern is that additional years of value-added predictors improve the predictions of future value-added—particularly the difference between having one score and having two scores. The lowest adjusted R-squared values come from models that predict a value-added score in one future year using one value-added score from a single prior year. For example, teachers' math value-added scores in the first year only explains 7.9 percent of the variance in value-added scores in the third year. The predictive power is even lower for ELA (2.5 percent). A second evident pattern in Table 4 is that value-added scores from the second year are typically two- to three times stronger predictors than value-added in the first year for both math and ELA.

Recall that elementary school teachers typically teach both math and ELA every year and thus we can estimate both a math and an ELA score for each teacher in each year. When we combine all available value-added scores from both subjects in both of the first two years, and also include cubic polynomial terms for theses scores, we can explain slightly more variance in future scores. Table 4 also shows that the measure of future score is as important as the measure of initial score. Initial scores do a far better job of predicting a teachers' average value-added over a group of years than of predicting value-added in any of the individual years. For math, when including all first and second year value-added measures, we explain about 26.1 percent of the variance in average future performance compared with no more than 17.6 percent of the variance in any of the individual future years. (For ELA, the comparable results are 17.8 percent and 11.3 percent.)

Table 3 shows early scores can explain up to approximately one-fourth to one-fifth of the variation in future scores; however, it is not necessarily clear whether this magnitude is relatively big or relatively small. For comparison, we estimate the predictive ability of measured characteristics of teachers during their early years. These include typically available measures: indicators of a teacher's pathway into teaching, available credentialing scores and SAT scores, competiveness of undergraduate institution, teacher's race/ ethnicity, and gender. When we predict math mean value-added scores in years three through five using this set of explanatory factors, we explain only 2.8 percent of the variation in the math outcome and 2.5 percent of the variation in the ELA outcomes.[9] The measured teacher characteristics that district leaders

[9] These results not shown, available upon request.

typically have at their disposal to predict who will be the most or least effective teachers clearly do not perform as well as value-added scores from the first two years.

RQ 3. How Accurately do Measures of Initial Performance Predict Future Performance?

The prior analyses provide evidence that future performance depends in part on initial performance; however, the analyses also imply that this predictive ability is far from perfect. In this section we further describe the degree of accuracy associated with these predictions. One shortcoming of the mean improvement trajectories by quintile shown above in Figure 3 is that it may obscure further important within-quintile variance. That is, it provides little information about whether *any* initially high-performing teachers become among the lowest-performing teachers in the future (or vice versa). In Table 4, we present a quintile transition matrix that tabulates the number of teachers in each initial quintile (rows) by the number of teachers in each quintile of the mean of their following three years (columns), along with row percentages.[10] The majority—61.9 percent—of the initially lowest quintile math teachers ultimately show up in the bottom two quintiles of future performance. On the other end, the initially highest-performing teachers exhibit even more consistency: About 68.9 percent of these teachers remain in the top two quintiles of mean math performance in the following years. Movements from one extreme to the other are comparatively rare. About 21.0 percent of bottom- and 10.2 percent of top- quintile initial performers end up in the opposite extreme two quintiles. Results are similar for ELA teaching. Overall, the transition matrix suggests that measures of value-added in the first two years predict future performance for most teachers.

[10] We use the mean of years 3, 4, and 5 rather than just the fifth year to absorb some of the inherently noisy nature of value-added scores over time.

To provide another perspective on our ability to predict future value-added scores, we return to Equation (4) above, in which we model mean value-added scores in years three through five as cubic polynomial functions of value-added scores in both subjects in the first two years. Using this model, we can predict future performance and present a conservative confidence interval for each forecasted prediction point (see Figure 4).

As Figure 4 shows, even 80 percent confidence intervals are quite large for individual predictions. The mean squared error for teachers in this sample is about 0.14, which is approximately equivalent to a standard deviation in the overall distribution of teacher effectiveness. The degree of error for individual predictions is substantively large, and we can see that teachers' predicted future value-added scores differ markedly from the observed scores based on distance from the y=x line. That said, recall that the adjusted r-squared from this simple model of future performance is high—about 27.8 percent of the variance in future performance can be accounted for using value-added scores in the first and second years. Certainly the value-added based predictions of future performance are imprecise, and accordingly most policy makers argue that value-added scores should not be used in isolation to reward or sanction teachers. The Measures of Effective Teaching (MET) study explores the potential benefits to combining multiple measures to generate more reliable teacher effectiveness estimates. Nonetheless, the movement towards a more strategic approach to human capital management in the K-12 setting drives us to consider the utility of the tools at hand in light of the current lack of strong alternatives on which to base predictions of how teachers will serve students throughout their career.

Given the confidence intervals shown in Figure 4, a policy that uses value-added scores to group teachers based on performance will produce groups that are not entirely distinct from one another in future years. Figure 5 presents the complete distribution of future value-added scores by initial quintile. These depictions provide a more complete sense of how groups based on initial effectiveness overlap in the future.[11] For each group, we have added two reference points, which are helpful for thinking critically about the implications of these distributions relative to one another. First, the "+" sign located on each distribution represents the mean of future performance in each respective initial-quintile group. The color-coded vertical lines represent the mean *first* year performance by quintile. This allows the reader to compare distributions both to where the group started on average, as well as to where other groups have ended up on average in future years.

The vast majority of policy proposals based on value-added target teachers at the top (for rewards, mentoring roles, etc.) or at the bottom (for support, professional development, or dismissal). Thus, even though the middle quintiles are not particularly distinct in Figure 5, it is most relevant that the top and bottom initial quintiles are. In both math and ELA, there is some overlap of the extreme quintiles in the middle—some of the initially lowest-performing teachers appear to be just as skilled in future years as initially high-performing teachers. However, the majority of these two distributions are distinct from one another.

We can take a closer look at the initially lowest quintile of performance relative to some meaningful comparison points. For example in math, the large majority (76.5 percent) of the density of the red distribution lies to the left of the mean of the distribution of future scores for the middle quintile (the comparable percentage is 74.4 percent for ELA). Thus, three fourths of the initially lowest performers never match the performance of an average fifth year teacher (of course this implies that about a quarter of the initially-lowest performing quintile—those who

[11] The value-added scores depicted in each distribution are each teacher's mean value-added score in years three, four, and five. For brevity, we refer to these scores as "future" performance.

appear at the very top of the red distribution of future performance— do surpass the mean of the middle quintile). One can conduct a similar analysis using smaller groupings of teachers than the quintiles described here. For example, one could examine what percentage of the top/bottom decile (or even bottom twentieth) out-perform an average teacher in the future. We address this below by making use of more fine-grained groupings of teachers.

RQ 4: When Predictions are Not Accurate, What are the Tradeoffs Associated with Making Errors?

This discussion lends itself naturally to a consideration of the tradeoffs associated with identifying teachers as low-performing based on imperfect measurements from a short period of time in the early career. The goal is to maximize the percentage of teachers for whom we accurately predict future performance based on early performance. There are two possible errors—Type I and Type II—that one could make in service of this goal. We begin with the null hypothesis that a given teacher is *not* ineffective in the long run (for the sake of simplicity, think of this as assuming a teacher is effective). Type I error is rejecting a true null hypothesis, which in this case means to falsely identify a teacher as low-performing when she turns out to be at least average in the long run. The degree of Type I error could be quantified by examining the percentage of teachers who are initially identified as ineffective who turn out to be effective in future years. This type of error typically dominates the value-added debate, because this error negatively and unfairly penalizes teachers who would be identified as ineffective even though they *would have* emerged as effective over time. On the other hand, Type II error is often overlooked even though it directly affects students' instructional experiences. In the case of Type II error, one fails to reject a false null hypothesis. For the case at hand, this implies that one fails to identify a teacher as ineffective when she actually is ineffective in the long run. This error might be quantified as the percentage of teachers who were not identified as low-performing initially but nonetheless perform poorly in the long run. Students who are assigned to teachers who persist as a result of Type II error receive a lower quality of instruction than they would have had the teacher been replaced. In practice, school districts typically seek to identify only a small proportion of the workforce as either very effective of ineffective. In this scenario, Type I errors are minimized, though likely at the expense of Type II errors. At the low end of the distribution, this penalizes students with more ineffective teachers.

While we have framed the discussion of Type I and Type II error in terms of identifying ineffective teachers, a parallel approach can be taken to identifying excellent teachers. In this case, the null hypothesis is that a given teacher is *not* high performing in the long run. Type I error is rejecting a true null hypothesis—predicting that a teacher will be excellent when he or she is not. Type II error is not rejecting the null when it is true—thinking that a teacher will not be excellent when he or she is. To the extent that excellent teachers deserve recognition, Type II errors could impact teachers individually and collectively.

In practice, identifying Type I and Type II errors is complex, in part because it requires a clear criterion for identifying future ineffectiveness and excellence. The measures we have of future quality are imprecise; narrow, as they are based only on student test performance in math and ELA; and relative instead of absolute, as they compare teacher to each other rather than to a set standard. We have addressed to some extent the measurement error in a teacher's value-added measure in a given year by using Bayes shrunk estimates which attenuates extreme measures in proportion to their imprecision, as well as averaging across multiple future years to lessen the influence of any one outlier result. We, however, cannot address the narrowness of the value-added measure, nor its relative nature. Again, we return to the idea that using multiple measures of teacher effectiveness—e.g., value-added augmented by rigorous observation protocols and other measures—would increase reliability and broaden the domains that are measured. In the end, policy makers will establish thresholds for teacher effectiveness to differentiate teachers depending on the particular human resource objective at hand.

To illustrate the potential tradeoffs between Type I and Type II errors, we use the current data as an opportunity to examine how well one could have predicted teachers' future performance based on early career value-added measures. There are a number of reasons why district leaders might try to make such predictions. For example, if one can identify early teachers who are likely to struggle in their future careers, a policy could target this set of teachers for professional development or additional support. Another possibility would be to delay tenure decisions for teachers who perform relatively low in their first year or two.[12] In the current example, we describe a generic policy which identifies a certain percentage of new teachers as initially low-performing, inherently predicting that these teachers are likely to be low-performing in the future. We compare those who are identified by this generic policy (i.e., below some initial performance threshold) to those who are not identified (above that threshold), and we see the frequency with which Type I and Type II errors are made.

For this analysis, we calculate the mean of a teacher's value-added scores in years one and two and translate that into percentiles of initial performance. Figure 6 plots future terciles of performance as a function of these initial performance percentiles. Moving from left to right

[12] There are reasons to identify high-performing teachers early, as well. For example, these teachers might themselves be strong mentors to other new teachers. In addition, if initially highly-effective teachers are likely to continue to be among the highest performing in the future, then a policy might attempt to compensate these teachers to encourage their continued participation in the teacher workforce. In practice, one could analyze the impact of *any* number of strategic policy responses using this same approach of balancing Type I and Type II errors (e.g., support, professional development, mentoring, compensation, tenure, dismissal). In our example, we describe a generic policy which merely identifies a teacher who is predicted to be low-performing in the future, but we are not suggesting a particular policy response to these teachers.

along the x-axis represents an increase in the threshold for identifying a teacher as ineffective based on these percentiles. In the left panel of Figure 6, we depict the set of teachers who fall *below* a given threshold and thus are identified as low-performing. The y-axis depicts the percentage of each group—those who fall either below the threshold (left) or above the threshold (right)—who subsequently appear in each tercile of future performance, with separate lines for the bottom, middle, and highest third of the distribution. If we imagine a vertical line that passes through X=10 on the horizontal axis, this line would provide information on the results of classifying the lowest ten percent of teachers as low performing. The solid red line shows that approximately 64 percent of these teachers would fall in the lowest tercile. That is, 64 percent would be in the bottom third of future performers. The dashed yellow line show that approximately 24 percent would be in the middle third of future performers, while the dotted green line shows that the remaining approximately 13 percent would be in the top third of future performers.

In the right panel of Figure 6, we depict the corresponding set of teachers who fall *above* that same threshold (i.e., the other 90 percent who are *not* identified as low-performing). Of the 90 percent of teachers not identified as low performing in the above example, approximately 39 percent would be in the top third, another 39 percent would be in the middle third, and approximately 22 percent would be in the bottom third.

We can garner a great deal of information from this figure. First, it is clear that while there are errors in identifying ineffective teachers even when initial ineffectiveness is defined at a low level, most of the teachers identified as low-performing also show up in the bottom third of the distribution of future performance. Type I errors—captured by the green line on the left panel—are thus relatively infrequent. These are the set of teachers who were initially identified as low-performers but who in the future appear in the top third of the performance distribution. Type I errors become slightly more frequent as one raises the threshold of initial performance and thus aims to identify a higher proportion of teachers as ineffective.

Type II errors are depicted on the right panel based on the red line: These are the teachers who were not initially identified as low-performing based on the given threshold (x-axis), but who ultimately appear in the bottom third (red) of future performance. When the threshold for low-performance is the bottom ten percent, then by definition the other 90 percent of teachers are *not* identified as low-performing. The right panel shows that group of unidentified teachers are about equally likely to appear in the top two terciles of future performance. Here, the red line summarizes the rate of Type II errors.

Consider another hypothetical policy that identifies the bottom 5 percent of teachers in initial value-added as low-performing and thus eligible for some policy response (e.g., mentoring, PD, additional oversight). In this case, we are attempting to test a hypothesis about whether a teacher will be ineffective or not (the null hypothesis). For math, Figure 6 indicates that, among the 5 percent of teachers identified, 75.0 percent subsequently appear in the bottom third of the distribution of future performance, 16.7 percent appear in the middle third of the distribution, and only 8.3 percent appear in the top third. At the 5 percent threshold, the top 95 percent of teachers are not identified as ineffective. Of those, 37.6 percent appear in the top third of the future performance distribution, 38.3 percent appear in the middle third, and 24.1 percent appear in the lowest third. At this threshold, the Type I error rate among those identified as low-performing is 8.3 percent, and the Type II error rate among those not identified is 24.1 percent. However, it is also important to keep in mind the relative size of these groups—8.3 percent of the bottom 5 percent of teachers is less than 1 percent of the overall group, while 24.1 percent of the top 90 percent of teachers is about 20 percent of the overall group. In the current analytic sample of new elementary teachers with at least five years of value-added scores (966 teachers in math), these error rates imply a Type I error for fewer than ten teachers, but a Type II error with approximately 200 teachers.

Overall, Figure 6 graphically displays the inherent tradeoffs that come along with making policy decisions based on imperfect information in the early career (first two years). We do see evidence of Type I error—in the range depicted in the graph we see the virtually no Type I errors are made when the identification threshold is low (e.g., below 5 percent of teachers). As one identifies an increasing percentage of teachers as low performing, we see that Type I error rate increase, but only slightly. Even among the bottom 40 percent of teachers identified—the highest threshold depicted in the graph—we see that only 15 percent are observed in the top third in the future. When we look at the right panel, however, we do also see that as Type I error rates increase, Type II error rates go down among teachers who fall above the selected threshold. This illustrates a classic balance at play here between false identifications and failures to identify. Figure 6 also depicts the corresponding rate of making "accurate" predictions at these same thresholds, by looking at the other two lines in each panel.

In the example above, we posited that the top and bottom third of the distribution of future performance could be characterized as high- and low-performing respectively; however one could debate about the appropriate criteria for future effectiveness. Another reasonable assertion might be to characterize every teacher who is ultimately less effective than an *average* teacher and then retained as a Type II error, and every teacher who would have become significantly more effective than an average teacher but is inappropriately identified as a Type I error. We are agnostic about what should be used by policy makers in practice as the "right" criteria, however we acknowledge the very real need to provide evidence for those who must make such decisions. In Table 5, we therefore describe the frequency of transitions to the top, middle, and bottom third of the distribution of future performance, alongside the same information but instead simply by top and bottom half of the distribution. We also now focus on teachers who are in the extremes of the initial performance distribution—that is, the top 5, 10, 15, and 20 percent (the initially highest performers), as well as the bottom 5, 10, 15, and 20 percent. While Figure 6 focuses only on initially low-performing teachers, Table 5 also reports on long term performance of initially high performers. Table 5 shows that these teachers are even more likely to remain consistent in terms of future performance than their initially low-performing counterparts. The row percentages reported in Table 5 for the bottom 5, 10, 15, and 20 percent of initial performers correspond perfectly with the visual relationship depicted in Figure 6; the table simply provides concrete numbers at specific thresholds and allows the reader to look for one's self at different ways of defining adequate future performance.

Ultimately, policymakers will need to make their own decisions about what criteria are used to characterize levels of teacher performance. We have explored quintiles, terciles and top/bottom half of the distribution in this paper thus far. Another possibility is to compare a novice teacher's ongoing performance to that of an average first year teacher, as this represents an individual that could serve as a feasible replacement. In fact, among the teachers in the bottom 5 percent of the initial math performance distribution, the vast majority—83.3 percent—do not perform in their future third, fourth and fifth years as well as an average first year teacher in math. The corresponding figure is 72.2 percent for ELA. In other words, had students who were assigned to these initially lowest-performing teachers instead been assigned to an average new teacher, they would have performed at much higher levels on their end-of-year tests.

More concretely, the average math value-added score of a third-year teacher who initially performed in the bottom 5 percent in years one and two is about -0.15 standard deviation units. The average first-year teacher, on the other hand, has a math value-added score of -0.03 standard deviation units. The difference between the two is almost a full standard deviation in effectiveness for teachers in our data. We therefore expect a large negative difference (around 0.11 standard deviations) in the potential outcomes for students assigned to these initially very low-performing teachers as opposed to an average new teacher, even in the third year alone. Further, an ineffective teacher retained for *three* additional years imposes three years of below-average performance on students. The longer a teacher with low true impacts on students is retained, the expected differential impact on students will be the *sum* of the difference between an average new teacher and the less effective teacher across years of additional retention.

The same logic can be applied to teachers at the high end of the teacher effectiveness spectrum. The average math value-added score of a third-year teacher who initially performed in the top 5 percent in years one and two is 0.24 standard deviation units. Imagine a scenario in which a school system cannot manage to retain this high-performing teacher, and as a result the students who would have been assigned to this teacher are instead assigned to her replacement—an average first year teacher (who would typically have a mean math value-added score of -0.03 standard deviation units). The impacts for these students would be dramatic in magnitude.

One final concern arose as we thought about the implications of any policy that attempts to predict future performance based on imperfect information from the early career. We worried that the value-added measures used to detect early performance might also identify teachers in other systematic ways. For example, it might be possible that value-added scores tend to be lower for teachers of certain demographic backgrounds and thus subgroups of teachers might be disproportionately identified by such a policy. In the case that being identified early as low-performing increases the likelihood that a teacher exits the profession, it would be possible to see a demographic shift in the composition of the teacher workforce toward less diversity.

To explore this concern, we examine the racial/ethnic breakdown of teachers at different points in the distribution of initial effectiveness (again, according to a teacher's mean value-added in the first two years). Table 6 follows the basic structure as the preceding table. We examine characteristics of teachers who are in the extremes of the initial performance distribution—that is, the top 5, 10, 15, and 20 percent (the initially highest performers), as well as the bottom 5, 10, 15, and 20 percent. For example, we find that 38 of the 59 teachers who are in the top 5 percent of the initial performance distribution for math are white, 9 are black, 5 are Hispanic, and 7 are of another or unknown race. Table 6 also contains the corresponding row percentages for these groups. Of course, the row percentages are not equal to one another—indeed, there are simply far more white teachers in New York City than any other group. Instead, we examine whether relative proportions vary across the initial performance distribution. We find that these proportions are quite similar at the top and bottom of the distribution. White teachers make up about 64.4 percent of the top five percent, 61.9 percent of the top 10 percent, 62.9 percent of the top 15 percent, and 66.7 percent of the top 20 percent. Proportions are again similar for black, Hispanic, and teachers of other or unknown race/ethnicity. Importantly, this is also the case among the lowest performing teachers—relative stability in the demographics of teachers in the bottom 5, 10, 15, and 20 percent of the distribution. If anything, it appears that white teachers are slightly *less* likely to be in the top quintile of performance than in the bottom quintile.[13] These findings suggest that a policy based on early career value-added scores would

[13] In a separate analysis (not shown), we conduct a similar analysis examining the racial breakdown by initial performance, however we separate results across all five quintiles of the distribution of initial performance, rather

not also incidentally identify higher proportions of minority teachers, at least in the case of New York City.

Conclusions

From a policy perspective, the ability to predict future performance is most useful for inexperienced teachers because policies that focus on development (e.g. mentoring programs), dismissal, and promotion are likely most relevant during this period. In this paper we describe the trajectory of teachers' performance over their first five years as measured by their value-added to ELA and math test scores of students and how this trajectory varies across teachers. Our goal is to assess the potential for predicting future performance (performance in years 3, 4, and 5) based on teachers' performance in their first two years. We focus particularly on Type I and Type II error where Type I error is falsely classifying teachers into a group to which they do not belong (e.g. ineffective or excellent) and Type II error is failing to classify teachers into a group to which they belong.

We find that, on average, initial performance is quite predictive of future performance, far more so than measured teacher characteristics such as their own test performance (e.g. SAT) or education. On average the highest fifth of teachers remain the highest fifth of teachers; the second fifth remains the second fifth; the third fifth remains the third fifth; and so on. Predictions are particularly powerful at the extremes. Initially excellent teachers are far more likely to be

than simply the top/ bottom 5, 10, 15, and 20 percent. The findings are similar: There is no evidence that minority teachers are more likely to appear in lower quintiles—there are only slight fluctuations in the racial/ demographic breakdown of quintiles but for black and Hispanic teachers there is no clear pattern in those fluctuations. Again, white teachers appear to be slightly *more* likely to be identified as initially low-performing rather than high-performing, but the differences across quintiles are not large: 63.6, 62.4, 67.4, 67.0, and 76.0 percent of each quintile—top to bottom respectively—are white. Results are available upon request.

excellent teachers in the future than are teachers who were not as effective in their first few years.

This said, any predictions we make about teachers' future performance are far from perfect. The predicted future scores we estimated were, on average, about 0.14 standard deviation units off from actual scores (RMSE), which represents a substantial range of possible effectiveness. Certainly, when it comes to making policy based on imprecise measures of teacher effectiveness, there is no avoiding that some mistakes will be made. Thinking about these errors using the lens of Type I versus Type II errors emphasizes the fact that there are tradeoffs to be made in practice. While most attention has been paid to the former—falsely identifying teachers as ineffective when they ultimately are not—the latter represents the failure to identify and address teaching that does not serve students well in terms of their academic outcomes. The paper highlights the balance between these two kinds of error and also sheds light on how complex it is to definitively know when these mistakes are made.

Tables

Table 1:
Analytic Sample Sizes by Cumulative Restrictions

	MATH		ELA	
	# Tchrs	**# Obs**	**# Tchrs**	**# Obs**
All Teachers Tied to Students in NYC	18,919	62,779	19,567	63,632
Started Teaching in 2000- 2007	16,502	57,603	17,053	58,413
Modal Grade in First Five Years is 4 or 5	5,099	23,633	5,099	23,613
In HR Dataset for At Least 5 Years	3,734	20,641	3,731	20,649
Has VA Score in At Least 1st Year	3,360	16,102	3,307	15,954
Has at Least 2 VA Scores in Next 4 Years	2,333	14,232	2,298	14,080
Has VA in At Least Years 1 thru 3	2,053	12,697	2,026	12,562
Has VA in At Least Years 1 thru 5	966	7,548	972	7,597
Has VA in At Least Years 1 thru 7	376	3,681	390	3,786
Has VA in At Least Years 1 thru 9	145	1,626	148	1,650

Table 2:
Difference in Mean Value Added and Numbers of Final Analytic Sample Teachers in each Quintile of Initial Performance, by Approach to Quintile Construction

		Q1	Q2	Q3	Q4	Q5
Math Quintiles….						
… of All Teacher-Years (1)	n	224	207	194	219	122
	mean	-0.165	-0.049	0.015	0.092	0.222
... After Limiting to Teachers in First Year (2)	n	171	171	198	212	214
	mean	-0.224	-0.100	-0.018	0.063	0.227
… And Limiting to Elementary Teachers (3)	n	150	187	207	213	209
	mean	-0.235	-0.107	-0.018	0.065	0.230
… And Limiting to Teachers with 5+ VA score (4)	n	194	193	193	193	193
	mean	-0.214	-0.083	-0.002	0.077	0.239
ELA Quintiles…						
… of All Teacher-Years (1)	n	246	196	208	181	141
	mean	-0.156	-0.059	0.002	0.066	0.158
... After Limiting to Teachers in First Year (2)	n	214	163	185	198	212
	mean	-0.206	-0.088	-0.022	0.046	0.180
… And Limiting to Elementary Teachers (3)	n	185	176	201	208	202
	mean	-0.217	-0.098	-0.025	0.048	0.185
… And Limiting to Teachers with 5+ VA score (4)	n	195	194	195	194	194
	mean	-0.213	-0.090	-0.016	0.054	0.188

Note: We construct quintiles of performance in a teacher's first two years. The final analytic sample of teachers is restricted to the teachers who taught primarily fourth or fifth grade and for whom we observe at least five consecutive years of VA scores, beginning in the teacher's first year of teaching. Note that method (3) above is the preferred approach for this paper.

Table 3:
Adjusted R-Squared Values for Regressions Predicting Future (Years 3, 4, and 5) VA Scores as a Function of Sets of Value-Added Scores from the First Two Years

	Outcome			
Early Career VA Predictor(s)	VA in Y3	VA in Y4	VA in Y5	Mean($VA_{Y3\text{-}5}$)
Math				
Math VA in Y1 Only	0.079	0.052	0.077	0.111
Math VA in Y2 Only	0.153	0.149	0.117	0.223
Math VA in Y1 & Y2	0.176	0.158	0.146	0.256
VA in Both Subjects in Y1 & Y2	0.176	0.171	0.147	0.262
VA in Both Subjects in Y1 & Y2 (cubic)	0.178	0.171	0.146	0.261
ELA				
ELA VA in Y1 Only	0.025	0.019	0.016	0.040
ELA VA in Y2 Only	0.058	0.080	0.042	0.117
ELA VA in Y1 & Y2	0.068	0.084	0.048	0.131
VA in Both Subjects in Y1 & Y2	0.085	0.102	0.058	0.161
VA in Both Subjects in Y1 & Y2 (cubic)	0.090	0.113	0.061	0.168

Table 4:
Quintile Transition Matrix from Initial Performance to Future Performance, By Subject (Number, Row Percentage, Column Percentage)

		Quintile of Future Math Performance					
Math Initial Quintile		Q1	Q2	Q3	Q4	Q5	Row
Q1	n	47	47	26	25	7	152
	(row %)	(30.9)	(30.9)	(17.1)	(16.4)	(4.6)	
	(col %)	(39.8)	(24.7)	(11.2)	(10.6)	(3.6)	
Q2	n	28	47	60	33	16	184
	(row %)	(15.2)	(25.5)	(32.6)	(17.9)	(8.7)	
	(col %)	(23.7)	(24.7)	(25.8)	(14.0)	(8.2)	
Q3	n	24	47	44	59	34	208
	(row %)	(11.5)	(22.6)	(21.2)	(28.4)	(16.3)	
	(col %)	(20.3)	(24.7)	(18.9)	(25.0)	(17.3)	
Q4	n	14	32	58	64	46	214
	(row %)	(6.5)	(15.0)	(27.1)	(29.9)	(21.5)	
	(col %)	(11.9)	(16.8)	(24.9)	(27.1)	(23.5)	
Q5	n	5	17	45	55	93	215
	(row %)	(2.3)	(7.9)	(20.9)	(25.6)	(43.3)	
	(col %)	(4.2)	(8.9)	(19.3)	(23.3)	(47.4)	
	Column Total	118	190	233	236	196	**973**

		Quintile of Future ELA Performance					
ELA Initial Quintile		Q1	Q2	Q3	Q4	Q5	Row
Q1	n	49	51	44	26	16	186
	(row %)	(26.3)	(27.4)	(23.7)	(14.0)	(8.6)	
	(col %)	(39.2)	(25.1)	(19.0)	(11.0)	(8.6)	
Q2	n	31	40	45	40	22	178
	(row %)	(17.4)	(22.5)	(25.3)	(22.5)	(12.4)	
	(col %)	(24.8)	(19.7)	(19.5)	(16.9)	(11.9)	
Q3	n	19	52	44	58	31	204
	(row %)	(9.3)	(25.5)	(21.6)	(28.4)	(15.2)	
	(col %)	(15.2)	(25.6)	(19.0)	(24.5)	(16.8)	
Q4	n	13	41	48	59	47	208
	(row %)	(6.3)	(19.7)	(23.1)	(28.4)	(22.6)	
	(col %)	(10.4)	(20.2)	(20.8)	(24.9)	(25.4)	
Q5	n	13	19	50	54	69	205
	(row %)	(6.3)	(9.3)	(24.4)	(26.3)	(33.7)	
	(col %)	(10.4)	(9.4)	(21.6)	(22.8)	(37.3)	
	Column Total	125	203	231	237	185	**981**

Table 5:
Movements of Initially Highest- and Lowest- Performing Teachers to Groups of Future of Performance (by Thirds, and by Top and Bottom Half)

	Initial Percentage Identified		Bottom Third	Middle Third	Top Third	Bottom Half	Top Half
MATH	top	5%	1	5	53	2	57
			(1.69)	(8.47)	(89.83)	(3.39)	(96.61)
	top	10%	7	22	89	17	101
			(5.93)	(18.64)	(75.42)	(14.41)	(85.59)
	top	15%	12	37	121	27	143
			(7.06)	(21.76)	(71.18)	(15.88)	(84.12)
	top	20%	16	64	148	42	188
			(7.02)	(28.07)	(64.91)	(18.42)	(82.46)
	bottom	5%	18	4	2	20	4
			(75.00)	(16.67)	(8.33)	(83.33)	(16.67)
	bottom	10%	47	18	10	56	19
			(62.67)	(24.00)	(13.33)	(74.67)	(25.33)
	bottom	15%	68	29	16	86	27
			(60.18)	(25.66)	(14.16)	(76.11)	(23.89)
	bottom	20%	83	44	25	114	39
			(54.61)	(28.95)	(16.45)	(75.00)	(25.66)
ELA	top	5%	6	15	44	10	55
			(9.23)	(23.08)	(67.69)	(15.38)	(84.62)
	top	10%	12	39	63	24	91
			(10.53)	(34.21)	(55.26)	(21.05)	(79.82)
	top	15%	20	56	89	35	132
			(12.12)	(33.94)	(53.94)	(21.21)	(80.00)
	top	20%	26	78	111	50	167
			(12.09)	(36.28)	(51.63)	(23.26)	(77.67)
	bottom	5%	19	13	4	29	7
			(52.78)	(36.11)	(11.11)	(80.56)	(19.44)
	bottom	10%	37	29	20	58	28
			(43.02)	(33.72)	(23.26)	(67.44)	(32.56)
	bottom	15%	64	44	29	95	42
			(46.72)	(32.12)	(21.17)	(69.34)	(30.66)
	bottom	20%	85	63	38	126	60
			(45.70)	(33.87)	(20.43)	(67.74)	(32.26)

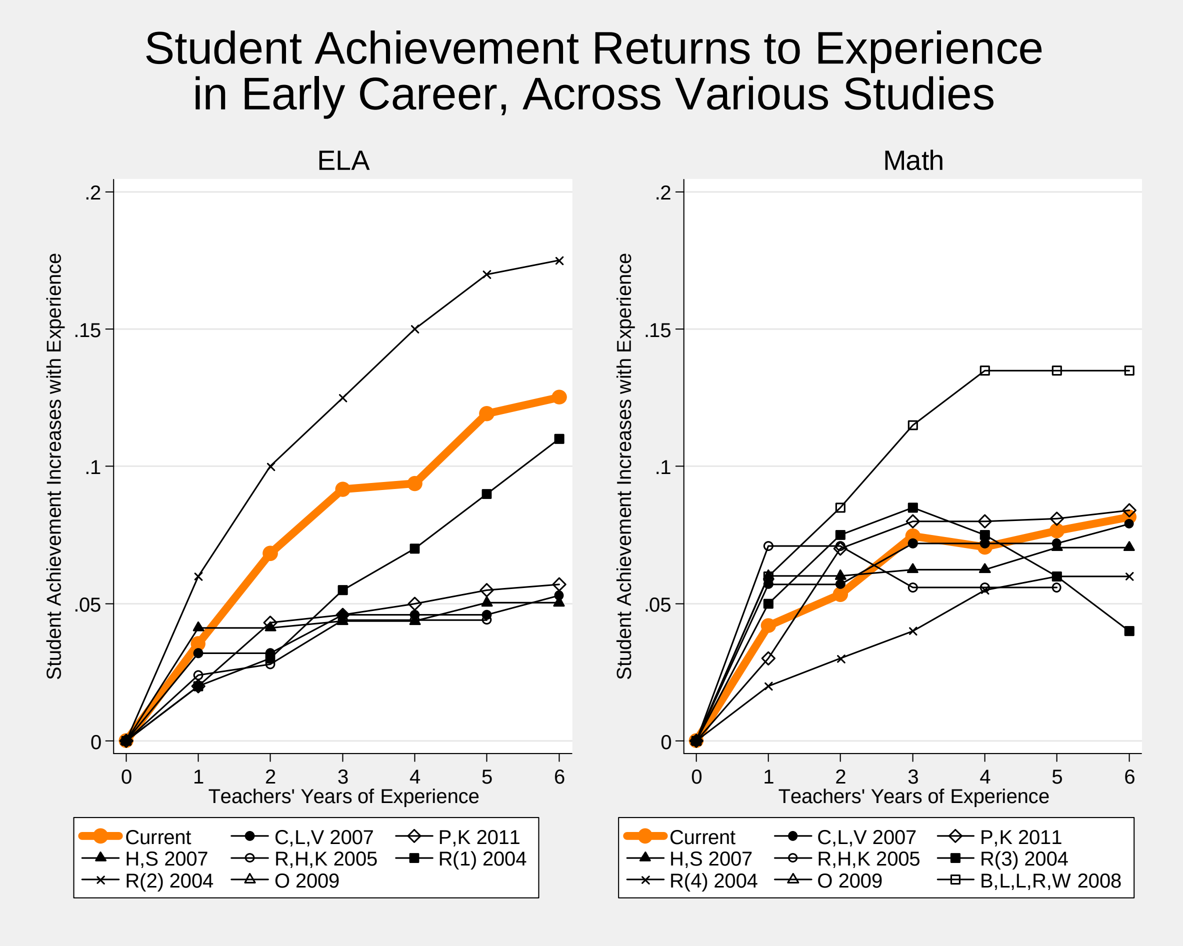
Table reports the number of teachers in each cell, along with corresponding row percentages (below each number, in parentheses). Note that the first three column percentages correspond to the bottom, middle, and top third of the distribution of future performance (as measured by the teacher’s mean value-added score in years 3 through 5), and these three percentages sum to 100 percent. The final two columns break the distribution of future performance into to bottom and top half only, and they also sum to 100 percent.

Table 6:
Teacher Demographics (Count & Row Percentage), by Groups of Initially Highest- and Lowest-Performing Teachers and Subject

		Initial Percentage Identified	**White**	**Black**	**Hispanic**	**Other**	**Row Total**
MATH	top	5%	38	9	5	7	59
			(64.41)	(15.25)	(8.47)	(11.86)	
	top	10%	73	19	13	13	118
			(61.86)	(16.10)	(11.02)	(11.02)	
	top	15%	107	29	19	15	170
			(62.94)	(17.06)	(11.18)	(8.82)	
	top	20%	152	35	23	18	228
			(66.67)	(15.35)	(10.09)	(7.89)	
	bottom	5%	16	3	3	2	24
			(66.67)	(12.50)	(12.50)	(8.33)	
	bottom	10%	56	7	8	4	75
			(74.67)	(9.33)	(10.67)	(5.33)	
	bottom	15%	83	13	11	6	113
			(73.45)	(11.50)	(9.73)	(5.31)	
	bottom	20%	109	22	13	8	152
			(71.71)	(14.47)	(8.55)	(5.26)	
ELA	top	5%	48	6	4	7	65
			(73.85)	(9.23)	(6.15)	(10.77)	
	top	10%	77	18	10	9	114
			(67.54)	(15.79)	(8.77)	(7.89)	
	top	15%	108	27	14	16	165
			(65.45)	(16.36)	(8.48)	(9.70)	
	top	20%	143	34	21	17	215
			(66.51)	(15.81)	(9.77)	(7.91)	
	bottom	5%	23	7	3	3	36
			(63.89)	(19.44)	(8.33)	(8.33)	
	bottom	10%	57	13	9	7	86
			(66.28)	(15.12)	(10.47)	(8.14)	
	bottom	15%	85	27	14	11	137
			(62.04)	(19.71)	(10.22)	(8.03)	
	bottom	20%	120	34	17	15	186
			(64.52)	(18.28)	(9.14)	(8.06)	

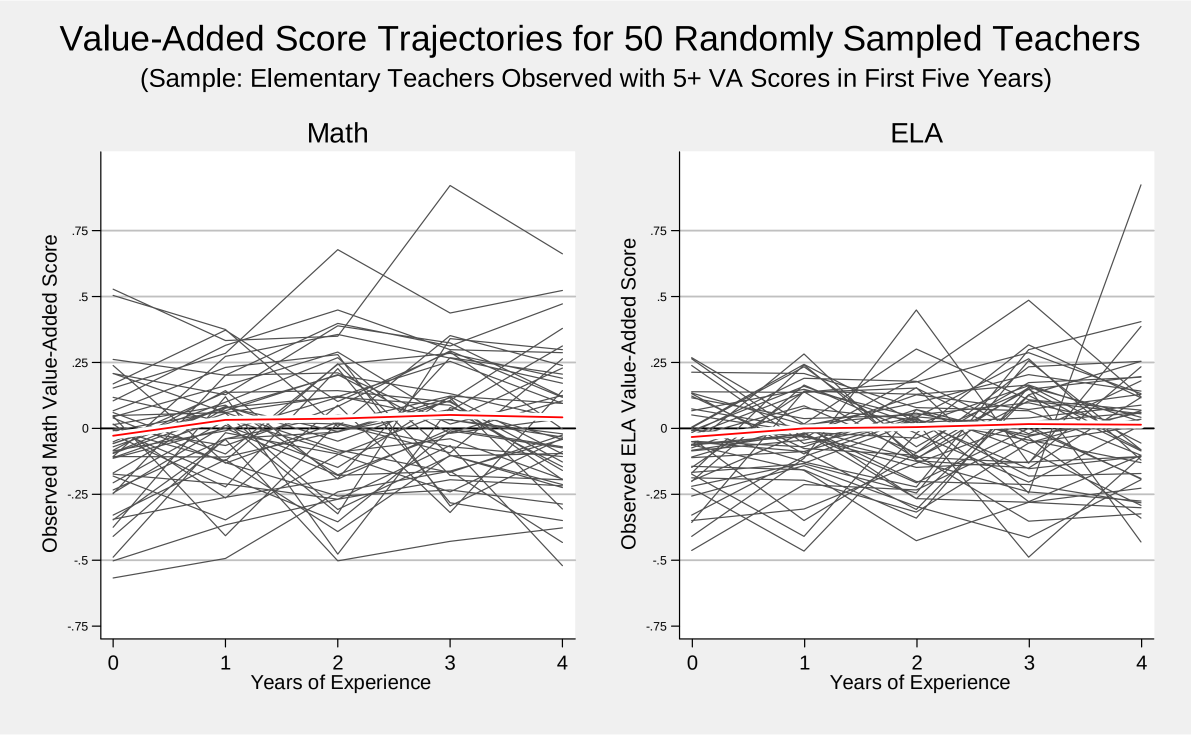
Figures

Figure 1:
Student Achievement Returns to Teacher Early Career Experience, Preliminary Results from Current Study (Bold) and Various Other Studies



Results are not directly comparable due to differences in grade level, population, and model specification, however Figure 1 is intended to provide some context for estimated returns to experience across studies for our preliminary results. Current= Results for grade 4 & 5teachers who began in 2000+ with at least 9 years of experience. For more on model, see Technical Appendix. C,L V 2007= = Clotfelter, Ladd, Vigdor (2007; Rivkin, Hanushek, & Kain, 2005), Table 1, Col. 1 & 3; P, K, 2011 = Papay & Kraft (2011), Figure 4 Two-Stage Model; H, S 2007 = Harris & Sass (2011), Table 3 Col 1, 4 (Table 2); R, H, K, 2005= Rivkin, Hanushek, Kain (2005), Table 7, Col. 4; R(A-D) 2004 = Rockoff (2004), Figure 1 & 2, (A= Vocab, B= Reading Comprehension, C= Math Computation, D= Math Concepts); O 2009 = Ost (2009), Figures 4 & 5 General Experience; B,L,L,R,W 2008 = Boyd, Lankford, Loeb, Rockoff, Wyckoff (2008).

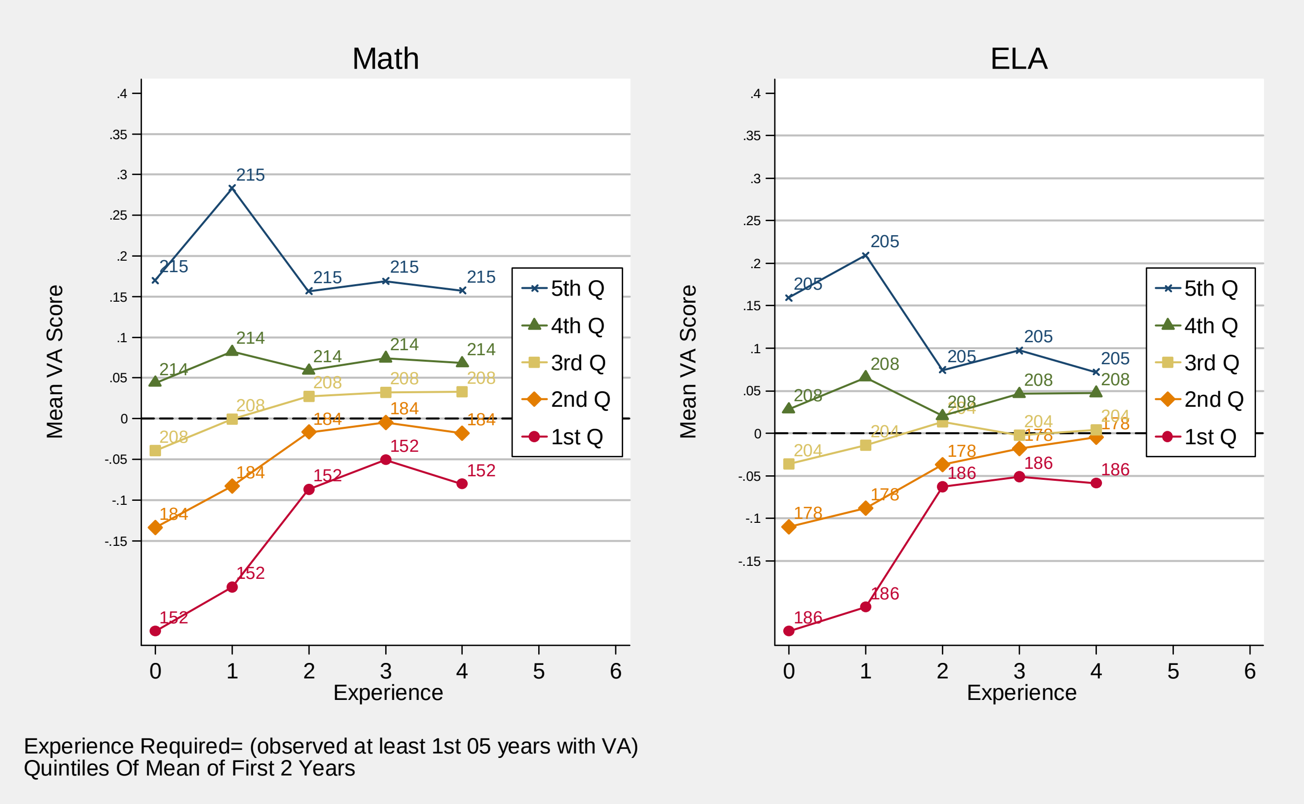
Figure 2:
Variance across Teachers in Quality (VA) over Experience, by Subject and Attrition Group.



Supplement to Figure 2.
Standard Deviation of Estimated Value Added Scores, by Levels of Experience in Figure 2
(Across All Teachers in the Sample, versus 100 Teachers Randomly Sampled for the Figure)

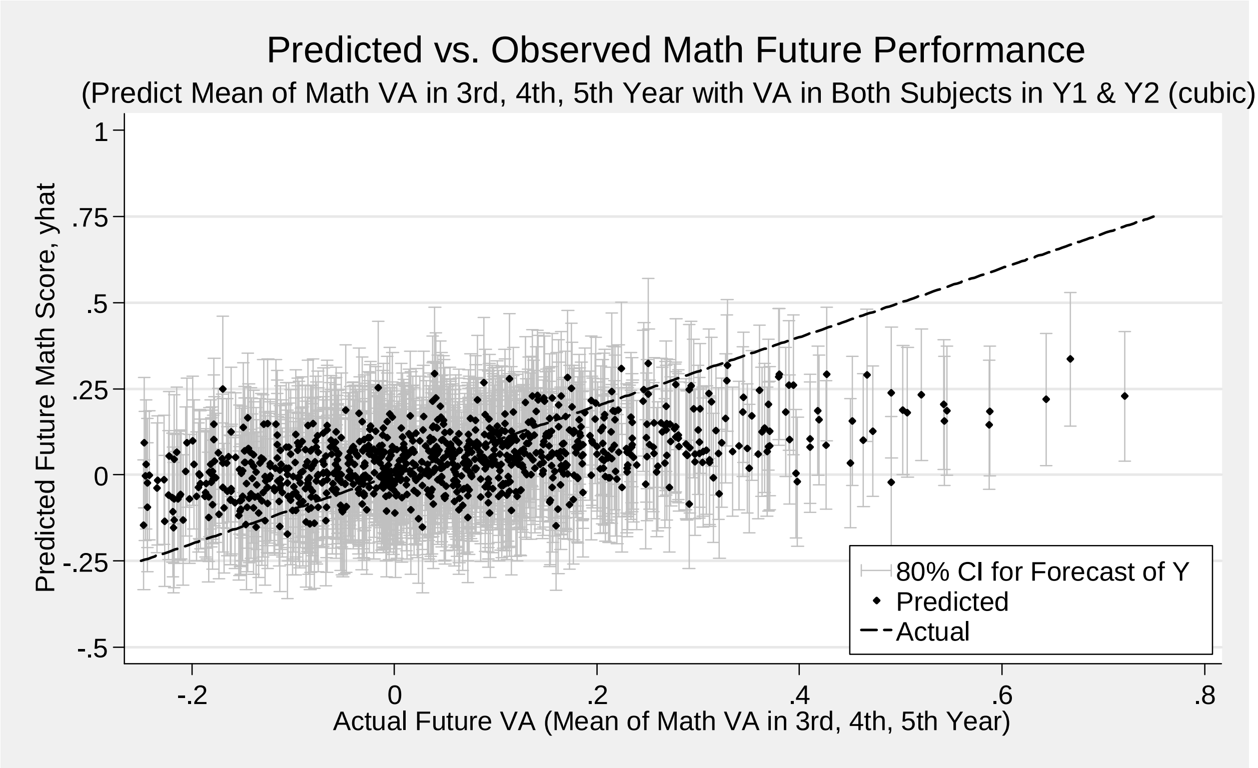
	Math					ELA				
	E= 0	E=1	E=2	E=3	E=4	E= 0	E=1	E=2	E=3	E=4
Full Sample	0.192	0.207	0.213	0.224	0.220	0.177	0.184	0.190	0.193	0.196
100 Teachers	0.209	0.229	0.226	0.238	0.241	0.177	0.188	0.195	0.202	0.205

Figure 3:
Mean VA Scores, by Subject (Math or ELA), Quintile of Initial Performance, and Years of Experience for Elementary School Teachers with VA Scores in at Least First Five Years of Teaching.



Experience Required= (observed at least 1st 05 years with VA)
Quintiles Of Mean of First 2 Years

Figure 4:
Predicted Future Value-Added Scores (Mean of Years, 3,4, and 5) based on Observed Valued-Added Scores in Years 1 and 2, by Actual Future Value-Added Scores, with 80% Confidence Intervals Around Individual Predictions.



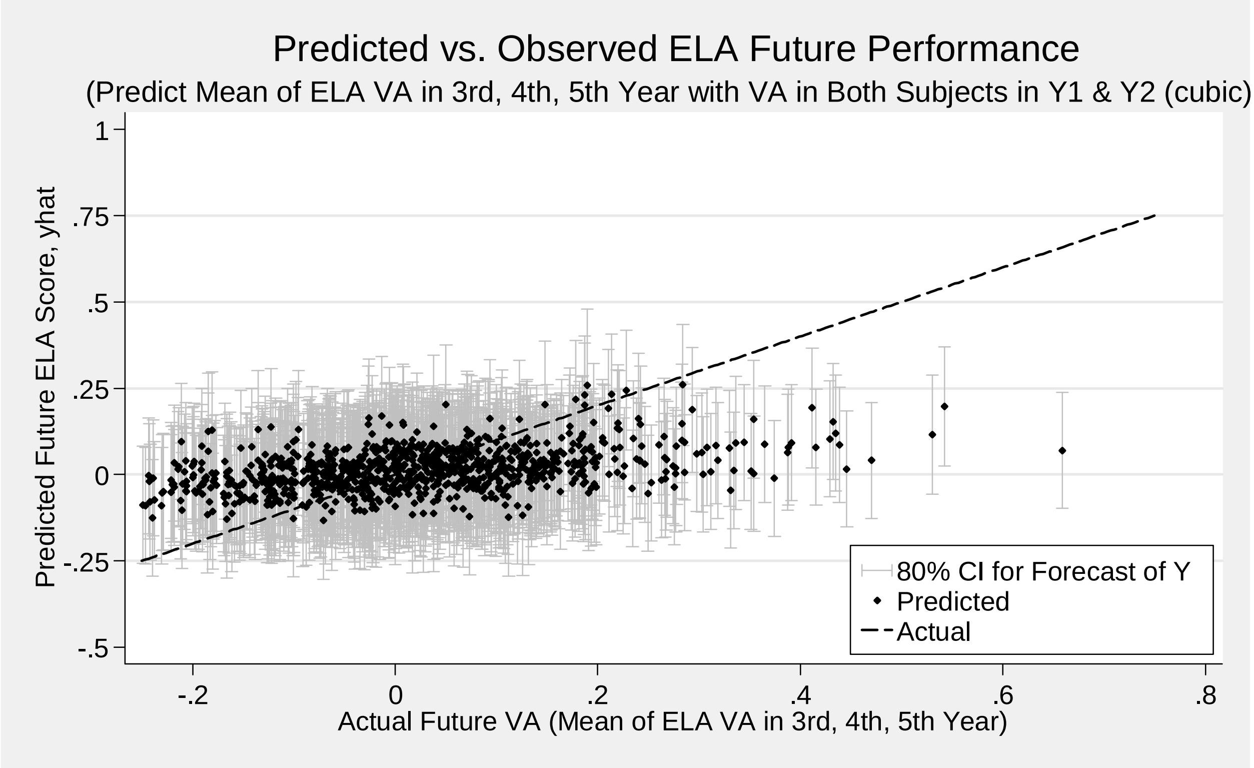


Figure 5:
Distribution of Future Value-Added Scores, by Initial Quintile of Performance

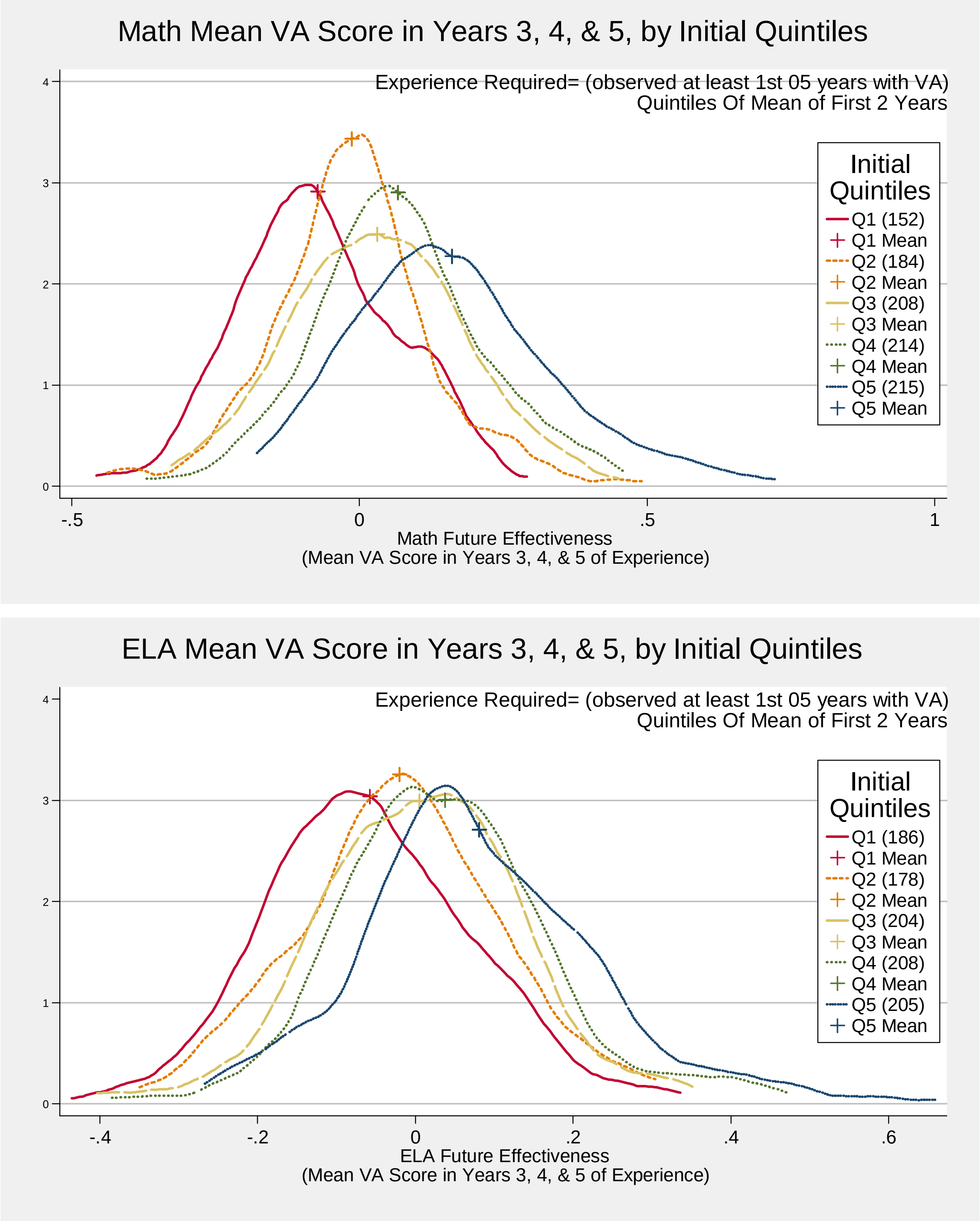
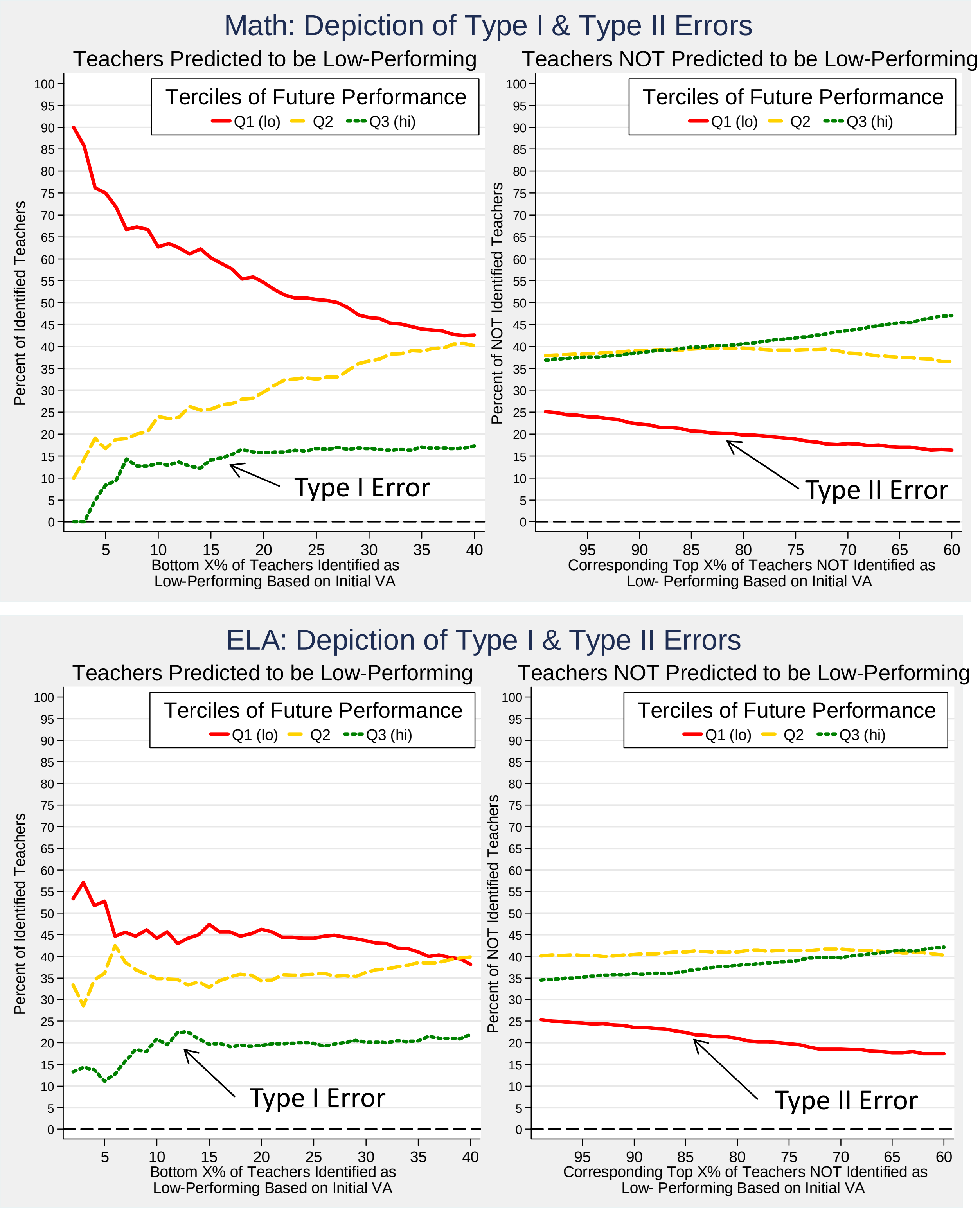


Figure 6:


Math: Depiction of Type I & Type II Errors
Teachers Predicted to be Low-Performing
Teachers NOT Predicted to be Low-Performing
Terciles of Future Performance
Q1 (lo)
Q2
Q3 (hi)
Percent of Identified Teachers
Percent of NOT Identified Teachers
0
5
10
15
20
25
30
35
40
45
50
55
60
65
70
75
80
85
90
95
100
Type I Error
Type II Error
5
10
15
20
25
30
35
40
Bottom X% of Teachers Identified as Low-Performing Based on Initial VA
95
90
85
80
75
70
65
60
Corresponding Top X% of Teachers NOT Identified as Low- Performing Based on Initial VA
ELA: Depiction of Type I & Type II Errors
Teachers Predicted to be Low-Performing
Teachers NOT Predicted to be Low-Performing
Terciles of Future Performance
Q1 (lo)
Q2
Q3 (hi)
Percent of Identified Teachers
Percent of NOT Identified Teachers
Type I Error
Type II Error
Bottom X% of Teachers Identified as Low-Performing Based on Initial VA
Corresponding Top X% of Teachers NOT Identified as Low- Performing Based on Initial VA

Appendix A

The most straightforward approach to making quintiles would be to simply break the full distribution of teacher-by-year fixed effects into five groups of equal size. However, we know that value-added scores for first year teachers are, on average, lower than value-added scores for teachers with more experience. For the purposes of illustration, imagine that first year teacher effects comprise the entire bottom quintile of the full distribution. In this case, we would observe no variability in first year performance—that is, all teachers would be characterized as "bottom quintile" teachers, thus eliminating any variability in initial performance that could be used to predict future performance. We thus chose to center a teacher's first year value-added score around the mean value-added for first year teachers and then created quintiles of these centered scores. By doing so, quintiles captured whether a given teacher was relatively more or less effective than the average *first* year teacher, rather than the average teacher in the district.

In order to trace the development of teachers' effectiveness over their early career, we limited the analytic sample to teachers with a complete set of value-added scores in the first five years. As is evident from Table 1 above, relatively few teachers meet this restrictive inclusion criterion. We hesitated to first restrict the sample and then make quintiles solely within this small subset, because we observed that teachers with a more complete value-added history tended to have higher initial effectiveness. In other words, a "bottom quintile" first year teacher in the distribution of teachers with at least five consecutive years of value-added might not be comparable to the "bottom quintile" among all first years teachers for whom we might wish to make predictions. For this reason, we made quintiles relative to the sample of all teachers regardless of the number of value-added scores they possessed, and subsequently limited the sample to those with at least five years of value-added. As a result of this choice, we observe slightly more top quintile teachers than bottom quintile teachers in the initial year. However by making quintiles before limiting the sample, we preserve the absolute thresholds for those quintiles and thus ensure that they are consistent with the complete distribution of new teachers. In addition, it is simply not feasible for any districts to make quintiles in the first year or two depending on how many value-added scores *will* have in the first five years.

Finally, our ultimate goal is to use value-added information from the early career to produce the most accurate predictions of future performance possible. Given the imprecision of any one year of value-added scores, we average a teacher's value-added scores in years one and two and make quintiles thereof. We present some specification checks by examining our main results using value-added from the first two years in a variety of ways (e.g., first year only, second year only, a weighted average of the first two years, teachers who were consistently in the same quintile in both years). In Table 2, we present the number of teachers and mean of value-added scores in each of five quintiles of initial performance, based on these various methods for constructing quintiles. One can see that the distribution of the teachers in the analytic sample (fourth and fifth grade teachers with value-added scores in first five years) depends on quintile construction.

Appendix B

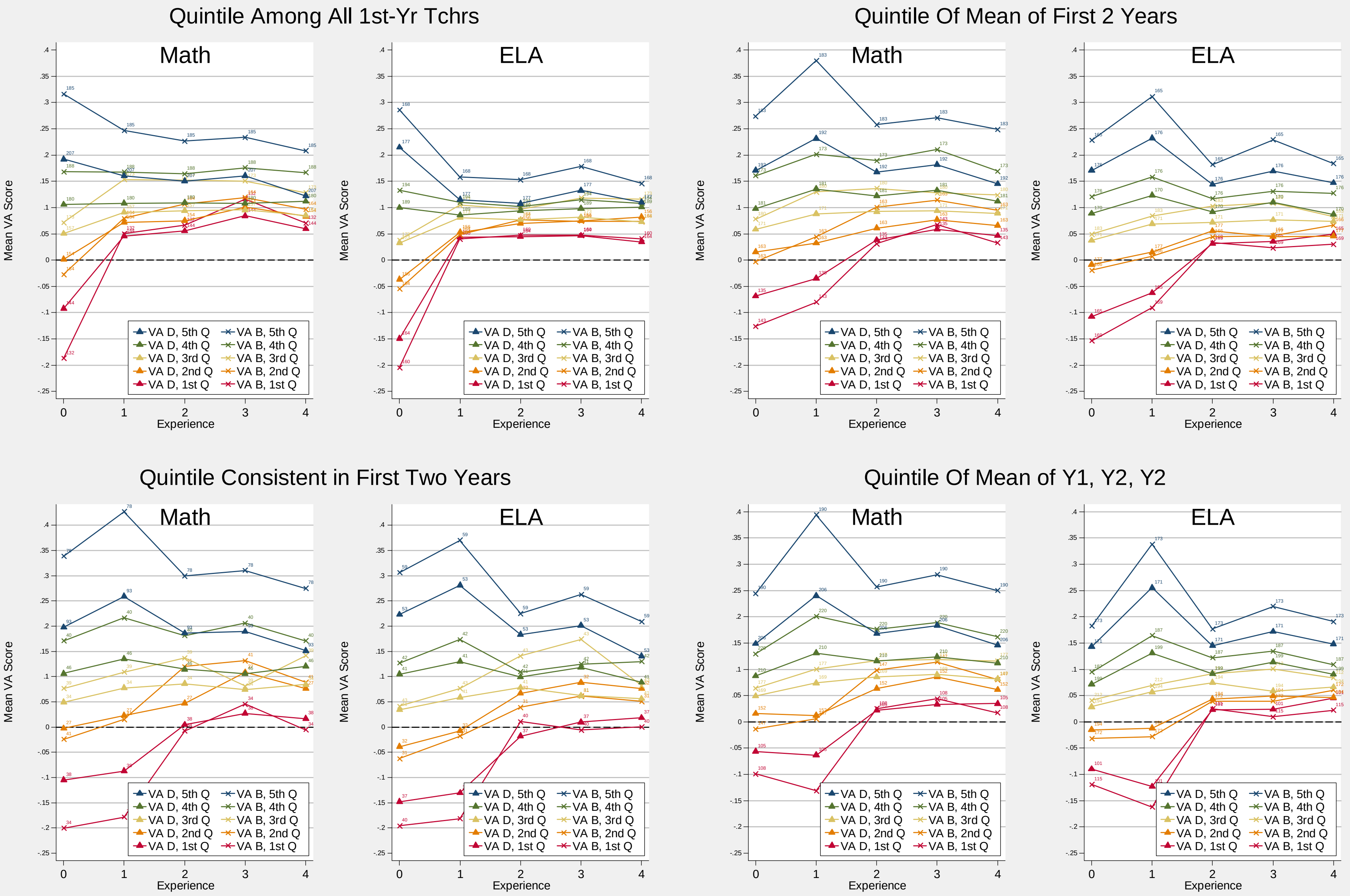
In Figure 3 of the paper, we present mean value-added scores over the first five years of experience, by initial performance quintile. Here we recreate these results across three dimensions: (A) minimum value-added required for inclusion in the sample, (B) how we defined initial quintiles, and (3) specification of the value-added models used to estimate teacher effects:

(A) We examine results across two teacher samples based on minimum value-added required for inclusion. The first figure uses the analytic sample used throughout the main paper—teachers with value-added scores in at least all of their first five years. The second widens the analytic sample to the set of teachers who are consistently present in the dataset for at least five years, but only possess value-added scores in their 1st , and 2 of the next 4 years.

(B) We examine results across four possible ways of defining quintiles: (1) "Quintile of First Year"—this is quintiles of teachers' value-added scores in their first year alone; (2) "Quintile of the Mean of the First Two Years"—this is quintiles of teacher's *mean* value-added scores in the first *two* years and is the approach we use throughout the paper; (3) "Quintile Consistent in First Two Years"—here we group teachers who were consistently in the same quintiles in first and second year (i.e., top quintile both years); and (4) "Quintile of the Mean of Y1, Y2, & Y2"—the quintiles of teacher's mean value added score in first and second year, double-weighting the second year.

(C) Finally, we examine results using two alternative value-added models to the one used in the paper. "VA Model B" uses a gain score approach rather than the lagged achievement approach used in the paper. "VA Model D" differs from the main value-added model described in the paper in that it uses student-fixed effects in place of time-invariant student covariates such as race/ ethnicity, gender, etc. See next page for results.

Elem Teachers with VAM in All of First Five Years



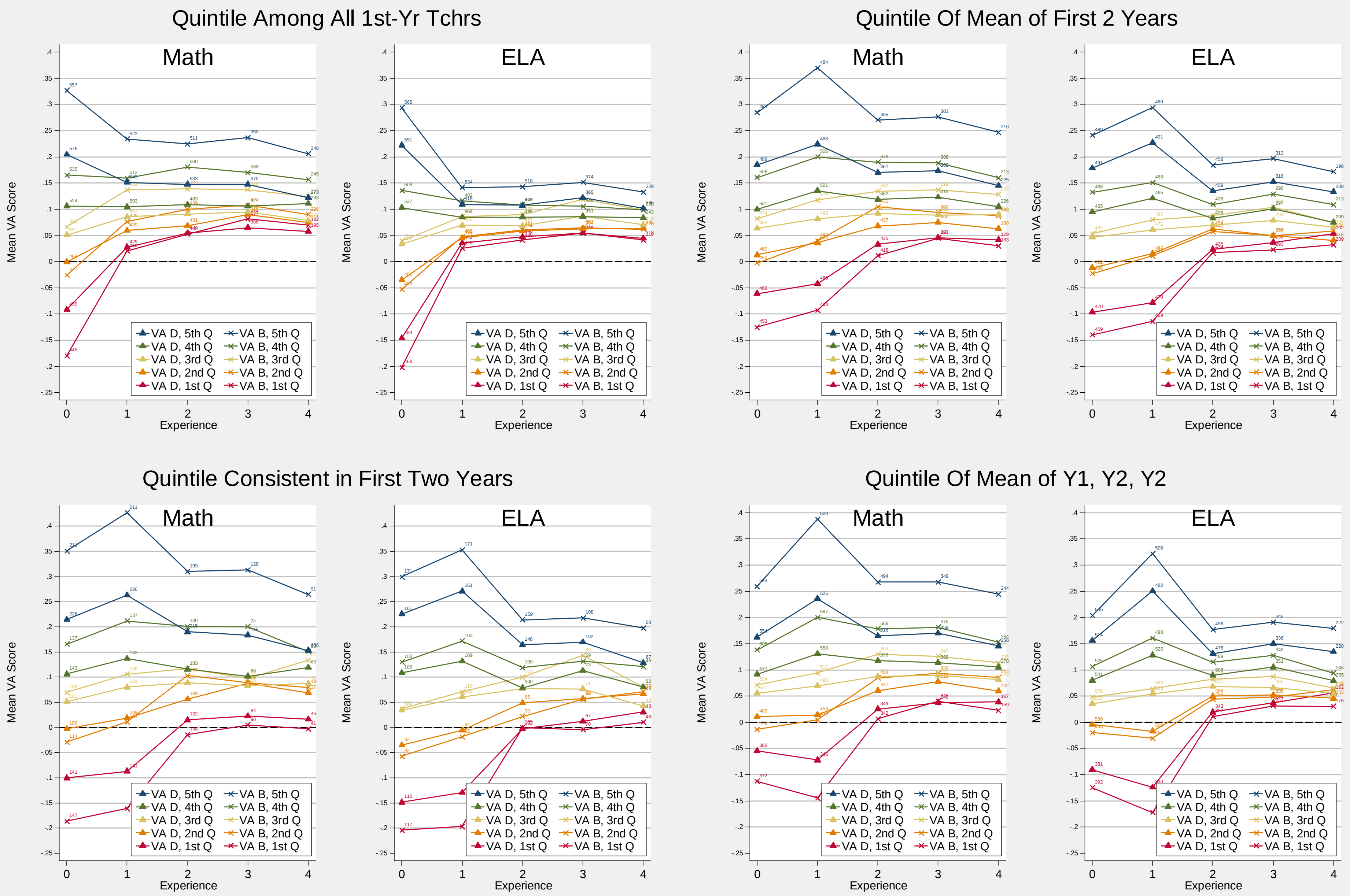
Quintile Among All 1st-Yr Tchrs

Quintile Of Mean of First 2 Years

Quintile Consistent in First Two Years

Quintile Of Mean of Y1, Y2, Y2

Elem Teachers with VAM in 1st, and 2 of Next 4 Years



References

Aaronson, D., Barrow, L., & Sander, W. (2007). Teachers and student achievement in the Chicago public high schools. *Journal of Labor Economics, 25*(1).

Atteberry, A. (2011). *Stacking up: Comparing Teacher Value Added and Expert Assessment*. Working Paper.

Atteberry, A., Loeb, S., & Wyckoff, J. (2013). *Teacher Attrition from High Stakes Testing: Strategic Behavior or the Normal Chaos?* CEPWC Working Paper. University of Virginia.

Boyd, D. J., Grossman, P., Lankford, H., Loeb, S., Wyckoff, J., & Urban, I. (2007). Who leaves. *The implications of teacher attrition for school achievement. Retrieved April, 17*, 2007.

Boyd, D. J., Lankford, H., Loeb, S., Rockoff, J. E., & 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. J., 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.

Chetty, R., Friedman, J. N., & Rockoff, J. E. (2011). *The long-term impacts of teachers: Teacher value-added and student outcomes in adulthood*. National Bureau of Economic Research.

Clotfelter, C., Ladd, H. F., & Vigdor, J. L. (2006). Teacher-student matching and the assessment of teacher effectiveness. *Journal of Human Resources, 41*(4), 778.

Clotfelter, C., Ladd, H. F., & Vigdor, J. L. (2007). Teacher credentials and student achievement: Longitudinal analysis with student fixed effects. *Economics of Education Review, 26*(6), 673-682.

Goldhaber, D., Gross, B., & Player, D. (2011). Teacher career paths, teacher quality, and persistence in the classroom: Are public schools keeping their best? *Journal of Policy Analysis and Management, 30*(1), 57-87.

Goldhaber, D., & Hansen, M. (2010). Is It Just a Bad Class? Assessing the Stability of Measured Teacher Performance. Center for Education Data and Research.

Grossman, P. L., Loeb, S., Cohen, J., Hammerness, K. M., Wyckoff, J., Boyd, D. J., & Lankford, H. (2010). Measure for Measure: The relationship between measures of instructional practice in middle school English Language Arts and teachers' value-added scores. *NBER Working Paper*.

Hanushek, E. A. (1971). Teacher characteristics and gains in student achievement: Estimation using micro data. *The American Economic Review, 61*(2), 280-288.

Hanushek, E. A., Kain, J., O'Brien, D., & Rivkin, S. (2005). The market for teacher quality. *NBER Working Paper*.

Hanushek, E. A., & Rivkin, S. G. (2010). Constrained Job Matching: Does Teacher Job Search Harm Disadvantaged Urban Schools? : National Bureau of Economic Research.

Hanushek, E. A., Rivkin, S. G., Figlio, D., & Jacob, B. (2010). Generalizations about using value-added measures of teacher quality. *American Economic Review, 100*(2), 267-271.

Harris, D. N., & Sass, T. R. (2011). Teacher training, teacher quality and student achievement. *Journal of Public Economics, 95*(7), 798-812.

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.

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. (2002). The promise and pitfalls of using imprecise school accountability measures. *Journal of Economic Perspectives, 16*, 91-114.

Kane, T. J., & Staiger, D. O. (2008). *Are Teacher-Level Value-Added Estimates Biased? An Experimental Validation of Non-Experimental Estimates*. Working Paper. Retrieved from http://isites.harvard.edu/fs/docs/icb.topic245006.files/Kane_Staiger_3-17-08.pdf

Kane, T. J., & Staiger, D. O. (2008). Estimating teacher impacts on student achievement: An experimental evaluation: National Bureau of Economic Research.

Kane, T. J., & Staiger, D. O. (2012). Gathering Feedback for Teaching, Measures of Effective Teaching Project: Bill and 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., & Betts, J. R. (2007). *Re-examining the role of teacher quality in the educational production function*. University of Missouri Department of Economics Working Paper, (708).

Koedel, C., & Betts, J. R. (2011). Does student sorting invalidate value-added models of teacher effectiveness? An extended analysis of the Rothstein critique. *Education Finance and Policy, 6*(1), 18-42.

McCaffrey, D. F., Lockwood, J., Koretz, D., Louis, T. A., & Hamilton, L. (2004). Models for value-added modeling of teacher effects. *Journal of Educational and Behavioral Statistics, 29*(1), 67.

McCaffrey, D. F., Sass, T. R., Lockwood, J., & Mihaly, K. (2009). The intertemporal variability of teacher effect estimates. *Education Finance and Policy, 4*(4), 572-606.

Milanowski, A. (2004). The relationship between teacher performance evaluation scores and student achievement: Evidence from Cincinnati. *Peabody Journal of Education, 79*(4), 33-53.

Murnane, R., & Phillips, B. (1981). What do effective teachers of inner-city children have in common?* 1. *Social Science Research, 10*(1), 83-100.

Nye, B., Konstantopoulos, S., & Hedges, L. V. (2004). How large are teacher effects? *Educational Evaluation and Policy Analysis, 26*(3), 237-257.

Ost, B. (2009). How Do Teachers Improve? The Relative Importance of Specific and General Human Capital.

Papay, J. P., & Kraft, M. A. (2011). Do Teachers Continue to Improve with Experience? Evidence of Long-Term Career Growth in the Teacher Labor Market. *Working Paper*.

Rivkin, S., Hanushek, E. A., & Kain, J. (2005). Teachers, schools, and academic achievement. *Econometrica, 73*(2), 417-458.

Rockoff, J. E. (2004). The impact of individual teachers on student achievement: Evidence from panel data. *American Economic Review, 94*(2), 247-252.

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

Taylor, E. S., & Tyler, J. H. (2011). The effect of evaluation on performance: Evidence from longitudinal student achievement data of mid-career teachers: National Bureau of Economic Research.

Weisberg, D., Sexton, S., Mulhern, J., & Keeling, D. (2009). The widget effect. *Brooklyn, NY: The New Teacher Project*.

Yoon, K. S. (2007). *Reviewing the evidence on how teacher professional development affects student achievement*: National Center for Educational Evaluation and Regional Assistance, Institute of Education Sciences, US Dept. of Education.