NATIONAL
CENTER *for* ANALYSIS *of* LONGITUDINAL DATA *in* EDUCATION RESEARCH

TRACKING EVERY STUDENT'S LEARNING EVERY YEAR

A program of research by the American Institutes for Research with Duke University, Northwestern University, Stanford University, University of Missouri-Columbia, University of Texas at Dallas, and University of Washington

Do First Impressions Matter? Improvement in Early Career Teacher Effectiveness

ALLISON ATTEBERRY,
SUSANNA LOEB AND
JAMES WYCKOFF

WORKING PAPER 90 • FEBRUARY 2013

Do First Impressions Matter? Improvement in Early Career Teacher Effectiveness

Allison Atteberry
University of Virginia

Susanna Loeb
Stanford University

James Wyckoff
University of Virginia

Contents

Acknowledgements

We appreciate helpful comments from Matt Kraft and Eric Taylor 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 to the American Institutes for Research. The research reported here was also supported by the IES Grant R305B100009 to the University of Virginia. The views expressed in the paper are solely those of the authors and may not reflect those of the funders. Any errors are attributable to the authors.

CALDER • American Institutes for Research
1000 Thomas Jefferson Street N.W., Washington, D.C. 20007
202-403-5796 • www.caldercenter.org

Do First Impressions Matter? Improvement in Early Career Teacher Effectiveness

Allison Atteberry, Susanna Loeb, and James Wyckoff

CALDER Working Paper No. 90
February 2013

Abstract

There is increasing agreement among researchers and policymakers that 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 as well as later life outcomes. However, there is not similar agreement about how to improve teacher effectiveness. Several research studies confirm that on average novice teachers show remarkable improvement in effectiveness over the first five years of their careers. In this paper we employ rich data from New York City to explore the variation among teachers in early career returns to experience. Our goal is to better understand the extent to which measures of teacher effectiveness during the first two years reliably predicts future performance. Our findings suggest that early career returns to experience may provide useful insights regarding future performance and offer opportunities to better understand how to improve teacher effectiveness. We present evidence not only about the predictive power of early value-added scores, but also on the limitations and imprecision of those predictions.

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. The early career period represents a unique opportunity to identify struggling teachers, examine the likelihood of future improvement, and make strategic pre-tenure dismissals to improve teacher quality. In this paper, we explore how teacher value-added measures in the first two years predict future teacher performance. In service of this larger goal, we pursue 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. We use panel data from the New York City Department of Education that follows all new

teachers who began between the 1999-00 and 2006-07 school years to pursue the following research questions:

- To what extent do teachers vary around the mean pattern in returns to experience? We examine the degree of variability in the developmental trajectories of teacher in terms of effectiveness in the early career.
- To what extent do teachers with different initial value-added scores in the first two years exhibit different returns to experience during the first five years, and how well do these initial scores account for variability in future performance?
- To what extent do predictions made based on early value-added scores mischaracterize teachers' future performance?

In what follows we provide some background for the relevance of this research question, as well as a review of existing literature that helps frame our question. We next describe the data from New York City used in the analysis, as well as the analytic approach used to answer these three research questions. We follow with the results organized by question, and conclude.

Background and Prior Literature

Research documents the substantial impact of assignment to a high-quality teacher on student achievement (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). We also know that teachers are not uniformly effective. The difference between effective and ineffective teachers has substantial effects on 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 do not acknowledge these disparities in quality, nor do most districts tailor their responses to or compensation for teachers based on performance. In the *Widget Effect,* Weisberg, Sexton, Mulhern, & Keeling, (2009) surveyed twelve large districts across four states and found that no measures of performance were taken into account in recruitment, hiring/ 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 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 often little useful 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 observational measures might be useful for improving the overall quality of the teacher workforce. In the field, policy makers rarely propose to use value-added scores as the exclusive metric for teacher evaluation. 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. In this paper we focus on value-added scores as illustrative of teacher quality measures more broadly, not because we believe that value-added scores should be used in isolation. Practically speaking, there are very few places where other measures of teacher effectiveness are readily available at this point to study a panel of teachers throughout their first five years.

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; 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.

It is worth noting that measuring reliability based on stability is potentially more problematic in the first five years of a teacher's career. Inherent in this approach to measuring reliability is the belief that the latent phenomenon of interest—true effectiveness— is not changing over time. Yet, on average, teachers undergo the largest improvements of their careers in the first few years of teaching. Researchers have generally documented a leveling off of returns to experience after five to seven years, suggesting that many teachers reach their own plateau, whatever that may be, during this early career period (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 this early phase, one might expect teacher quality measures to be less stable during this time as evidenced by year-to-year correlation for example. At the same time, these measures may well be reliable in the sense that the scores consistently reflect latent true quality as it develops and, in theory these scores may be just as predictive of future scores despite their instability.

[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).

That said, there are many 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. In this paper we explore 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.

Data

The backbone of the data that we use 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 that we need 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

We explore 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 we can observe their later performance. 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 our sample threatens the validity of our estimates because we 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, Lankford, Loeb, and Wyckoff, 2007). As a result, our primary analyses focus on the set of New York City elementary teachers who began between 2000 and 2006 who have, at a minimum, 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, we must observe teachers with 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 for whom we observe 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. 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 for whom we can be sure they 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, in this project we define 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]

[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

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

The outcome A_{itgsy} is the achievement of student i, with teacher t, in grade g, in school s, at time y, and we model this 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 have made several important analytic choices about the best specification for our purposes herein. 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] We attend 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 analysis to

[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).

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. Estimating Mean and Variance in Returns to Experience

We first estimate the mean returns to experience for teachers in the 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-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

[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.

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 teachers vary around this mean trend. In order to explore this, we randomly sample 100 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.

RQ 2. Performance in the Initial Years of Teaching as a Predictor of Future Effectiveness

Our second research question asks: To what extent do teachers with different initial value-added scores in the first two years exhibit different returns to experience during the first five years, and how well do these initial scores account for variability in future performance? To build off our work 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 throughout the first five years separately by quintiles of teachers' initial performance, and we examine the likelihood that teachers transition from low to high quintiles (and vice versa) during the course of their early career. Policy makers often translate raw evaluation scores into four or five performance groups in order to facilitate direct action for top and bottom performers. Because this practice is so ubiquitous, 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 non-trivial analytic decisions on our part and thus we delineate some of our challenges in Appendix A. These analytic choices are likely at play for policy-makers in the real world as well, and thus the discussion of our process may be instructive to this larger audience.)

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.

We are also interested in whether *any* initially high-performing teachers become among the lowest-performing teachers in the future (or vice versa). We therefore also 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.

It is worth noting that quintile groupings may obscure large 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. This approach allows us to consider the ability to predict future scores with initial scores without introducing quintile groupings into the analysis.

We use the following equation 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:

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

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. Examining Errors in Prediction

Finally, because we know that errors in prediction are inevitable, we present evidence on the degree of confidence in our predictions, and the nature of the miscategorizations one might make based on value-added scores from a teacher's first two years. We examine confidence intervals around forecasted future scores from the most promising specifications of Equation (4) above. In addition, 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 pay for performance, professional development, probation, dismissal, etc.). We then follow teachers into the following five 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.

Results

Mean and Variance in Early Career Improvement by Experience

Researchers consistently have found that, on average, teachers become more effective at improving student test performance during their first few 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 5 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 about 15 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, this estimate of 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 100 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), alongside the mean value-added scores (red) in the same period. This graph illustrates notable variability around the mean

[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).

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 0.20 across teachers in their first year (experience = 0). For math, the standard deviation of first-year teacher value-added is approximately 0.21. The variance in both ELA and math value-added scores steadily increase with experience so that the standard deviation in value added is at least 0.23 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 and, thus, that there is likely to be meaningful variation in returns to experience across teachers.

Performance in the Initial Years of Teaching as a Predictor of Future Effectiveness

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 continue 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 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. However, the mean trajectories by quintile shown in Figure 3 may obscure further important within-quintile variance. That is, it provides little information about movements across quintiles in the future. In Table 3, 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.[8] The majority—62 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 73 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 19 percent of bottom- and 10 percent of top- quintile initial performers end up in the opposite extreme two quintiles. Results are similar for ELA teaching.

Taken together, the transition matrix in Table 3 and 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. Finally, despite this variability, the transition matrix suggests that measures of value-added in the first two years predict of future performance for most teachers. We next pin down more carefully the extent to which initial performance can provide accurate and meaningful predictions about teacher performance in the future.

[8] 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.

In Table 4, we present adjusted R-squared values from a 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 8.9 percent of the variance in value-added scores in the third year. The predictive power is even lower for ELA (2.9 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 are unique in that they 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 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 27.8 percent of the variance in average future performance compared with no more than 19.4 percent of the variance in any of the individual future years. (For ELA, the comparable results are 20.9 percent and 15.4 percent.)

Table 4 shows early scores can explain up to a 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 (2.5 percent for ELA).[9] The measured teacher characteristics that district leaders 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.

Potential Errors in Categorizing Teachers

The prior analyses provide evidence that initial performance is predictive of later performance; however, the analyses also imply that this predictive ability is far from perfect. In this section we further describe the error associated with these predictions. To provide one 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

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

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. 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.

A policy that uses value-added scores to group teachers based on performance will likely 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.[10] For each group, we have added two reference points. 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 4, 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.

[10] 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.

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 lowest (red) quintile 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, most 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 appear at the very top of the red distribution of future performance— do surpass the mean of the middle quintile).

Figure 5 also allows us to compare the distribution of initially lowest quintile math teachers to the average teacher in the first year of experience (yellow vertical line), as this is the expected performance of a teacher with whom one could replace a dismissed teacher. It turns out that 68.9 percent of math teachers do not exceed the comparison to the average first year teacher (66.6 percent for ELA). In addition, 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.

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 clarity, think of this as assuming a teacher is at least average). In this case, a Type I error is rejecting a true null hypothesis, which is to falsely identify a teacher as low-performing when she turns out to be at least average in the long run. This type of error typically dominates the value-added debate, because this error negatively and unfairly penalizes teachers who

would be dismissed even though they *would have* emerged as effective over time. On the other hand, Type II error is often overlooked even though it likely affects students' instructional experiences. In the case of Type II error, one fails to reject a false null hypothesis, which 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 perform poorly in the future who were not identified as low-performing based on initial performance. Depending on the definition of ineffective, 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 by an average new teacher.

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* excellent in the long run. Type I error is rejecting a true null hypothesis—thinking 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 error in this context impacts teachers. To the extent that by identifying excellent teachers schools can improve their quality of instruction, Type I error, in this context, impacts students.

In practice, identifying Type I and Type II errors is complex, in part because it requires a clear criterion for identifying future "ineffectiveness" or "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 ameliorated to some extent the measurement error in a teacher's value-added measure in a given year by (1) using Bayes shrunk estimates which attenuates extreme measures in proportion to their imprecision, (2) averaging across multiple future years to lessen the influence of any one outlier result, and (3) breaking effectiveness into quintiles, so that while teachers in the middle quintiles may be less

distinct, one can focus on teachers at the extremes of future performance using top and bottom quintiles. We, however, do not address the narrowness of the value-added measure, nor its relative nature.

Figure 6 helps to illustrate Type I and Type II error associated with identifying teachers as ineffective, perhaps for the purpose of dismissal. In practice, this same approach could be used for any number of strategic policy responses such as allocating additional support, mentoring, observation, or professional development. Simply for the clarity of the example, we describe a dismissal policy. We start by translating the mean value-added scores of teachers in years one and two into percentiles. Moving from left to right along the x-axis represents an increase in the percentage of teachers who are identified as ineffective, and as a result might be dismissed. The y-axis gives the corresponding percent from each of the top and bottom three *future* deciles (separate lines for each decile) that would be dismissed based on the x-axis value. For example, a vertical line at 10 would give the percent of each of the future deciles that would be identified as ineffective if we were too identify the 10 percent of the lowest value-added teachers in the first two years as ineffective.

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 very low level (e.g. the 5^{th} percentile), most of the teachers identified end up in the bottom part of the distribution of future performance. Second, not surprisingly, the errors get bigger as we aim to identify a higher proportion of teachers as ineffective. For example, a substantial portion of teachers in the bottom 50 percent of initial value added end up in the top three deciles of future value-added.

To make the example more concrete, consider a hypothetical dismissal policy of the bottom ten percent of teachers in initial value-added. In this case, we are attempting to test a hypothesis about whether a teacher will be ineffective or not (the null hypothesis). We see that this policy would eliminate 29.5 percent of teachers who would subsequently appear in the lowest decile of future

performance and another 22.1 percent of teachers who would appear in the second lowest decile. In contrast, none of the top decile of future performance would be (falsely) identified and only two percent of the second highest decile would be (falsely) identified. The latter two numbers can also be thought of as a quantification of the Type I error—teachers who were identified as low performing by the policy but ultimately appeared to be among the highest performers in the future.

Figure 6 also illustrates Type II error. At the ten percent threshold, while 29.5 percent of the lowest decile teachers would have been dismissed, the other 70.5 percent of the lowest decile were not (fail to reject a false null). If one believes that the bottom ten percent of the distribution of performance in years three through five is a good criteria for ineffectiveness, then the failure to identify these teachers can be viewed through the lens of Type II error.[11] As one moves to the right on the x-axis, dismissing a larger proportion of teachers based on initial value-added, these tradeoffs balance one another. At 20 percent dismissal rate, one loses half (51.5 percent) of the future bottom decile in math (and fails to eliminate the other half of that quintile), while the relative "cost" is 6.8 percent of the top decile.[12]

One could argue about the appropriate criteria for future effectiveness. Another reasonable assertion might be to characterize every teacher who is significantly less effective than an average teacher and then retained as a Type I error, and every teacher who becomes significantly more effective than an average teacher who is accidentally dismissed as a Type II error (a more extreme interpretation of these results). We are agnostic about what should be used by policy makers in practice as the "right" criteria.

[11] Of course, the ability to eliminate a large percentage of the bottom deciles of future performance is capped by the percentage of teachers one is willing to fire. Put more concretely, if one adopts a policy of firing the bottom five percent of teachers after the first two years, even a "perfect" measure could only dismiss at most 50 percent of the bottom decile (i.e., 5 percent of the whole sample equals 50 percent of one decile).

[12] It is worth noting that, at some point, firing an unreasonably high percentage of teachers may trigger a general equilibrium problem, and the assumption that there is a continuous supply of "average" new teachers will no longer be true. The further to the right we move along the x-axis in Figure 6, not only is the likelihood of making type I errors much greater, but the likelihood of encountering a shortage of qualified teachers also increases.

Conclusions

From a policy perspective, the ability to predict future performance is practically most important for inexperienced teachers because policies that focus on development (e.g. mentoring programs), dismissal, and promotion are likely most relevant during this period. Prior work has documented the relationship between a teacher's value-added in one year and his or her value-added in future years. These analyses have been based on teachers without any restriction based on teaching experience. However, there is reason to believe that the relationship between current performance and future performance might be different for novice teachers than for other teachers. In particular, substantial evidence suggests that on average teachers improve more (that is, change their performance more) over the first years of teaching than over subsequent years.

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 their educational experience. 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 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 any 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.

We see three immediate strands coming out of the work completed herein. First, we will expand our existing analysis to middle school teachers. There are reasons to believe that the training, structure, and organization of middle schools might produce a different growth experience than observed in the elementary teacher population. Indeed some preliminary work suggests that the relationship between initial and future performance may be less straight forward in higher grades.

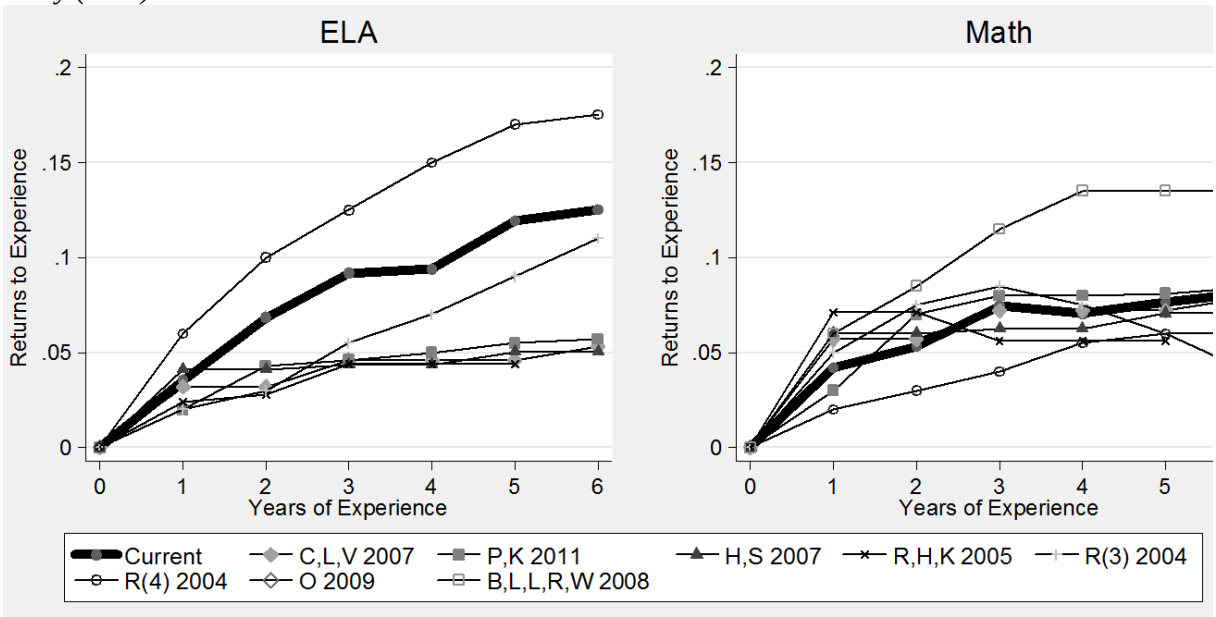
The second next step that arises from this work is to examine potential causes for the notable variability in growth rates in the early career. While the most effective teachers tend to remain the most effective and the least effective remain among the least effective, Figure 2 depicts a wide range of developmental patterns across the teachers in the first five years. Moreover, even when we break the mean value-added score trajectories over time into quintiles, there is undoubtedly important within-quintile variation. That is, even among the initially least effective teachers, some make up more ground than others. In future work we seek to identify correlates of teachers' growth over this time period. Our interest in this work is piqued by a variance decomposition of the growth in teacher effectiveness over the first five years of teaching indicating that 30 percent of the variance lies between schools, and 70

percent within schools. In our larger dataset, we observe a great deal about how teachers were trained, measures of their generic and teaching abilities, educational attainment details, and pathways into teaching. Further, once teachers begin teaching, they are undoubtedly influenced by (for better or worse) the organizational nature of the schools to which they are assigned, their colleagues, school leaders, and opportunities for professional development. For a few cohorts of New York City teachers, we can also look more deeply at the experiences of new teachers using in-depth survey data for teachers who have recently completed their first year. Work in this area is intended to help district leaders and policy makers understand how new teacher experiences might be modified to improve the quality of the existing teacher workforce.

Finally, we are interested in an observation that arose as an artifact of trying to follow teachers across multiple years with value-added scores: Of the 5,516 elementary math teachers who began teaching in or after the 1999-2000 school year and were present in the teacher database for at least their first five years, only 842 (about 15.3 percent) received value-added scores in every year. Some preliminary work suggests to us that teachers who possessed more value-added scores during their early career tended to be somewhat higher-performing in their initial year. Certainly there are a number of reasons that could account for missing value-added scores—e.g., switching to a non-tested subject or grade, insufficient numbers of tested students in a given year, leaves of absence. It is also possible that some of those explanations could be systematic or strategic on the part of teachers and principals. While that behavior is in itself of interest to those who wish to understand how teachers and schools might respond to evaluation policies, it is also interesting to note that we can evaluate such policies only for teachers who have at least some minimum amount of consistent evidence about how they perform over time. To the extent that this represents a somehow selective sample, the conclusions we reach about these policies may be less generalizable to all teachers. We think that examining the nature of the data patterns that arise in a district like New York might be instructive to the larger field.

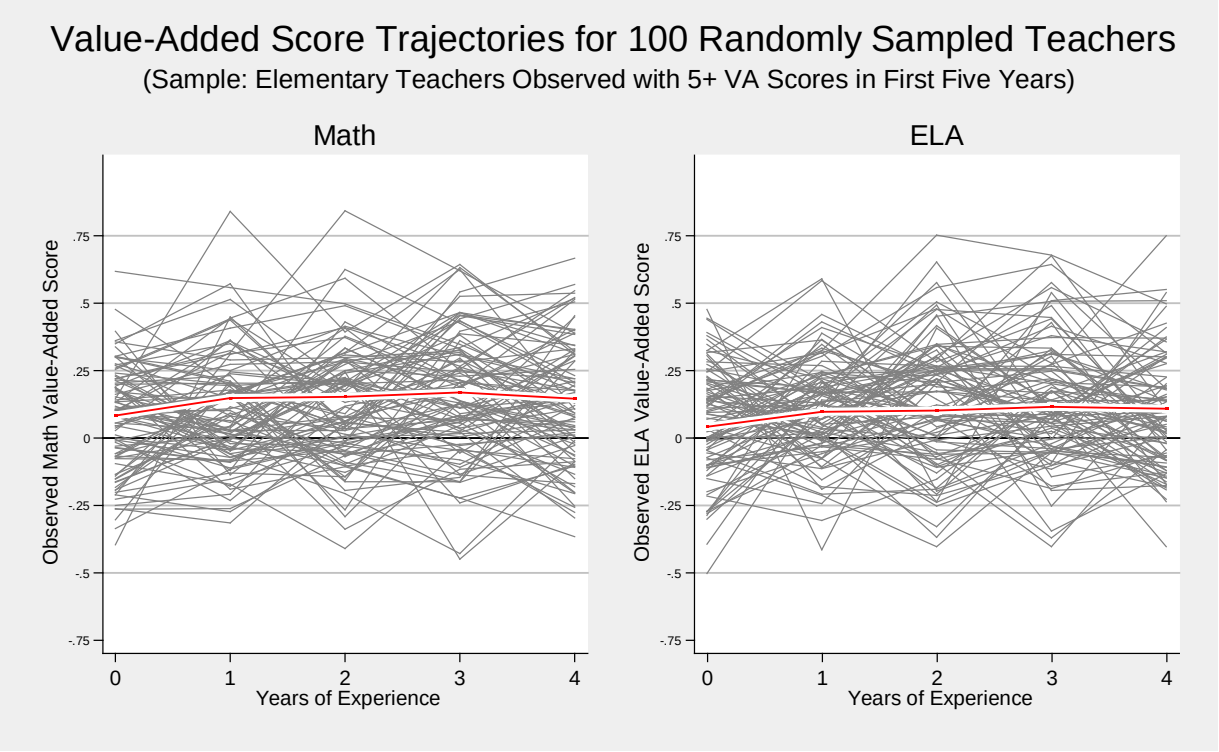
Figures

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	0.215	0.231	0.236	0.242	0.240	0.204	0.214	0.222	0.228	0.229
100 Teachers	0.211	0.232	0.230	0.243	0.241	0.192	0.204	0.220	0.231	0.230

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.

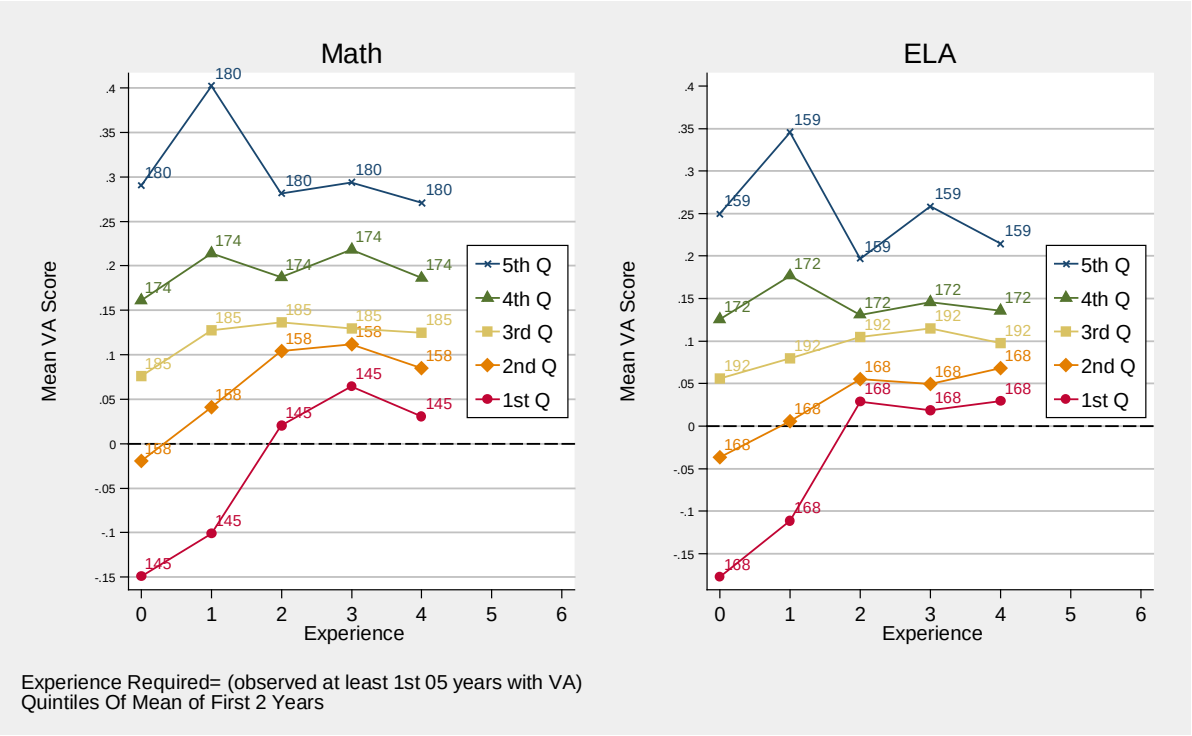
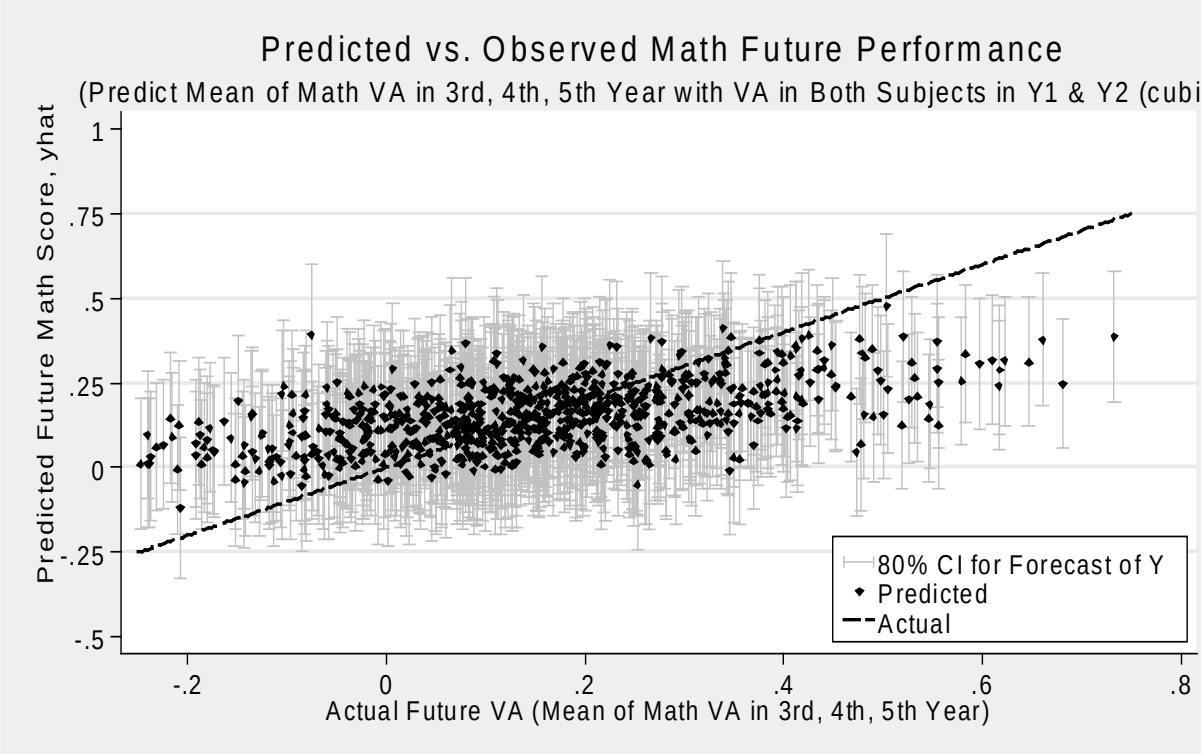


Figure 4:
Predicted Future Value-Added Scores (Mean of Years, 3,4, and 5) based on Observed Valued-Added Scores in Years 1 and 2, by Actual Future Value-Added Scores, with 80% Confidence Intervals Around Individual Predictions.



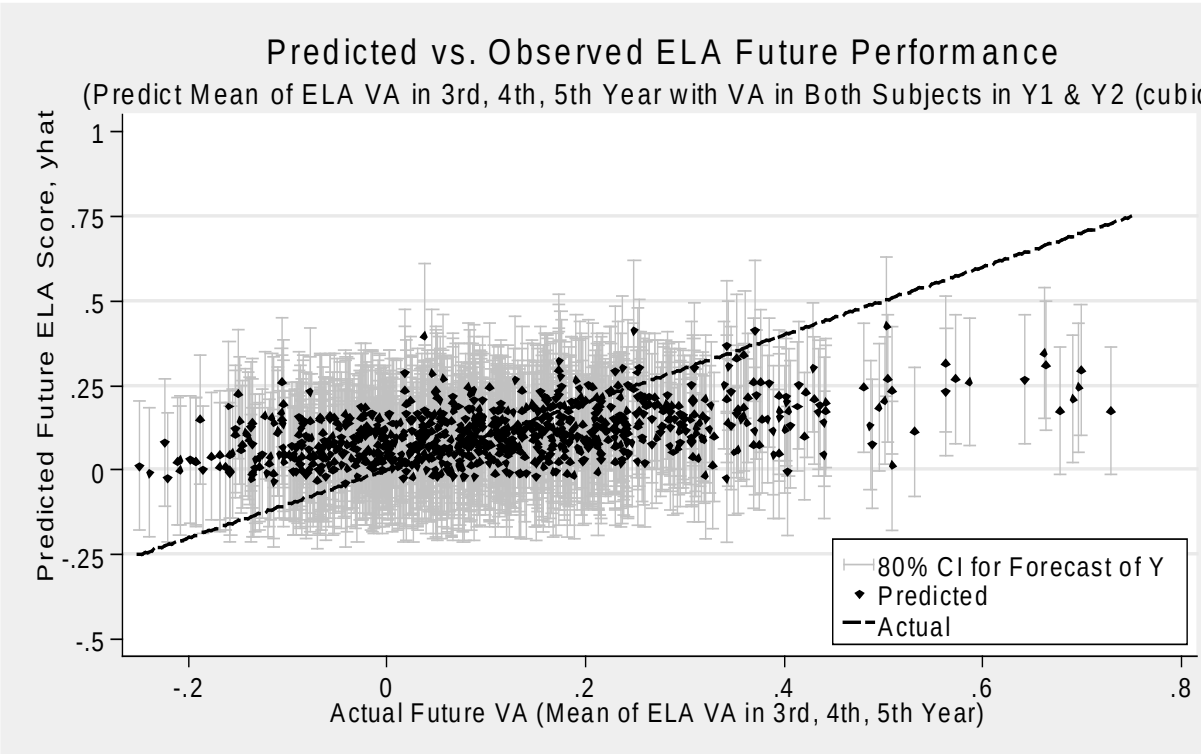
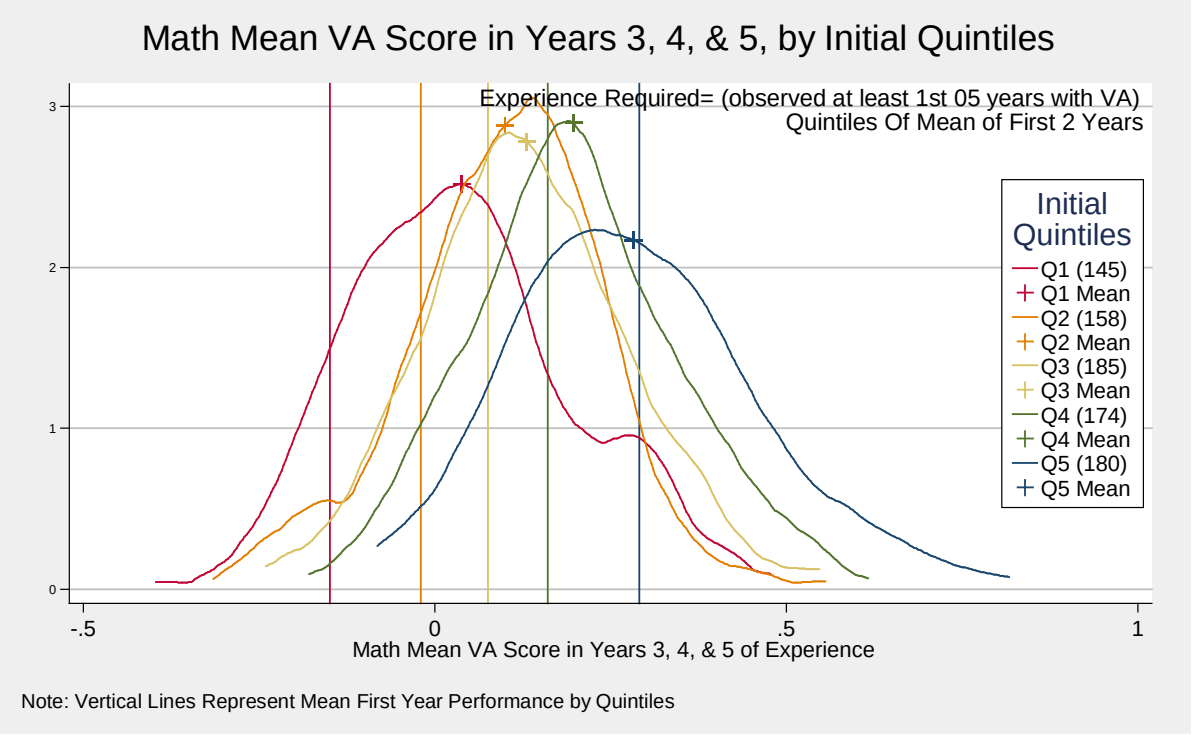
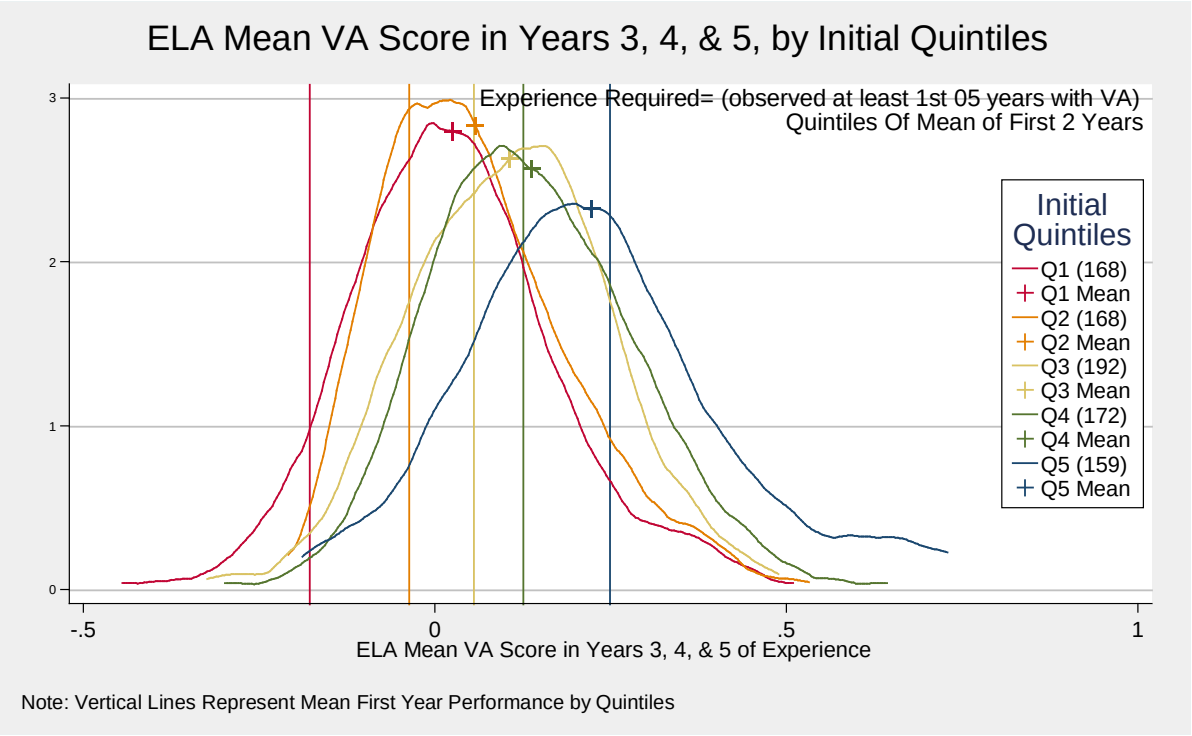


Figure 5:
Distribution of Future Value-Added Scores, by Initial Quintile of Performance

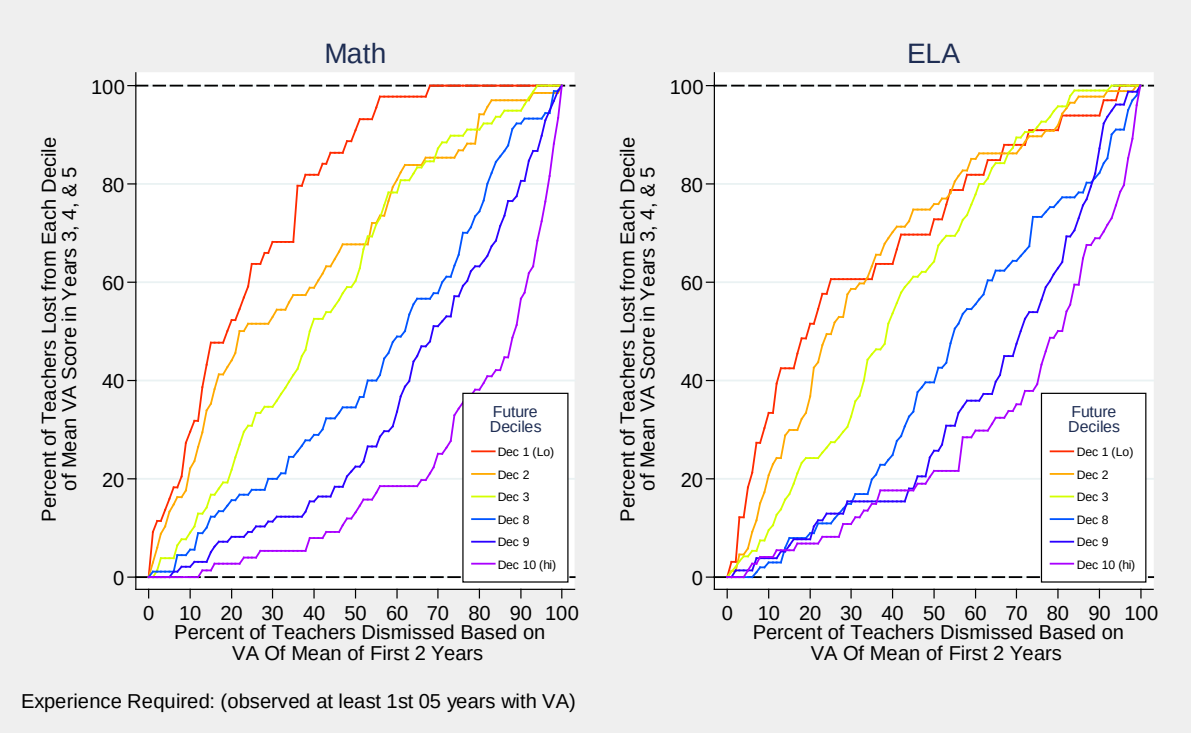


Note: Vertical Lines Represent Mean First Year Performance by Quintiles



Note: Vertical Lines Represent Mean First Year Performance by Quintiles

Figure 6:
Departures by Future Performance Quintile Based on Early Career Performance



Tables

Table 1:
Analytic Sample Sizes by Cumulative Restrictions

	Math		ELA	
	# Tchrs	**# Obs**	**# Tchrs**	**# Obs**
All Grade 4-8 Teachers Tied to Students in NYC since 2000	16,909	45,979	17,607	47,753
Started Teaching in 2000- 2006	13,355	39,367	13,942	41,041
Modal Grade in First Five Years is Grade 4 or 5	7,656	24,219	7,611	24,282
In HR Dataset for At Least 5 Years	5,516	20,790	5,482	20,860
Has VA Score in At Least 1st Year	4,170	14,085	4,180	14,226
Has VA in 1st and at Least 2 of Next 4 Years	2,068	10,853	2,073	10,967
Has VA in At Least Years 1 thru 3	1,792	9,544	1,798	9,642
Has VA in At Least Years 1 thru 5	842	5,685	859	5,822
Has VA in At Least Years 1 thru 7	329	2,780	346	2,918
Has VA in At Least Years 1 thru 9	135	1,324	139	1,362

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	104	158	207	219	154
	mean	-0.114	0.010	0.099	0.180	0.310
... After Limiting to Teachers in First Year (2)	n	51	125	165	242	259
	mean	-0.204	-0.069	0.031	0.132	0.306
… And Limiting to Elementary Teachers (3)	n	145	158	185	174	180
	mean	-0.125	0.011	0.102	0.188	0.346
… And Limiting to Teachers with 5+ VA score (4)	n	169	168	169	168	168
	mean	-0.112	0.028	0.113	0.196	0.354
ELA Quintiles…						
… of All Teacher-Years (1)	n	137	171	185	235	131
	mean	-0.107	-0.022	0.066	0.139	0.253
... After Limiting to Teachers in First Year (2)	n	81	127	179	236	236
	mean	-0.201	-0.079	0.008	0.100	0.258
… And Limiting to Elementary Teachers (3)	n	168	168	192	172	159
	mean	-0.144	-0.015	0.068	0.151	0.298
… And Limiting to Teachers with 5+ VA score (4)	n	172	172	172	172	171
	mean	-0.142	-0.012	0.067	0.145	0.291

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. Quintile Transition Matrix from Initial Performance to Future Performance, by Subject (Number, Row Percentage, Col Percentage)

		Quintile of Future Performance					
Math Initial Quintile		Q1	Q2	Q3	Q4	Q5	Row
Q1	n	53	37	28	17	10	145
	(row %)	(36.6)	(25.5)	(19.3)	(11.7)	(6.9)	
	(col %)	(47.3)	(23.4)	(14.1)	(8.5)	(5.7)	
Q2	n	23	37	49	38	11	158
	(row %)	(14.6)	(23.4)	(31.0)	(24.1)	(7.0)	
	(col %)	(20.5)	(23.4)	(24.7)	(19.0)	(6.3)	
Q3	n	22	45	50	42	26	185
	(row %)	(11.9)	(24.3)	(27.0)	(22.7)	(14.1)	
	(col %)	(19.6)	(28.5)	(25.3)	(21.0)	(14.9)	
Q4	n	10	25	40	55	44	174
	(row %)	(5.7)	(14.4)	(23.0)	(31.6)	(25.3)	
	(col %)	(8.9)	(15.8)	(20.2)	(27.5)	(25.3)	
Q5	n	4	14	31	48	83	180
	(row %)	(2.2)	(7.8)	(17.2)	(26.7)	(46.1)	
	(col %)	(3.6)	(8.9)	(15.7)	(24.0)	(47.7)	
Column Total		112	158	198	200	174	**842**
		Quintile of Future ELA Performance					
ELA Initial Quintile		Q1	Q2	Q3	Q4	Q5	Row
Q1	n	49	45	40	23	11	168
	(row %)	(29.2)	(26.8)	(23.8)	(13.7)	(6.5)	
	(col %)	(40.8)	(23.7)	(20.0)	(11.7)	(7.2)	
Q2	n	33	54	39	28	14	168
	(row %)	(19.6)	(32.1)	(23.2)	(16.7)	(8.3)	
	(col %)	(27.5)	(28.4)	(19.5)	(14.2)	(9.2)	
Q3	n	19	43	48	57	25	192
	(row %)	(9.9)	(22.4)	(25.0)	(29.7)	(13.0)	
	(col %)	(15.8)	(22.6)	(24.0)	(28.9)	(16.4)	
Q4	n	9	37	45	45	36	172
	(row %)	(5.2)	(21.5)	(26.2)	(26.2)	(20.9)	
	(col %)	(7.5)	(19.5)	(22.5)	(22.8)	(23.7)	
Q5	n	10	11	28	44	66	159
	(row %)	(6.3)	(6.9)	(17.6)	(27.7)	(41.5)	
	(col %)	(8.3)	(5.8)	(14.0)	(22.3)	(43.4)	
Column Total		120	190	200	197	152	**859**

Note: Initial quintiles are constructed by first restricting the sample to grade four and five teachers and then identifying five equally sized groups based on a teacher's mean value-added score in her first two years. The quintiles of future performance are constructed by first restricting the sample to grade four and five teachers and then identifying five equally-sized groups based on a teacher's mean value-added score in years three, four, and five. The sample is subsequently restricted to teachers with value-added scores in at least the first five years.

Table 4:
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_{Y3-5})
Math				
Math VA in Y1 Only	0.089	0.052	0.070	0.109
Math VA in Y2 Only	0.153	0.165	0.141	0.241
Math VA in Y1 & Y2	0.178	0.171	0.158	0.265
VA in Both Subjects in Y1 & Y2	0.179	0.188	0.166	0.277
VA in Both Subjects in Y1 & Y2 (cubic)	0.175	0.194	0.172	0.278
ELA				
ELA VA in Y1 Only	0.029	0.049	0.023	0.064
ELA VA in Y2 Only	0.062	0.114	0.069	0.154
ELA VA in Y1 & Y2	0.075	0.135	0.077	0.181
VA in Both Subjects in Y1 & Y2	0.090	0.145	0.087	0.203
VA in Both Subjects in Y1 & Y2 (cubic)	0.094	0.154	0.086	0.209

Table 5. Year-to-Year Correlations, by Subject and Experience

	Y+0	Y+1	Y+2	Y+3	Y+4	Y+5
Math						
All Teachers	1.000	0.436	0.386	0.343	0.308	0.291
New Teachers	1.000	0.373	0.328	0.288	0.246	0.175
ELA						
All Teachers	1.000	0.327	0.291	0.247	0.223	0.239
New Teachers	1.000	0.230	0.181	0.168	0.145	0.167

Note: The table presents pairwise correlations between value-added scores in a given year (Y) and the subsequent year (Y+1), two years later (Y+2), … , five years later (Y+5). We do this for two samples of teachers—the full sample of teachers who teach elementary grades in New York City (without regard to years of experience), and a subsample of teachers who began their career in year Y.

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:

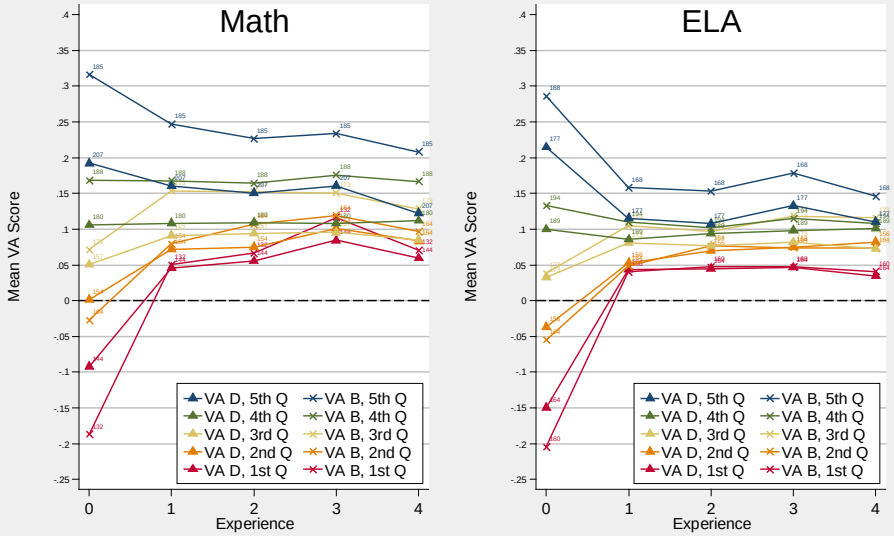
(A) We examine results across two teacher samples based on minimum value-added required for inclusion. The first figure uses the analytic sample used throughout the main paper—teachers with value-added scores in at least all of their first five years. The second widens the analytic sample to the set 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.

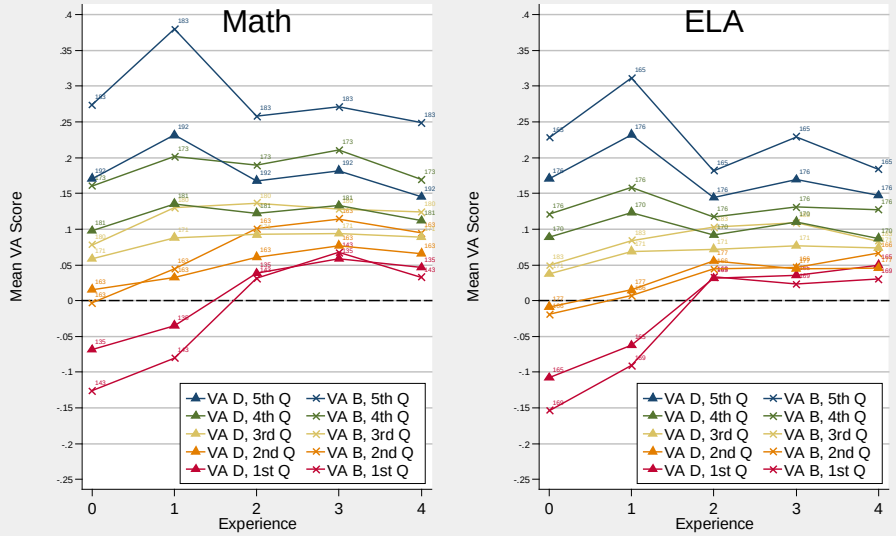
(C) Finally, we examine results using two alternative value-added models to the one used in the paper. "VA Model B" uses a gain score approach rather than the lagged achievement approach used in the paper. "VA Model D" differs from the main value-added model described in the paper in that it uses student-fixed 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

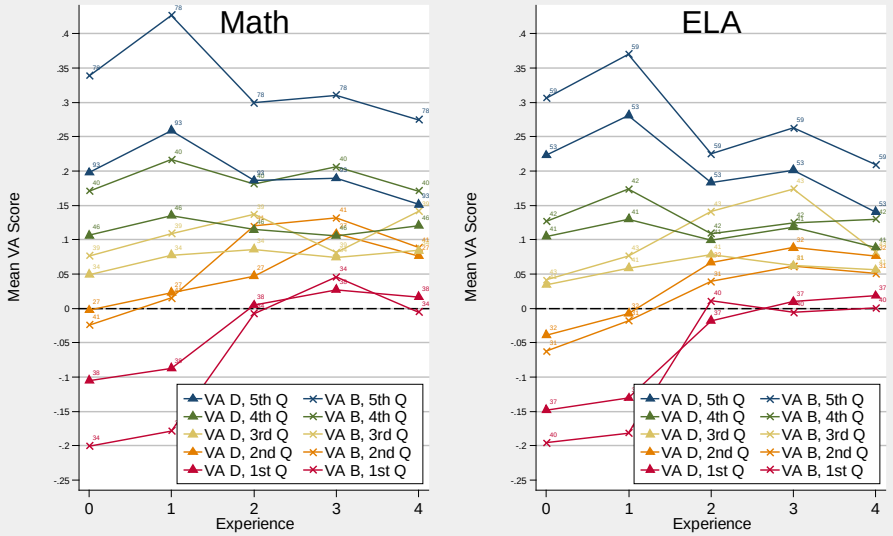
Quintile Among All 1st-Yr Tchrs



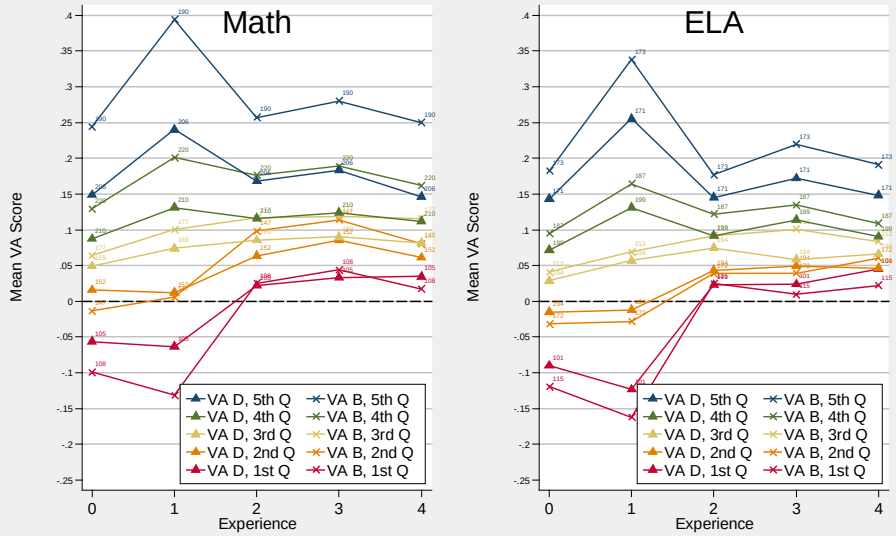
Quintile Of Mean of First 2 Years



Quintile Consistent in First Two Years

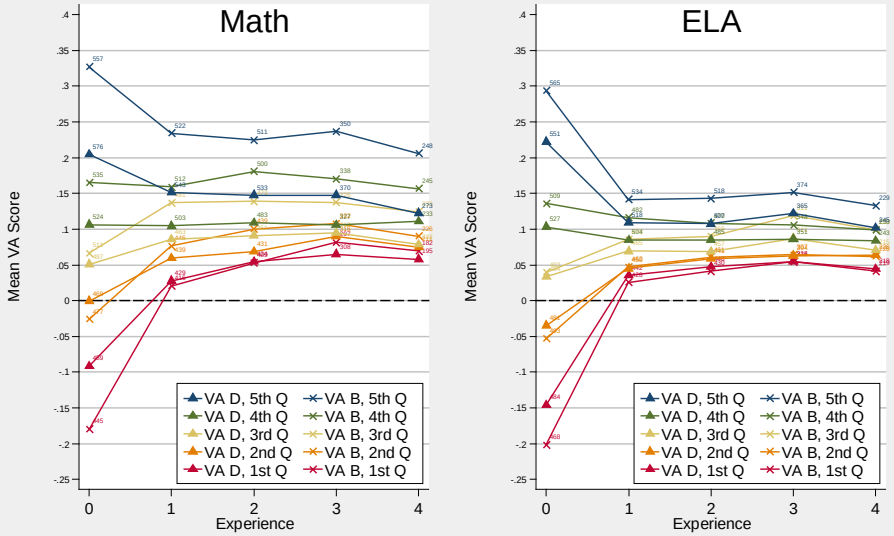


Quintile Of Mean of Y1, Y2, Y2

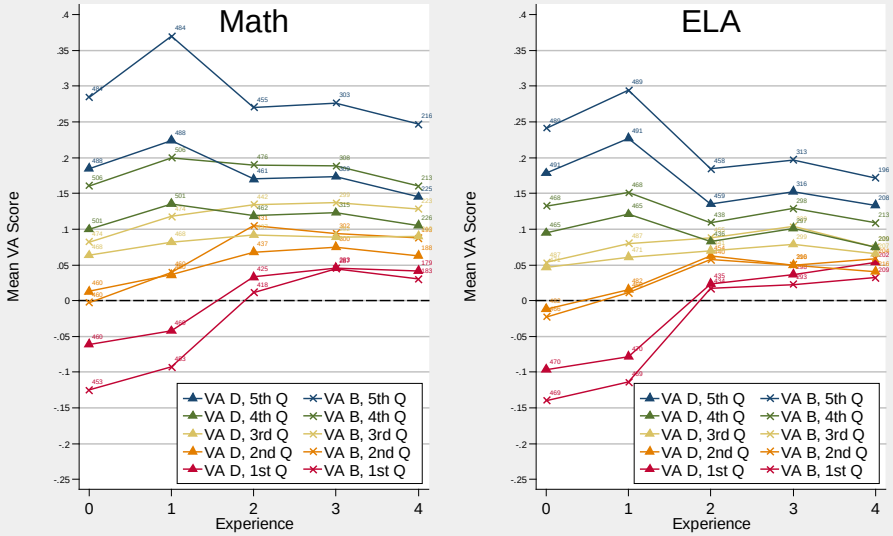


Elem Teachers with VAM in 1st, and 2 of Next 4 Years

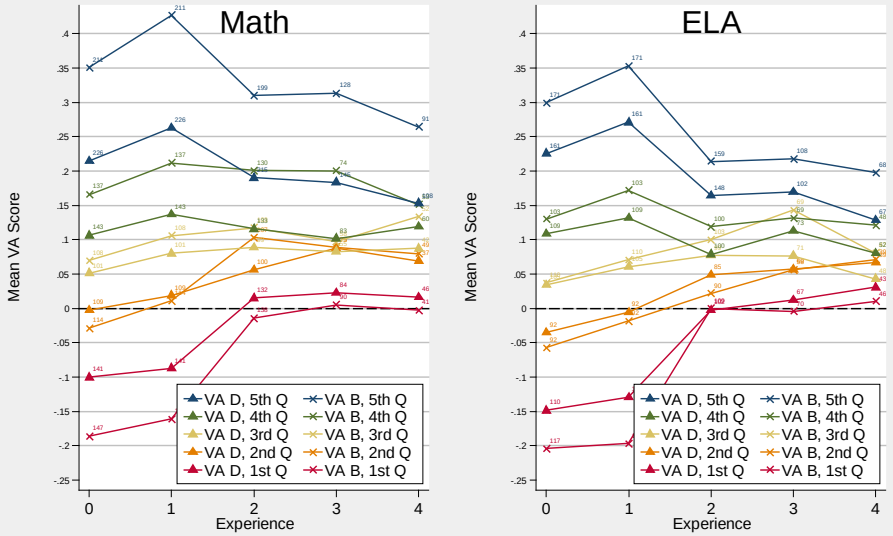
Quintile Among All 1st-Yr Tchrs



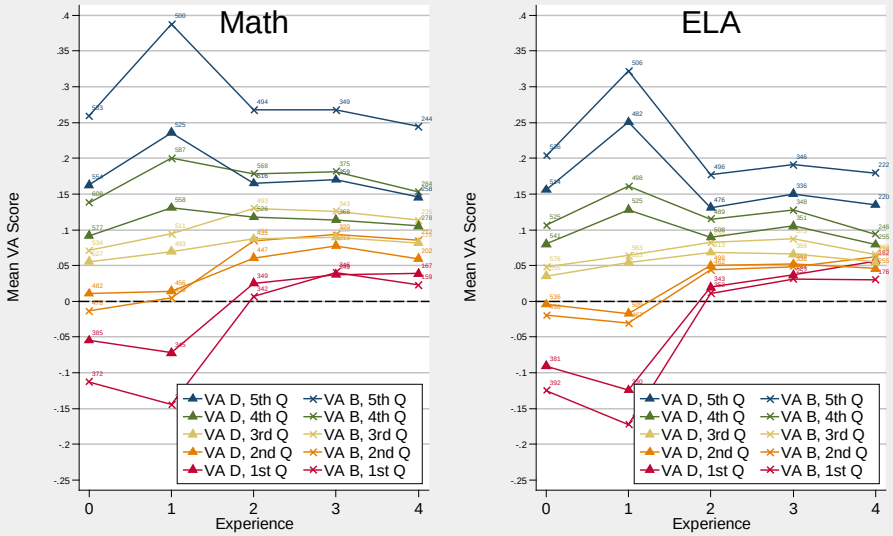
Quintile Of Mean of First 2 Years



Quintile Consistent in First Two Years



Quintile Of Mean of Y1, Y2, Y2



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.

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., & 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.