AERA Open
October-December 2015, Vol. 1, No. 4, pp. 1–23
DOI: 10.1177/2332858415607834

Do First Impressions Matter? Predicting Early Career Teacher Effectiveness

Allison Atteberry
University of Colorado, Boulder
Susanna Loeb
Stanford University
James Wyckoff
University of Virginia

As educational policy makers seek strategies to improve the teacher workforce, the early career period represents a unique opportunity to identify struggling teachers, examine the likelihood of future improvement, and make strategic pretenure investments in development or dismissals. It is also a useful time to identify particularly promising teachers for development and focus on high-needs areas. This article asks how much teachers vary in performance improvement during their first 5 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. This is particularly the case in math, while predictions about future English language arts (ELA) performance based on initial ELA value added are less precise. Predictions are most powerful at the extremes. We use these predictions to explore the likelihood that personnel actions based on initial performance would lead to inappropriate distinctions between teachers who would be high or low performing in future years. We also examine the much less discussed costs of failure to distinguish performance when meaningful differences exist. The results point to the potential of policies that make use of teachers' initial performance to inform personnel decisions.

Keywords: *policy makers, school districts, teaching effectiveness, value added*

Educational policy makers in most schools and districts face considerable pressure to improve student achievement. Principals and teachers recognize, and research confirms, that teachers vary considerably in their ability to improve student outcomes (Rivkin, Hanushek, & Kain, 2005; Rockoff, 2004). 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 teaching effectiveness, including but not limited to 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 (Dee & Wyckoff, 2015), to take on more challenging assignments, or to serve as models of expert practice (Glazerman & Seifullah, 2012). Others attempt to identify struggling teachers in need of mentoring or professional development to improve skills (Taylor & Tyler, 2011; Yoon, 2007). Because some teachers may never become effective, some researchers and policy makers are exploring dismissals of ineffective teaches as a mechanism for improving the teacher workforce (Boyd, Lankford, Loeb, Ronfeldt, & Wyckoff, 2011; Goldhaber & Theobald, 2013; Winters & Cowen, 2013).

Interest in measuring teacher effectiveness persists throughout teachers' careers but is particularly salient during the first few years when potential benefits are greatest. Attrition of teachers is highest during these years (Boyd, Grossman, Lankford, Loeb, & Wyckoff, 2008), and the ability to reliably differentiate more effective from less effective teachers would help target retention efforts. Moreover, less effective, inexperienced teachers may be able to sufficiently improve to become more effective than those with more experience. Targeting professional development to these teachers early allows benefits to be realized sooner and thus influence more students. Finally, nearly all school districts review teachers for tenure early in their careers (many states make this determination by the end of a teacher's third year). Tenure decisions can be more beneficial for students if measures of teaching effectiveness are considered in the process (Loeb, Miller, & Wyckoff, 2014).

The benefits to policy makers of early identification of teacher effectiveness are clear; the ability of currently available measures to accurately do so is much less well understood. Indeed, teachers often voice doubts about school and district leaders' ability to capture teacher effectiveness using admittedly crude measures such as value-added scores, intermittent observations, and/or principal evaluations. Their concerns are understandable, given that value-added scores are imprecise and districts are increasingly experimenting with linking important employment decisions to such measures, especially in the first few years of the career. A well-established literature examines the predictive validity of teacher value added for all teachers, which suggests that there is some useful signal among the noise, but the measures are imprecise for individual teachers (see, e.g., McCaffrey, 2012). Somewhat surprisingly, there is very little research that explores the predictive validity of measures of teacher effectiveness for early career teachers, despite good reasons to believe the validity would differ by experience.

In this article, we use value-added scores as one example of a measure of teaching effectiveness. We do so not because value-added measures capture all aspects of teaching that are important or because we think that value-added measures should be used in isolation. In fact, virtually all real-world policies that base personnel decisions on measures of teaching effectiveness combine multiple sources of information, including classroom observational rubrics, principal perceptions, or even student and parent surveys. Districts tend to use value-added measures (in combination with these other measures) when available, and because value-added scores often vary more than other measures, they can be an important component in measures of teaching effectiveness (Donaldson & Papay, 2015; Kane, McCaffrey, Miller, & Staiger, 2013). We focus on value-added scores in this article as an imperfect proxy for teaching effectiveness that is being used by policy makers today. Understanding the properties of value added for early career teachers is relevant in this policy context.

Measured value added for novice teachers may be more prone to random error than for more experienced teachers as their value-added estimates are based on fewer years of data and fewer students. Moreover, novice teachers on average tend to improve during the first few years of their careers, and thus their true effectiveness may change more across years than that for more experienced teachers. Figure 1 depicts returns to experience from eight studies, as well as our own estimates using data from New York City.[1] 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% of a standard deviation of student achievement higher than when he or she was a first-year teacher.[2] However, little is known about the *variability* of early career returns to experience. If some teachers with similar initial performance improve substantially and others do not, early career effectiveness measures will be weak predictors of later performance.

This article explores how well teacher performance, as measured by value added over a teacher's first 2 years, predicts future teacher performance. Toward this end, we address the following two research questions: (1) Does the ability to predict future performance differ between novice and veteran teachers? (2) How well does initial job performance predict future performance? We conclude the article with a more in-depth exploration of the policy implications and trade-offs associated with inaccurate predictions.

This article makes several contributions to existing literature on the use of measures of teaching effectiveness. Although an existing literature documents the instability of value added (see, e.g., Goldhaber & Hansen, 2010a; Koedel & Betts, 2007; McCaffrey, Sass, Lockwood, & Mihaly, 2009), that literature largely does not distinguish between novice and veteran teachers, and when it does (Goldhaber & Hansen, 2010b), the focus is specifically on tenure decisions. We build on this work, showing that value-added over the first 2 years is less predictive of future value added than later in teachers' careers. Nonetheless, there is still signal in the noise; early performance is predictive of later performance. We also develop and illustrate a policy-analytic framework that demonstrates the trade-offs of employing imprecise estimates of teacher effectiveness (in this case, value added) to make human resources policy decisions. How policy makers should use these measures depends on the policy costs of mistakenly identifying a teacher as low or high performing when the teacher is not versus the cost of not identifying a teacher when the identification would be accurate.

Background and Prior Literature

Research documents substantial impact of assignment to a high-quality teacher on student achievement (Aaronson, Barrow, & Sander, 2007; Boyd, Lankford, Loeb, Ronfeldt, et al., 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 proximal outcomes like standardized test scores, as well as distal outcomes such as college attendance, wages, housing quality, family planning, and retirement savings (Chetty, Friedman, & Rockoff, 2011).

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

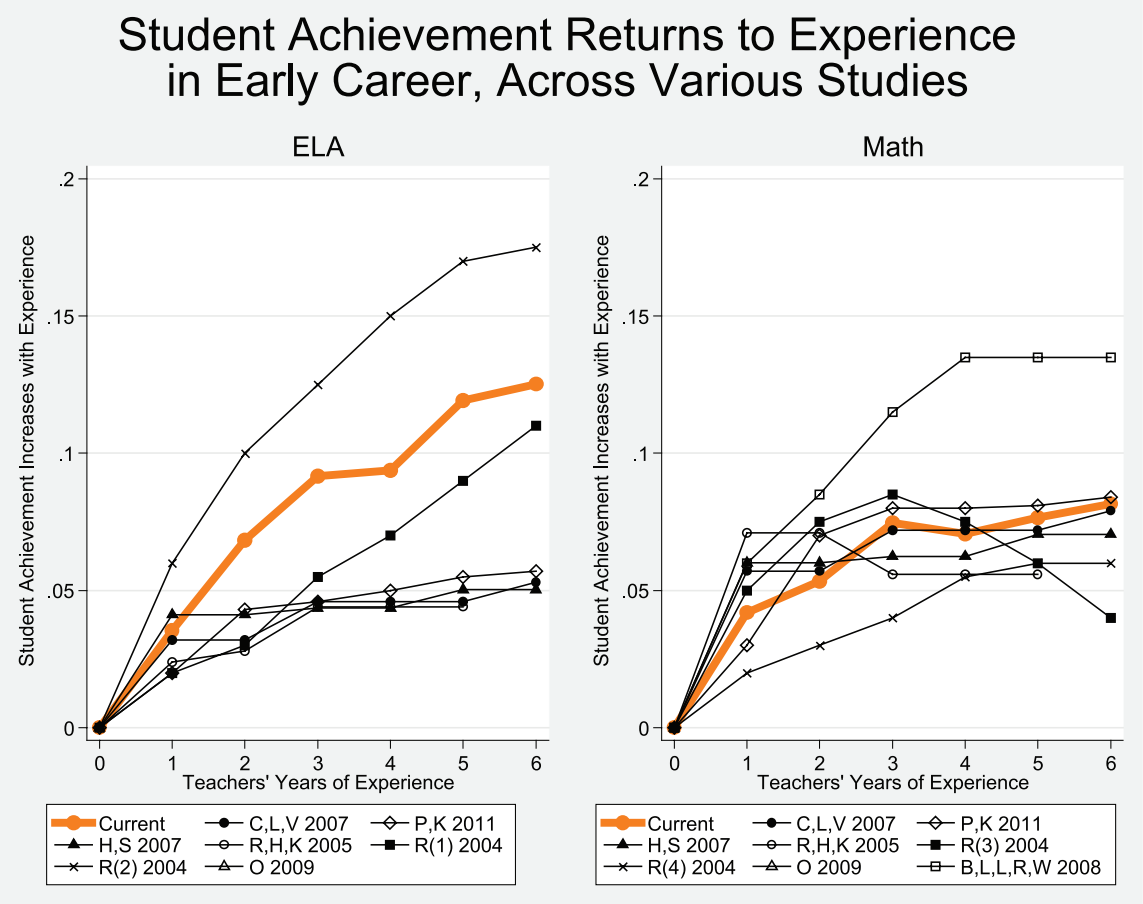


FIGURE 1. *Student achievement returns to teacher early career experience, preliminary results from current study (bold) and various other studies. Results are not directly comparable due to differences in grade level, population, and model specification, but Figure 1 is intended to provide some context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

conjunction with other evidence from the classroom, such as observational protocols or principal assessments, to inform policy discussions aimed at improving teaching.

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 used a test-retest reliability perspective, 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 across proximal years, reasoning that a reliable measure should be consistent with itself from one year to the next (Aaronson et al., 2007; Goldhaber & Hansen, 2010a; Kane & Staiger, 2002; Koedel & Betts, 2007; McCaffrey et al., 2009). When value-added scores fluctuate dramatically in adjacent years, this presents a policy challenge—the measures may reflect statistical imprecision (noise) more than true teacher performance. In this sense, stability is a highly desirable property in a measure of effectiveness, because measured effectiveness in one year predicts well effectiveness in subsequent years. Lockwood, Louis, and McCaffrey (2002) use simulations to explore how precise measures of performance would need to be to support inferences even at the tails of the distributions of teaching effectiveness and find that the necessary signal-to-noise ratio is perhaps unrealistically high. Schochet and Chiang (2013) also point out that the unreliability of teacher value-added estimates would lead to errors in identification of effective/ineffective teachers. They estimate error rates of about 25% among teachers of all experience levels

when comparing teacher performance to that of the average teacher. However, neither study focuses on differences between early career teachers and other teachers.

The perspective that stability and reliability are closely connected makes sense when true teaching effectiveness is expected to be relatively constant, as is the case of mid-career and veteran teachers. However, as shown in Figure 1, the effectiveness of early career teachers substantially changes over the first 5 years of teaching. Thus, teacher quality measures may reflect true changes over this period and, as a result, their measures could change from year to year in unpredictable ways. Anecdotally, one often hears that the first 2 years of teaching are a "blur" and that virtually every teacher feels overwhelmed and ineffective. If, in fact, first-year teachers' effectiveness is more subject to random influences and less a reflection of their true long-run abilities, their early evaluations would be less predictive of future performance than evaluations later in their career and would not be a good source of information for long-term decision making. Alternatively, even though value added tends to meaningfully improve for early career teachers, teachers' initial value added may predict their value added in the future quite well and thus be a good source of information for decision making.

We are aware of two related studies that explicitly focus on the early career period. Goldhaber and Hansen (2010b) explore the feasibility of using value-added scores in tenure decisions by running models that predict future student achievement as a function of teacher pretenure value-added estimates versus traditional teacher characteristics such as experience, master's degree obtainment, licensure scores, and college selectivity, and they find that the value-added scores are just as predictive as the full set of teacher covariates. We build on this work by exploring in more depth the implications of error in early career value-added scores for teachers. We model average trends in value-added scores by quintile of initial performance to examine propensity for improvement, and we explore the extent to which quintiles of initial performance overlap with quintiles of future performance. Staiger and Rockoff (2010) conduct Monte Carlo simulations to explore the feasibility of making early career decisions with information of varying degrees of imprecision. For example, they examine the possibility of dismissing some proportion of teachers after their *first* year on the job and find that it would optimize mean teacher performance to dismiss 80% of teachers after their first year—a surprisingly high threshold, although it does not account for possible effects on nondismissed teachers on the pool of available teacher candidates.

The current article distinguishes itself by providing an in-depth analysis of the real-world predictive validity of value added, with a distinct focus on teachers at the start of their career—a time when teacher performance is changing most rapidly and when districts have the greatest leverage to implement targeted human resource interventions and decisions. This article explores how actual value-added scores from new teachers' first 2 years would perform in practice if used by policy makers to anticipate and shape the future effectiveness of their teaching force. We are particularly interested in providing a framework through which policy makers might think about relevant policy design issues relative to current practice in most districts. Such issues include the following: What is an appropriate threshold for initial identification as highly effective or in need of intervention, how much overlap is there in the future performance of initially highly effective and ineffective teachers, and what are the trade-offs as one considers identifying more teachers as ineffective early in the career? We consider these questions in terms of early identification of both highly effective teachers (to whom districts might want to target retention efforts), as well as ineffective teachers (to whom the district might want to target additional support). Finally, we explore whether value-added scores in different subjects might be more or less useful for early identification policies—an issue not covered to date with regard to early career teachers but one that turns out to be important (see Lefgren & Sims [2012] for an analysis of using cross-subject value-added information for teachers of all levels of experience in North Carolina).

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) and the New York State Education Department (NYSED). These data include annual student achievement in math and English language arts (ELA) 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 ELA in Grades 3 through 8. However, for the current analysis, we restrict the sample to value added for elementary school teachers (Grades 4 and 5), because of the relative uniformity of elementary school teaching jobs compared with middle school teaching, where teachers typically 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 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 Grades 3 through 8. For most years, the data include scores for 65,000 to 80,000 students in each grade. We standardize 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

TABLE 1

Population of Teachers Who Began Teaching in SY 1999–2000 or After and Primarily Taught Grades 4 and 5: Descriptive Statistic on Three Relevant Analytic Samples Restrictions

	Math			ELA		
Has VA Scores in at Least . . .	(A) First Year	(B) 2 of Next 4 Years	(C) Years 1–5	(A) First Year	(B) 2 of Next 4 Years	(C) Years 1–5
Average VA score in first year	–0.035	–0.032	–0.030	–0.036	–0.034	–0.030
Proportion female, %	84.6	85.2	85.9	84.6	85.3	86.0
Proportion White, %	63.3	63.1	65.3	64.1	64.1	66.4
Proportion Black, %	17.5	17.7	17.8	17.8	18.2	18.3
Proportion Hispanic, %	11.5	11.6	10.3	10.6	10.5	9.1
Average standardized verbal SAT score	–0.107	–0.124	–0.120	–0.108	–0.129	–0.126
Average standardized math SAT score	–0.100	–0.123	–0.145	–0.110	–0.137	–0.148
Proportion attended most competitive UG, %	11.8	10.7	9.7	11.3	10.3	9.1
Proportion attended competitive UG, %	17.9	18.3	19.9	18.2	18.7	20.5
Proportion attended less competitive UG, %	39.6	40.2	40.7	40.2	40.9	41.8
Proportion attended not competitive UG, %	22.9	23.8	24.5	22.6	23.3	23.8
Proportion attended unknown UG, %	7.8	7.1	5.2	7.7	6.8	4.7
Pathway into teaching = college recommended path, %	54.0	55.7	56.7	54.3	56.1	57.4
Pathway into teaching = TFA path, %	1.6	0.7	0.1	1.6	0.9	0.2
Pathway into teaching = other nontraditional path, %	23.9	24.3	27.0	23.8	24.1	26.4
Pathway into teaching = unknown path, %	8.8	8.4	8.2	8.7	8.2	7.3
Number of teachers	**3,360**	**2,333**	**966**	**3,307**	**2,298**	**972**

Note: ELA = English language arts; SY = school year; TFA = Teach for America; VA = value added; UG = undergraduate institution.

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 3 through 8 in a given year. Data on teachers include teacher race, ethnicity, experience, and school assignment as well as a link to the classroom(s) in which that teacher taught each year.

Analytic Sample and Attrition

This article explores how measures of teacher effectiveness—value-added scores—change during the first 5 years of a teacher's career. For this analysis, we estimate teacher value added for the subset of teachers assigned to tested grades and subjects. Because we analyze patterns in value-added scores over the course of the first 5 years of a teacher's career, we can only include teachers who do not leave teaching before their later performance can be observed. Teachers with value-added scores typically represent about 20% of all teachers, somewhat more among elementary school teachers and less in other grades. As we indicate elsewhere, our analysis is intended to be illustrative of a process that could employ other measures of teacher effectiveness.

Table 1 provides a summary of three relevant analytic samples (by subject) and their average characteristics in terms of teacher initial value-added scores, demographics, and prior training factors, including SAT scores, competitiveness of their undergraduate institution, and pathway into teaching. In the relevant school years for this study, we observe 3,360 elementary school teachers who have a value-added score in their first year of teaching (3,307 for ELA). This is the population of interest—Group (A) in Table 1. Of these, about 29% (966 teachers) have value-added scores in *all* of the following 4 years, allowing us to track their long-run effectiveness annually. This sample—Group (C) in Table 1—becomes our primary analytic sample for the study. Limiting the sample to teachers with 5 consecutive years of value added addresses a possible attrition problem, wherein any differences in future mean group performance could be a result of a systematic relationship between early performance and the decision to leave within the first 5 years. The attrition of teachers from the sample may threaten the validity of the estimates because prior research shows evidence that early attriters can differ in effectiveness and thus maybe in their returns to experience (Boyd et al., 2007; Goldhaber, Gross, & Player, 2011; Hanushek et al., 2005). As a result, our 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 5 years (n = 966 for math, n = 972 for ELA).

Despite the advantages of limiting the sample in this way, the restriction of possessing value-added scores in every year introduces a potential problem of external validity. The notable decrease in sample size from Group (A) to Group (C) reveals that teachers generally do not receive value-added scores in every school year, and in research presented elsewhere, we examine this phenomenon (Atteberry, Loeb, & Wyckoff, 2013). That article shows there is substantial movement of teachers in and out of tested grades and subjects. Some of this movement may be identified as strategic—less effective teachers are moved out of tested grades and subjects. However, many of these movements appear less purposeful and therefore may reflect inevitable random movement in a large personnel management system. If teachers who are less effective leave teaching or are moved from tested subjects or grades during their first 5 years, the estimates of mean value added would be biased upward. That is, teachers who are consistently assigned to tested subjects and grades for 5 consecutive years may be different from those who are not. Because the requirement of having 5 consecutive years of value added scores is restrictive, we also examine results using a larger subsample of New York City teachers who have value-added scores in their first year and 2 of the following 4 years. This is Group (B) in Table 1 (2,333 teachers for math, 2,298 teachers for ELA). By using this larger subsample, we can run robustness checks using 70.1% of the 3,360 elementary teachers who have value-added scores in their first year (rather than the 28% when we use Group (A)). Table 1 shows that the average value-added scores, demographics, and training of teachers in these three groups are quite similar to one another, with few discernable patterns. In addition, while the primary analytic sample for the study is Group (A), we also replicate our primary analyses using Group (B) in Appendix C and find that the results are qualitatively very similar.

Methods

The analytic approach in this article is to follow a panel of new teachers through their first 5 years and retrospectively examine how performance in the first 2 years predicts performance thereafter. We estimate yearly value-added scores for New York City teachers in tested grades and subjects. We then use these value-added scores to characterize teachers' developing effectiveness over the first 5 years of their careers to answer the research questions outlined above. We begin by describing the methods used to estimate teacher-by-year value-added scores and then describe 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 study 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. We estimate teacher-by-year value added by employing a multistep residual-based method similar to that employed by the University of Wisconsin's Value-Added Research Center (VARC). VARC estimates value added for several school districts, including until quite recently New York City (see Appendix B). In Appendix C, we also 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 article. "VA Model C" differs from the main value-added model described in the article in that it uses student-fixed effects in place of time-invariant student covariates such as race/ethnicity, gender, and so on. In future work, others may be interested in whether teacher effectiveness measures derived from student growth percentile models would also garner similar results.

Research Question 1 (RQ1). Does the ability to predict future performance differ between novice and veteran teachers?

Previous research frequently characterizes the predictiveness of future value added based on current value added by examining correlations between the two or by examining the stability of observations along the main diagonal of a matrix of current and future performance quintiles. Although we explore other measures of predictiveness below, we employ these measures to assess whether there are meaningful differences between predictiveness of novice and veteran teachers.

Research Question 2 (RQ2). How well does initial job performance predict future performance?

The relationship between initial and future performance may be characterized in several ways. We begin by estimating mean value-added score trajectories during the first 5 years separately by quintiles of teachers' initial performance. We do so by modeling the teacher-by-year value-added measures generated by Equation 1 as outcomes using a nonparametric function of experience with interactions for initial quintile. Policy makers often translate raw evaluation scores into multiple performance groups 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.

Mean quintile performance may obscure the variability that exists within and across quintiles. For this reason, 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 or his value-added scores in the first 2 years of teaching. We use Equation 2 to predict each teacher's value-added score in a given "future" year (e.g., value-added score in years 3, 4, 5, or the mean of these) as a function of value-added scores observed in the first and second years. We present results across a number of value-added outcomes and sets of early career value-added scores, but Equation 2 describes the fullest specification, which includes a cubic polynomial function of all available value-added data in both subjects from teachers' first 2 years:

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

Equation 2 shows a teacher's math value-added score averaged in years 3, 4, and 5, $E\left[VA_{m,y=3,4,5}\right]$, predicted based on a cubic f\unction, f^3, of the teacher's math value-added scores from years 1 and 2, $\left(VA_{m,y=1}\right)$ and $\left(VA_{m,y=2}\right)$, as well as ELA value-added scores from years 1 and 2 $\left(VA_{e,y=1}\right)$ and $\left(VA_{e,y=2}\right)$. We summarize results from 40 different permutations of Equation 2—by subject and by various combinations of value-added scores used—by presenting the adjusted *R*-squared values that summarize the proportion of variance in future performance that can be accounted for using early value-added scores.

As policy makers work to structure an effective teaching workforce, they typically want to understand whether early career teachers will meet performance standards that place them in performance bands, such as highly effective, effective, or ineffective. Even if the proportion of the variance of future performance explained by early performance is low, it may still be a reliable predictor of these performance bands. We examine this perspective by examining mobility across performance levels of a quintile transition matrix of early and later career performance. For example, how frequently do initially high- (low-) performing teachers become low- (high-) performing teachers?

Finally, we examine the distribution of future performance scores separately by quintiles of initial performance. To the extent that these distributions are distinct from one another, it suggests that the initial performance quintiles accurately predict future performance.

Policy Implications and Trade-offs Associated With Inaccurate Predictions

Because we know that errors in prediction are inevitable, we present evidence on the nature of misidentification based on value-added scores from a teacher's first 2 years. We present a framework for thinking about the kinds of mistakes likely to be made and for whom those mistakes are costly, and we apply this framework to the 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 (e.g., targeted professional development, tenure decisions, or performance incentives). We then follow teachers through their fifth year, examining the frequency of accurate and inaccurate identifications based on early career designations. We use this approach to assess the benefits and costs of employing early career measures of value added to predict future value added. In addition, we examine whether such early career identification policies differentially affect teachers by race and ethnicity.

Results

RQ 1. Does the Ability to Predict Future Performance Differ Between Novice and Veteran Teachers?

The value added of novice teachers is less predictive of future performance than is value added of veteran teachers. Table 2 shows the correlations of value added of first-year teachers with their value added in successive years, as well as the correlation of value added of teachers with at least 6 years of experience with their value added in successive years. In all cases, value added is single year value added. In math, the correlations for novice teachers are always smaller than those for experienced teachers (differences are always statistically significant). Most relevant for our purposes is that the correlations with out-year value added diminish much more rapidly for novice than experienced teachers. For example, the correlation in "year + 5" is 37% of that in "year + 1" for novice teachers (0.132 vs. 0.356), while it is 75% for veteran teachers (0.321 vs. 0.421). A similar but somewhat less consistent and diminished pattern exists in ELA. Value added for early career teachers is meaningfully less predictive of future value added than it is for more experienced teachers. As we noted above, there is great conceptual appeal to employing value added in a variety of policy contexts for early career teachers. Just how misleading is early career value added of future performance? How might this affect policy decisions? We explore these questions below.

RQ 2. How Well Does Initial Job Performance Predict Future Performance?

Teachers with comparable experience can vary substantially in their effectiveness. For example, we estimate that the standard deviation in teacher math value added of first-year teachers is 0.21. Twenty percent of a standard deviation in student achievement is large relative to most educational interventions (Hill, Bloom, Black, & Lipsey, 2008) and produces meaningful differences in long-term outcomes for students (Chetty, Friedman, & Rockoff, 2014). Does this

TABLE 2

Cross-Year Correlation of Value-Added for Early Career Teachers and Veteran Teachers

	Math			ELA		
	Novice (Exp = 1)	Veteran (Exp > 5)	*p* Value	Novice (Exp = 1)	Veteran (Exp > 5)	*p* Value
Year + 1	0.356	0.421	.000***	0.206	0.293	.000***
Year + 2	0.302	0.384	.000***	0.160	0.258	.000***
Year + 3	0.275	0.333	.003***	0.131	0.207	.010**
Year + 4	0.191	0.321	.000***	0.090	0.171	.018*
Year + 5	0.132	0.321	.000***	0.152	0.185	.049*

Notes: The columns for Exp = 1 are the correlations of teachers' first-year value added with their value added in the subsequent 5 years (five rows). The columns for Exp > 5 are the correlations for teachers with at least 6 years of experience with their value added in the subsequent 5 years. The *p* values reported above are for the statistical test that the correlations for novice versus veteran teachers are statistically different from one another. Exp = experience; ELA = English language arts.

***$p < .001$, **$p < .01$, *$p < .05$.

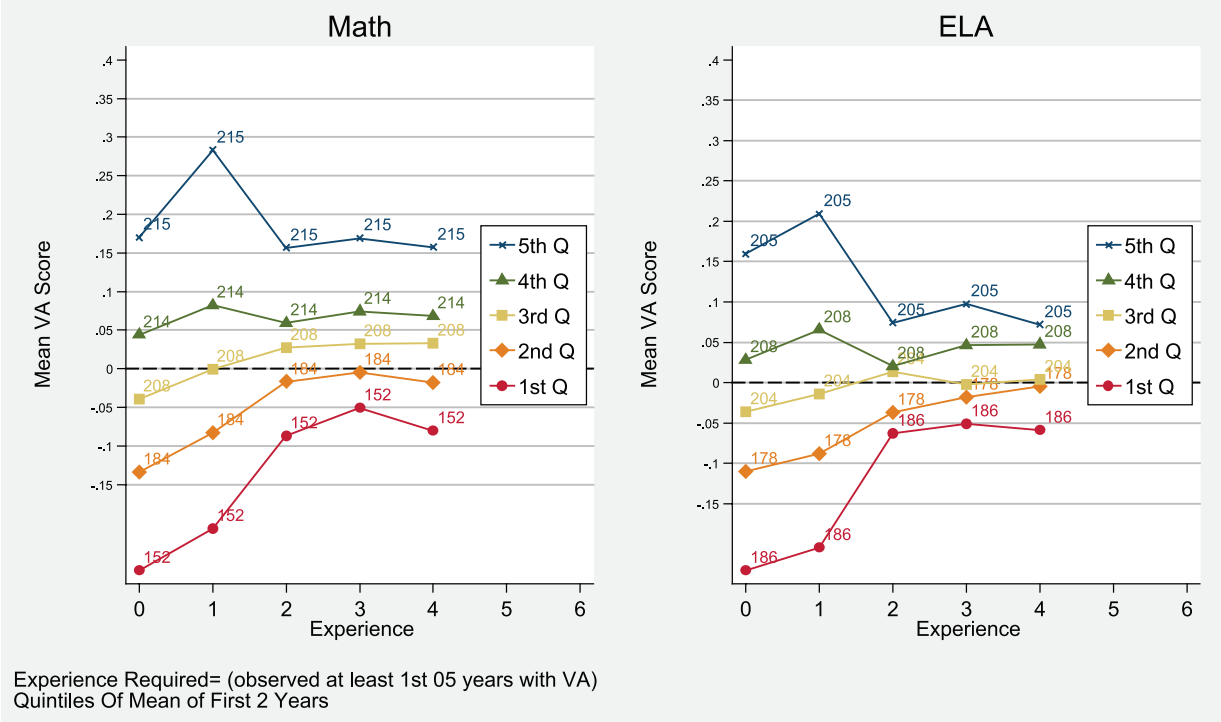


FIGURE 2. *Mean value-added (VA) scores, by subject (math or ELA), quintile of initial performance, and years of experience for elementary school teachers with VA scores in at least first 5 years of teaching. Numbers at each time point are sample sizes. These reflect the fact that quintiles are defined before limiting the sample to teachers with value added in all of their first 5 years. The sample sizes also reinforce the fact that patterns observed over time are among a consistent sample—changes over time are not due to any nonrandom attrition. The issues of defining quintiles and sample selection are discussed in greater detail in Appendices A and C. ELA = English language arts.*

variability in early career performance predict future differences? We assess the stability of early career differences from a variety of perspectives.

Figure 2 provides evidence of consistent differences in value added across quintiles of initial performance.[3] Although the lowest quintile does exhibit the most improvement (some of which may be partly due regression to the mean), this set of teachers does not, on average, "catch up" with other quintiles, nor notably are they typically as strong as the median first-year teacher even after 5 years. The issue of regression to the mean is somewhat mitigated by our choice to characterize initial performance by the mean value-added score in the first 2 years. To check the robustness of our findings to some of our main analytic choices, in Appendix C, we

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 2 Years

	Outcome			
Early Career VA Predictor(s)	VA in Year 3	VA in Year 4	VA in Year 5	Mean ($VA_{Years\ 3–5}$)
Math				
Math VA in year 1 only	0.079	0.052	0.077	0.111
Math VA in year 2 only	0.153	0.149	0.117	0.223
Math VA in years 1 and 2	0.176	0.158	0.146	0.256
VA in both subjects in years 1 and 2	0.176	0.171	0.147	0.262
VA in both subjects in years 1 and 2 (cubic)	0.178	0.171	0.146	0.261
ELA				
ELA VA in year 1 only	0.025	0.019	0.016	0.040
ELA VA in year 2 only	0.058	0.080	0.042	0.117
ELA VA in years 1 and 2	0.068	0.084	0.048	0.131
VA in both subjects in years 1 and 2	0.085	0.102	0.058	0.161
VA in both subjects in years 1 and 2 (cubic)	0.090	0.113	0.061	0.168

Note: ELA = English language arts; VA = value added.

re-create Figure 2 across three dimensions: (A) minimum value added required for inclusion in the sample, (B) how we defined initial quintiles, and (C) specification of the value-added models used to estimate teacher effects. Findings are quite similar in a general pattern, suggesting that these results hold up whether we use the less restrictive subset of teachers (based on number of available value-added scores) or had used other forms of the value-added model.

While useful for characterizing the mean pattern in each quintile, Figure 2 potentially masks meaningful within-quintile variability. To explore this issue, we present adjusted *R*-squared values from various specifications of Equation 2 in Table 3. This approach uses the full continuous range of value-added scores and does not rely on quintile definitions and their arbitrary boundaries. 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. For example, teachers' math value-added scores in the first year explain 7.9% of the variance in value-added scores in the third year. The predictive power is even lower for ELA (2.5%). Employing value added for the first 2 years explains 17.6% of value added in the third year (6.8% for ELA). A second evident pattern in Table 3 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 employ math value added in both of the first 2 years, we explain slightly more than a quarter of the variation in future math value added averaged across years 3 through 5 (0.256). Adding reading value added improves the explanatory power, but not by much (0.262).

The predictive power of early value-added measures depends on which future value-added measure they are predicting. Not surprisingly, given the salience of measurement error in any given year, early scores explain *averaged* future scores better than they explain future scores in a particular year. For example, for math, our best prediction model for year 3 value added (column 1) explains only 17.6% of the variation (8.5% for ELA). In contrast, when predicting variation in mean performance across years 3 through 5 (column 4), the best model predicts up to about 26% of the variance in math (16.8% in ELA).

Teacher's early value added is clearly an imperfect predictor of future value added. To benchmark these estimates, we compare them to predictiveness of other characteristics of early career teachers and to other commonly employed performance measures. As one 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 3 through 5 (same outcome as column 4 of Table 3) using this set of explanatory factors, we explain less than 3% of the variation in the math or ELA outcomes.[4] Another way of benchmarking these findings is to compare them to the predictive validity of other commonly accepted measures used for high-stakes evaluation. For example, SAT scores, often employed in decisions to predict college performance and grant admission, account for about 28% of the variation in first-year college grade point average (GPA) (Mattern & Patterson, 2014).

For a noneducation example, surgeons and hospitals are also often rated based on factors that are only modestly correlated with patient mortality (well below 0.5), but the field publishes these imperfect measures because they are better than other available approaches to assessing quality (Thomas & Hofer, 1999). (See also Sturman, Cheramie, & Cashen, 2005, for a meta-analysis of the temporal consistency of performance measures across different fields.) Although early career value added is far from a perfect predictor of future value added, it is far better than other readily available measures of teacher performance and is roughly comparable to the SAT as a predictor of future college performance.

These analyses suggest that initial value added is predictive of future value added; however, they also imply that accounting for the variance in future performance is difficult. Each of the prior illustrations provides useful information but also has shortcomings: The mean improvement trajectories by quintile shown in Figure 2 may obscure the mobility of teachers across quintiles. The explained variation measures reported in Table 3 provide much more detailed information regarding the relationship between early and future performance but may not inform a typical question confronting policy makers—how frequently do teachers assigned to performance bands (e.g., high or low performing), based on initial value added, remain in these bands when measured by future performance?

To illustrate the potential of value added to address this type of question, Table 4 shows a transition matrix that tabulates the number of teachers in each quintile of initial performance (mean value added of years 1 and 2) (rows) by how those teachers were distributed in the quintiles of future performance (mean value added of years 3–5) (columns), along with row percentages.[5] The majority—62%—of the initially lowest quintile math teachers are in the bottom two quintiles of future performance. Thus, a teacher initially identified as low performing is quite likely to remain relatively low performing in the future. About 69% of initially top quintile teachers remain in the top two quintiles of mean math performance in the following years. Results for ELA are more muted: About 54% of the initially lowest quintile are in the bottom two quintiles in the future, and 60% of the initially highest quintile remain in the top two quintiles in the future. Overall, the transition matrix suggests that measures of value added in the first 2 years predict future performance for most teachers, although the future performance of a sizable minority of teachers may be mischaracterized by their initial performance.

Broadening the transition matrix approach, we plot the distribution of future teacher effectiveness for each of the quintiles of initial performance (Figure 3). These depictions provide a more complete sense of how groups based on initial effectiveness overlap in the future.[6] The advantage, over the transition matrix shown above, is to illustrate the range of overlapping skills for members of the initial quintile groups. We can examine these distribution with various key comparison points in mind. 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 future performance in each respective initial-quintile group. Second, the diamond ("♦") represents the mean *initial* performance by quintile. This allows the reader to compare distributions both to where the group started on average, as well as to the mean future performance of each quintile.

Most 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 3, 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 are just as skilled in future years as initially highest performing teachers. However, most of these two distributions are distinct from one another.

How do the mischaracterizations implied by initial performance quintiles (Figure 3) compare to meaningful benchmarks? For example in math, 69% of the future performance distribution for the initially lowest performing quintile lies to the left of the mean performance of a new teacher (the comparable percentage is 67% for ELA). Thus, the future performance of more than two thirds of the initially lowest performing quintile does not rise to match the performance of a typical new teacher. A more policy relevant comparison would likely employ smaller groupings of teachers than the quintiles described here.[7] We examine the mischaracterizations and the loss function for such a policy below.

Policy Implications: What Are the Trade-offs Associated With Inaccurate Predictions?

District leaders may want to use predictions of future effectiveness to assign teachers to various policy regimes for a variety of reasons. For example, assigning targeted professional development and support to early career teachers who are struggling represents potentially effective human resources policy. Another possibility would be to delay tenure decisions for teachers who have not demonstrated their ability to improve student outcomes during their first 2 years. Alternatively, if high-performing teachers could be identified early in their careers, just when attrition is highest, district and school leaders could target intensive retention efforts on these teachers. In our analysis, initial performance is a meaningful signal of future performance for many teachers; however, the future performance of a number of other teachers is not reflected well by their initial performance. What does this imprecision imply about the policy usefulness of employing initial value-added performance to characterize teacher effectiveness?

Figure 4 provides a framework for empirically exploring the potential trade-offs in identifying teachers when the measures employed imprecisely identify teachers. It plots

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**

Note: ELA = English language arts.

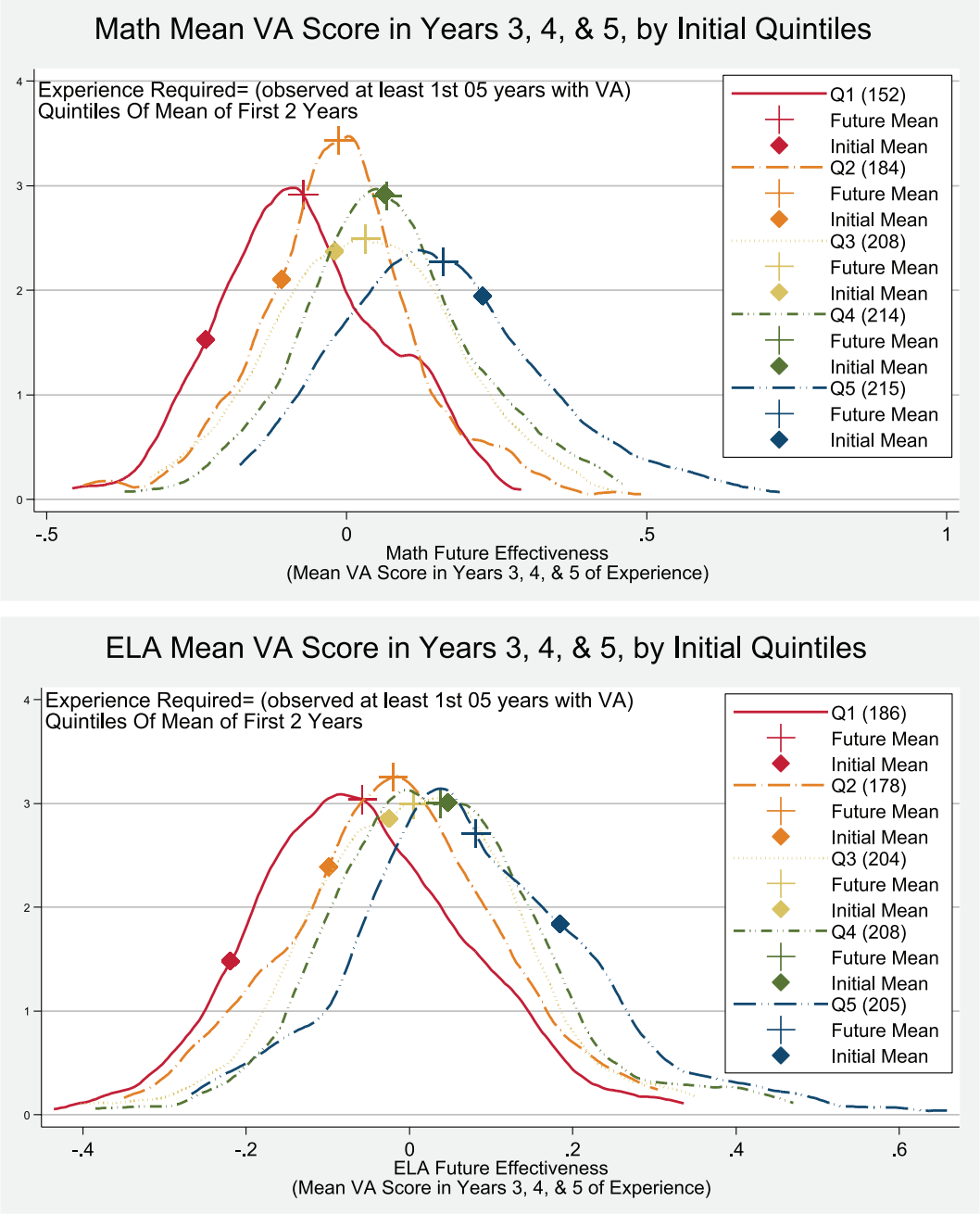


FIGURE 3. *Distribution of future value-added (VA) scores, by initial quintile of performance.*

future performance as a function of initial performance percentiles. Moving from left to right along the x-axis represents an increase in the threshold for identifying a teacher as ineffective (i.e., candidates for intervention) based on initial performance. For this exercise, we capture initial performance by calculating the mean of a teacher's value-added scores in years 1 and 2 and translating that into percentiles (x-axis). The y-axis depicts the associated percentage of teachers who appear in each tercile of future performance.[8] The figure plots the extent to which those initially identified as low performing are in the bottom third of future performance (red portion), are in the middle (yellow), or are among the top third of future performance (green). It is, of course, somewhat arbitrary to use "bottom third" as the cut-off for teachers who continue to be low performing in the future. Below, we also explore defining a teacher as low performing if he or she continues to perform below the average teacher (or the average first-year teacher)—an approach that would identify more teachers as ineffective. Of these two options, we begin with our somewhat more conservative definition of relatively low performing in the future (i.e., bottom third).

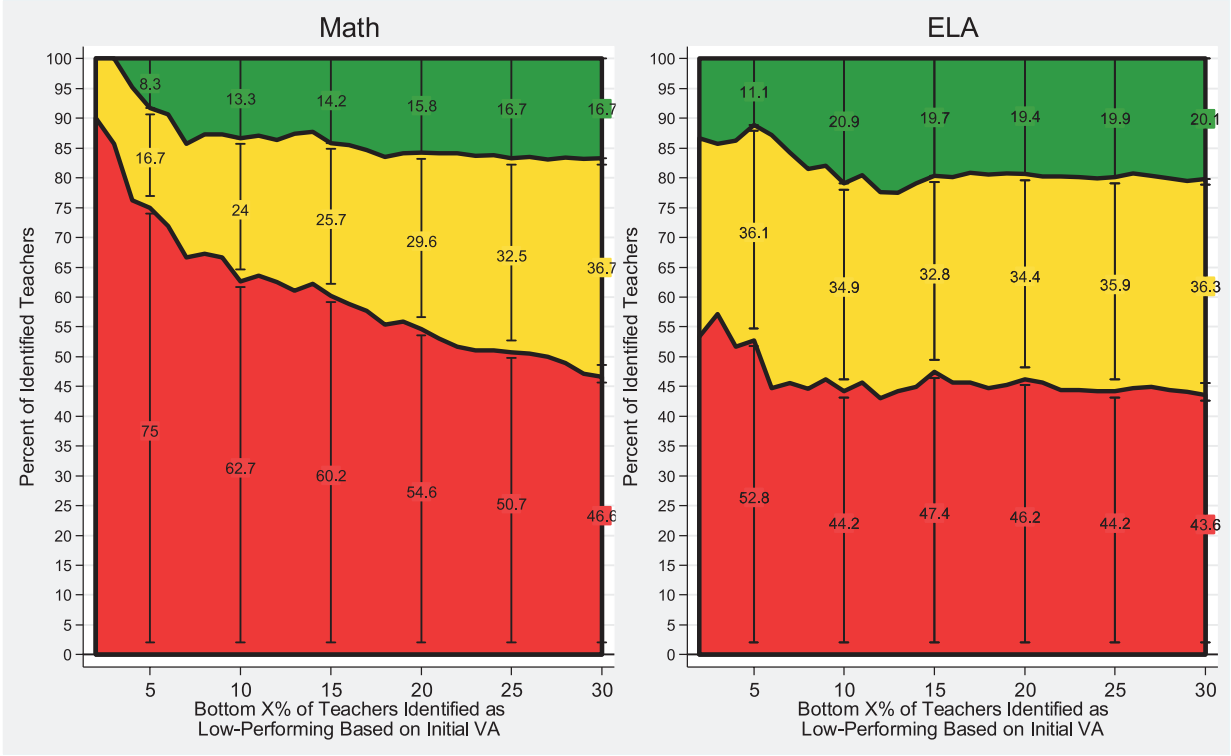


FIGURE 4. *Depiction of error rate for initially identifying a percentage of teachers as low-performing based on initial value-added scores in the first 2 years, by subject. ELA = English language arts.*

To illustrate the utility of this figure, we begin by focusing on the vertical line that passes through X = 5 on the horizontal axis, which indicates the effects of identifying the lowest 5% of teachers as ineffective. In this first example, we are considering a proposal to move from a current policy where no teachers are identified as ineffective to a new policy in which the bottom 5% of the initial performance distribution are identified as ineffective.[9] For instance, those 5% of the initially lowest performing teachers could receive targeted professional development in the early career. Does such a move constitute a policy improvement? We know the new policy will misidentify some teachers who are not low performing in the future, and Figure 4 allows us to quantify that rate of misidentification. At that level (X = 5), 75% (red) of teachers initially identified as ineffective subsequently perform in the lowest third of future performers. In other words, for three fourths of the initially identified set of teachers, the professional development intervention would have been warranted given that they continued to struggle into their fifth year. On the other hand, 16.7% of the initially identified teachers are in the middle tercile of future performance (yellow), and 8.3% (green) end up in the top third of future performance. In this hypothetical scenario, the middle- and high- tercile teachers would therefore receive targeted professional development that they may not have needed because they would have moved out of the bottom third without it.[10]

In our first example, about a quarter of the 5% of teachers initially identified as low performing are not among the bottom third in the future, and about 8% are actually among the most effective teachers in the future. It is worth keeping in mind here that this misidentification occurs for 8% *of the bottom 5%* of the overall distribution of teachers—that is, if there were 1,000 teachers, 50 would have been initially identified as low performing, and 4 of those would have subsequently appeared in the top tercile. The 5% threshold therefore made some—but not very many—egregious errors, in which teachers who would become among the most effective would have been inappropriately identified based on their early career value-added scores. On the flip side, we also know that 75% of the time, the ineffective label correctly identifies teachers who will be low performing in the future. Depending on the consequences, the identification process may accrue benefits to identified teachers, nonidentified teachers, or students. It is important to compare this to the original policy in which no teachers were identified as ineffective. While the percentage of initially identified teachers who are ultimately *mis*identified goes from 0% to 8% (the original policy vs. new 5% policy), we also know that the original policy failed to identify *any* teachers correctly,[11] whereas the new policy accurately captured future performance for 75% of teachers. This latter aspect of the policy comparison is often overlooked, since we are often

more concerned with what could go wrong (for teachers) than what could go right (for students).

Thus far, we have examined findings at the 5% identification level, but we will see that future misidentification rates depend on the size of the initially identified group. Figure 4 allows one to compare any two potential identification policies. If instead of identifying the bottom 5% of initial performers, one identifies the bottom 10%, the proportion of identified teachers who fall in the bottom tercile of future performance declines to 62.7%, with the attendant increase in serious misclassification from 8.3% to 13.3%. The 10% identification policy is getting more predictions right than wrong, but the error rate is somewhat higher. One might be concerned that a 10% identification rate is unrealistically high, but we argue that it depends on the policy one has in mind. For instance, if considering a proportion of teachers to dismiss, 10% seems high given the high cost to teachers of this misidentification. However, if considering a policy to target professional development to teachers who are struggling, misidentification has very little cost, and therefore 10% may be appropriate.

The framework we present here is useful for examining the differential rates of accurate and inaccurate identification at different initial levels of early identification. As we discuss below, district leaders and policy makers must then think through the potential costs of misidentification and benefits of accurate identification that would be associated with whatever policy response they may be considering. In the case of targeting professional development, the costs to misidentification (a small percentage of teachers receiving professional development they do not need) may not be particularly high. Indeed, it is common practice in districts for *all* teachers to receive professional development without regard to their performance. However, if the targeted policy intervention was, instead, delaying tenure (thus opening up the possibility that the teacher might not ultimately be retained), then the costs to misidentification would certainly be higher. Ultimately, the question of costs and benefits is policy and context specific. The information provided by our analysis can help policy makers think about the frequency of accurate predictions but not the relative utility of those judgments.

Above we have walked through the results on the left-hand side of Figure 4 (math), but the empirical outcomes for ELA are substantively different. As shown in the second panel of Figure 4, identifying ineffective teachers based on the lowest 5% of initial ELA performance leads to only 52.8% of these teachers being in the lowest tercile of future performance, implying that 47% are *not* among the bottom third in the long run (11.1% become among the top third of teachers in the future). Thus, employing initial value added to identify the future performance of teachers based on ELA value added leads to many more misidentifications. This pattern is entirely consistent with our earlier analysis that showed future ELA effectiveness was less predictive than math effectiveness. As is evident, for ELA, extending identification beyond 6% of teachers leads to more misidentified than correctly identified teachers. Again, how problematic this is depends on the benefits of correct identification and the costs of misidentification.

In Figure 4, we opted to think of future low performance as continued presence in the bottom third of the distribution of teaching, but districts may have different thresholds for evaluating whether a teacher should be identified as low performing. For example, let us consider a very different policy mechanism in which initially identified teachers become candidates for dismissal. In this case, the relevant comparison would be between the identified teachers and the teacher who likely would be hired in her or his place—an average first-year teacher. We therefore also compare a novice teacher's ongoing performance to that of an average first-year teacher, as this represents an individual who could serve as a feasible replacement. In fact, among the teachers in the bottom 5% of the initial math performance distribution, the vast majority—83.3%—do not perform in the future as well as an average first-year teacher in math. The corresponding number is 72.2% 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 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% in years 1 and 2 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 two thirds of a full standard deviation in teacher effectiveness. 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. Furthermore, an ineffective teacher retained for *3* additional years imposes 3 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% in years 1 and 2 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 or his 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.

Accurately identifying the effectiveness of early career teachers has large benefits to improving educational outcomes for students. Misidentifying teacher effectiveness has potentially large costs for teachers and in some cases students. The analysis has shown how misclassifying teachers

increases as the portion of teachers initially identified increases. What is the "right" level of identification? The answer is complex and depends on several factors. First, as the current debate would imply, the predictiveness of future effectiveness by initial effectiveness is important. If initial value added were a perfect predictor of future value added, most policy makers would be much more comfortable with using it. Predictiveness also requires a clear criterion for identifying future ineffectiveness and excellence. Would policy makers be willing to agree that a teacher identified as initially high performing was correctly identified as long as she or he was in the top tercile of future performance, as we have employed here? The smaller the future performance band, the less predictive any measure will be. Second, value-added measures of future performance are imprecise, narrowly defined, and relative. Where possible, we mitigate the problems of measurement error by employing Bayes shrunk estimates as well as averaging across multiple future years. We, however, cannot address the narrowness of the value-added measure relative to the way most people conceptualize student outcomes or its relative nature. For these purposes, employing 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.

Third, tolerance for imprecision in identifying teachers also depends on the benefits of correct identification and the costs of misidentification. For example, if the benefit to students from more effective teachers was substantial, as suggested by Chetty et al. (2011), in terms of college-going and future earnings (e.g., they estimate that having a top 5% teacher for just 1 year raises a child's lifetime income by $80,000), that would imply an increased tolerance of the misclassification of increasing the identification of ineffective teachers, other things equal. As the district becomes known for rewarding effectiveness and intervening to help teachers who underperform, teachers who think they can be rewarded in such a system may be attracted to the profession. On the other hand, if these identification systems seem to teachers to be driven by noise and—due to the imprecision of any given measures of teaching effectiveness—teachers do not perceive that these early identifications are reliable, then even effective teachers may be discouraged from remaining in the profession. Little is known about how increasing the link between imperfect teacher performance measures and teacher pay or retention policies would affect the general equilibrium of the teacher labor market, but Rothstein (in press) explores these issues in a simulation that suggests that both bonus and tenure policies outlined in his study have greater benefits than costs, although he acknowledges that these analyses are based on hypothesized parameters that the field has not yet estimated well. Indeed, this may depend very much on local labor market conditions.

One final but important concern with policies that attempt to predict future performance based on imperfect information from the early career is potentially nonrandom error in measurement. Value-added measures used to detect early performance might systematically favor one group of teachers over another based on characteristics unrelated to true effectiveness. For example, value-added scores could be lower for teachers working in disadvantaged schools or with certain student populations. Most research on value added shows relatively small systematic bias of this type, but the research is likely incomplete (Raudenbush, 2013). It is also possible some group of teachers has lower value added but greater effectiveness on a different dimension of teaching. If this were the case, then if the use of value added discouraged teachers with these other important skills from teaching, then the teaching force could suffer. More generally, narrowly defined measures of quality could reduce the diversity of the teaching force along multiple dimensions.

As a first attempt 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 2 years). We examine characteristics of teachers who are in the extremes of the initial performance distribution—that is, the top 5% and 10% (the initially highest performers), as well as the bottom 5% and 10%. This analysis, unlike analyses elsewhere in this article, include all teachers for whom we are able to calculate initial performance (value added in the first 2 years of teaching) because the question is whether early career identification disproportionately affects teachers by race/ethnicity. Table 5 shows that the relative percentages of teachers identified at each of these thresholds are quite similar by race. The differences for both math and ELA are small and never statistically significantly different from each other.[12] These findings suggest that a policy of identification of early career teachers by value-added scores would not incidentally identify minority teachers at differential rates as either high or low performers, 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 retention, development (e.g., mentoring programs), tenure, dismissal, and promotion are likely most relevant during this period. In this article, we describe the trajectory of teachers' performance over their first 5 years as measured by their value added to math and ELA test scores of students. 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 2 years. We find that, on average, initial performance is 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, and so on. Predictions are more

TABLE 5

Identification of Top and Bottom Performers by Race/Ethnicity and Subject (column percentages)

Initial Percentage Identified	White	Black	Hispanic
Math			
Top 5%	5.37	5.67	5.00
Top 10%	9.73	11.17	10.00
Bottom 5%	4.44	4.50	4.77
Bottom 10%	9.44	6.67	11.36
ELA			
Top 5%	5.69	4.83	5.58
Top 10%	9.89	10.50	9.90
Bottom 5%	4.79	3.83	4.82
Bottom 10%	9.59	8.67	8.63

Note: Unlike other tables in the article, this sample is *not* restricted to the set of teachers who have value added in all of their first 5 years. Because this analysis relies only on knowledge of value added in the first 2 years, the sample is restricted only to the set of teachers who have value added in the first 2 years of teaching. Omitted from this table are teachers of other races, which in total constitute 5.2% of the sample. ELA = English language arts.

powerful at the extremes of the performance distribution and also in math relative to ELA. Finally, it appears to be somewhat easier to identify teachers who will be excellent in the future, as opposed to teachers who will be very ineffective.

This said, predictions about teachers' future performance based on initial value added are far from perfect. Initial performance accounts for about 25% of the overall variance in future math performance and 16% in future ELA performance. Thus, basing human resource decisions on initial value-added estimates will lead to errors since some teachers who would, in fact, be highly effective will be treated as ineffective, and some teachers who would be ineffective will be treated as if they were effective. However, it is important to keep in mind that most districts have a default policy in which they make little attempt in the first few years to use value-added information to anticipate future effectiveness. This *absence* of policy also inherently misidentifies teachers: In this instance, identification rates are zero—no teachers are judged to be ineffective when in fact some do turn out to be relatively low performing. The costs of these mistakes by omission are typically not discussed and are borne by students who are subjected to less effective teachers who could have been identified for intervention early in their careers. The relevant question for educational policy is not whether predictions of teacher effectiveness misidentify teachers—any prediction of human behavior will inevitably contain errors—but rather what is the level of identification that maximizes the net benefits? Is it nearly zero, which is the default policy in many districts, or some level above that?

Ultimately, that is a question that policy makers must resolve as they weigh benefits and costs of alternative levels of identifying high- and low-performing teachers in their state or district. Crucial to that decision is an empirical understanding of the accuracy of alternative predictions. Our data from New York City suggest that at relatively low levels of identifying teachers (e.g., the bottom and top 5%), initial performance accurately categorizes future performance for most math teachers (about 75%), but for ELA teachers, predictions are closer to being accurate about half the time. Compare this to a district that makes no attempt to identify struggling teachers at the start of their careers and therefore makes no accurate predictions about future ineffective teachers. Even 50% accuracy seems an improvement over making no predictions at all.

At the same time, it is important to acknowledge that any policy based on value-added scores would necessarily be limited to the 20% to 30% of district employees who possess value-added scores in multiple years, and then on top of that, it would at most identify a small subset of those teachers (we discuss herein identification rates of 5% to 10%, but this would depend on the given policy). On one hand, if it is the case that only 8% of the 5% who are initially identified are *mis*identified (as we find in math), then one positive note is that very few teachers overall would experience this misclassification. On the other hand, the benefits of correctly identifying teachers early in their career is also necessarily limited to a subset of that small group of 5% to 10% of teachers. Depending on how productively a district can respond to a teacher's early identification (can we help them improve? can we recruit better replacements?), the evaluation system may not have a large impact on the overall distribution of teaching effectiveness. Given that these systems are controversial and not always favored by teachers, it is unclear whether the overall benefits are worth the costs. That said, for the set of students who are assigned to teachers that the district could have predicted would continue to struggle, the intervention might be quite valuable.

Finally, it is also important to acknowledge that the stability of value-added scores could vary from district to district (McCaffrey et al., 2009), which in turn could mean that the predictive power could be stronger or weaker in other places. Therefore, districts may be interested in implementing their own analyses of the predictive power of the teaching effectiveness measures as an informative step to shaping policies based on those measures.

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

APPENDIX TABLE A1

Analytic Sample: 966 Teachers Who Have Value-Added Scores in All of the First 5 Years: Difference in Mean Value Added and Numbers of Final Analytic Sample Teachers in Each Quintile of Initial Performance, Depending on Approach to Quintile Construction

Approach to Quintile Construction		Q1	Q2	Q3	Q4	Q5	Total
Math							
Quintiles of all teacher-years (1)	*n*	224	207	194	219	122	966
	Mean	–0.165	–0.049	0.015	0.092	0.222	
Quintiles made after limiting to teachers in first year (2)	*n*	171	171	198	212	214	966
	Mean	–0.224	–0.100	–0.018	0.063	0.227	
. . . And limiting to elementary teachers (3)	*n*	150	187	207	213	209	966
	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	966
	Mean	–0.214	–0.083	–0.002	0.077	0.239	
ELA							
Quintiles of all teacher-years (1)	*n*	246	196	208	181	141	972
	Mean	–0.156	–0.059	0.002	0.066	0.158	
Quintiles made after limiting to teachers in first year (2)	*n*	214	163	185	198	212	972
	Mean	–0.206	–0.088	–0.022	0.046	0.180	
. . . And limiting to elementary teachers (3)	*n*	185	176	201	208	202	972
	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	972
	Mean	–0.213	–0.090	–0.016	0.054	0.188	

Note: We construct quintiles of performance in a teacher's first 2 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 5 consecutive years of value-added (VA) scores, beginning in the teacher's first year of teaching. Note that method (3) above is the preferred approach for this article. ELA = English language arts.

future performance. We thus chose to center a teacher's first-year value-added score on 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.

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 5 years. As is evident from Table 1 in the article, about 29% of elementary teachers with value-added scores in their first year 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 5 consecutive years of value added might not be comparable to the "bottom quintile" among all first-year 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 5 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 we will have in the first 5 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 1 and 2 and make quintiles thereof. In Appendix C, we present some specification checks by examining our main results using value added from the first 2 years in a variety of ways (e.g., first year only, second year only, a weighted average of the first 2 years, teachers who were consistently in the same quintile in both years). In Appendix Table A1, 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 the first 5 years) depends on quintile construction.

Appendix B

Estimation of Teacher Value Added

We estimate teacher-by-year value added by employing a multistep residual-based method similar to that employed by

the University of Wisconsin's Value-Added Research Center (VARC). VARC estimates value added for several school districts, including until quite recently New York City.

We initially estimate Equation 1, which regresses achievement (Y_{icsjt}) for student i in class c at school s taught by teacher j in time t as a function of prior achievement $(Y_{icsjt_{-1}})$, student attributes (X_{icsjt}), and class fixed effects (α_{csjt}). In this model, the class fixed effects subsume both the teacher-by-year fixed effect (τ_{jt}) and any other class- (Z_{csjt}) or school-level (S_{st}) predictors of student achievement.

$$Y_{icsjt} = \lambda Y_{icsjt_{-1}} + \beta' X_{icsjt} + \alpha_{csjt} + \varepsilon_{icsjt}, \quad (1)$$

where $\alpha_{csjt} = \gamma' Z_{csjt} + \varphi' S_{st} + \tau_{jt}$

Employing these estimates, we calculate the residuals $(\hat{\varepsilon}_{icsjt})$ from this regression without accounting for α_{csjt} and then estimate Equation 2, which regresses this residual on class and school characteristics as well as a class random effect (ζ_{jt}) to reflect the grouping of students into classrooms.

$$\hat{\varepsilon}_{icsjt} = \alpha_{csjt} + \varepsilon_{icsjt} = \gamma' Z_{csjt} + \varphi' S_{st} + \zeta_{jt} + \omega_{icsjt}. \quad (2)$$

Employing these estimates, we calculate the residuals (q_{icsjt}) from this model and calculate teacher-by-year value added by averaging across the student-level residuals within a teacher and year.

$$\hat{\tau}_{jt} = \bar{q}_{icsjt} \quad (3)$$

The teacher-by-experience fixed effects become the value-added measures that serve as the outcome variable in our later analyses. They capture the average achievement of teacher j's students in year t, 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.

The standard errors of the teacher-by-year value-added estimates are estimated as shown in Equation 4 using the student-level errors $(e_{icsjt} = q_{icsjt} - \hat{\tau}_{jt})$ from Equation 3 and number of observations for each teacher-by-year group.

$$SE\left(\hat{\tau}_{jt}\right) = \sqrt{var\left(e_{icsjt}\right) / N_{jt}}. \quad (4)$$

We then employ a standard empirical Bayes shrinkage method to account for the varying uncertainty associated with each teacher-by-year value-added estimate.

In the teacher-by-year value-added model presented above, we make several important analytic choices about model specification. Our preferred model uses a lagged achievement as opposed to modeling gain scores as the outcome.[13] 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 analysis to comparisons only between teachers who have taught at least some students in common.[14] 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 allow only valid comparisons within the same school. In Appendix C, we examine results across a variety of value-added models, including models with combinations of gain score outcomes, student, and school fixed effects.

Appendix C

In Figure 2 of the article, we present mean value-added scores over the first 5 years of experience, by initial performance quintile. We do so on a somewhat restrictive sample of about 29% of the elementary teachers with value-added scores in their first year of teaching who also have value-added scores in all of their first 5 years (see Table 1). As discussed in the article, we also rerun analyses on a less restrictive sample of teachers who have value-added scores in only 2 of the 4 years following their first. This is but one example of an analytic choice made in this article that could affect our findings. In this appendix, we examine whether our primary findings are robust to three analytic choices made in the article: (A) minimum value-added required for inclusion in the sample, (B) how we defined initial quintiles, and (C) specification of the value-added models used to estimate teacher effects:

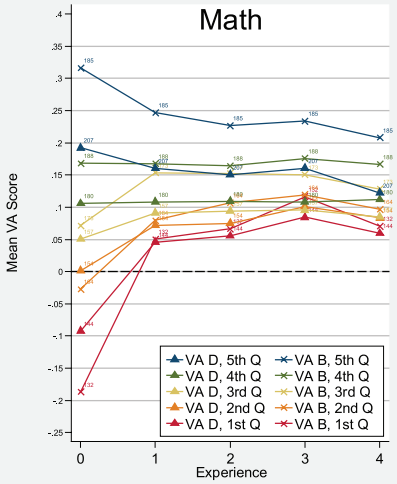
(A) We examine results across two teacher samples based on minimum value added required for inclusion. The first figure uses the analytic sample used throughout the main article—teachers with value-added scores in all of their first 5 years. The second widens the analytic sample to the set of teachers who are consistently present in the data set for at least 5 years but only possess value-added scores in years 1 and 2 of the next 4 years.

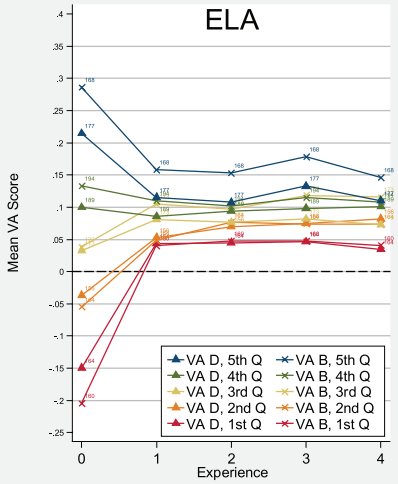
(B) We examine results across four possible ways of defining quintiles: (1) "quintile of first year"—this includes quintiles of teachers' value-added scores in their first year alone; (2) "quintile of the mean of the first 2 years"—this includes quintiles of teachers' *mean* value-added scores in the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

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

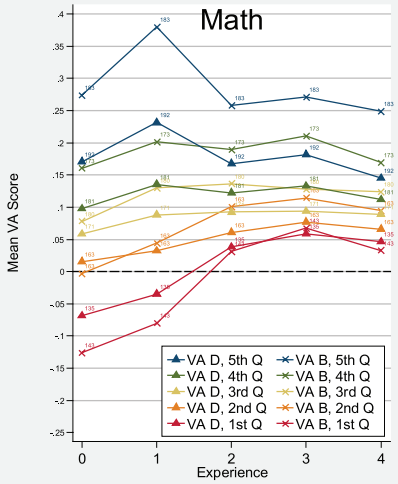
Elem Teachers with VAM in All of First Five Years

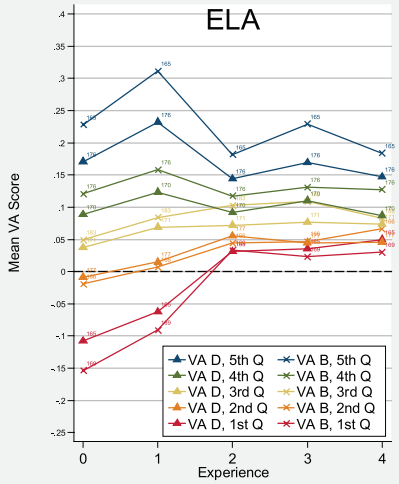
Quintile Among All 1st-Yr Tchrs



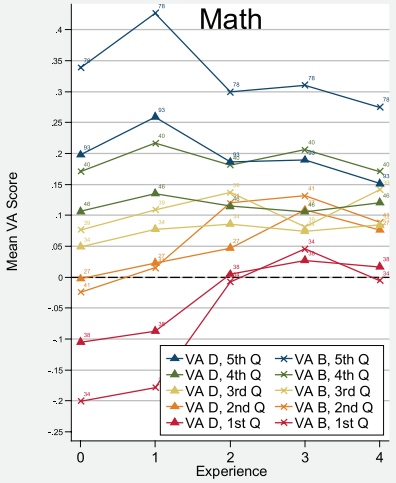


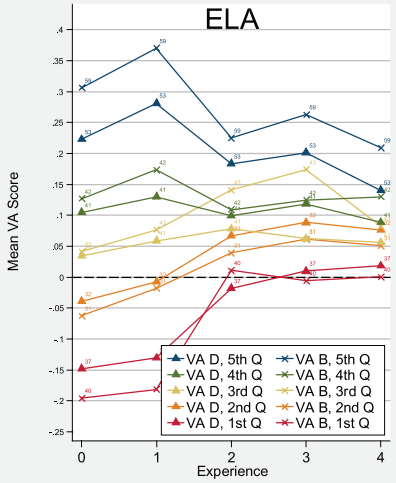
Quintile Of Mean of First 2 Years



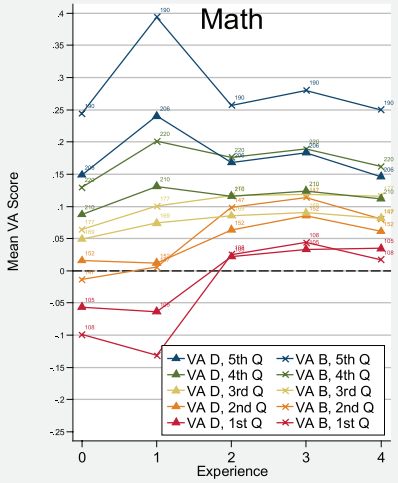


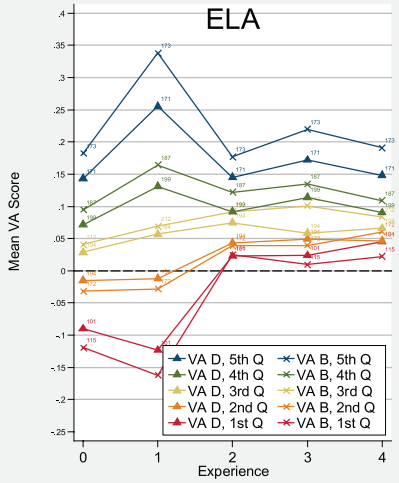
Quintile Consistent in First Two Years





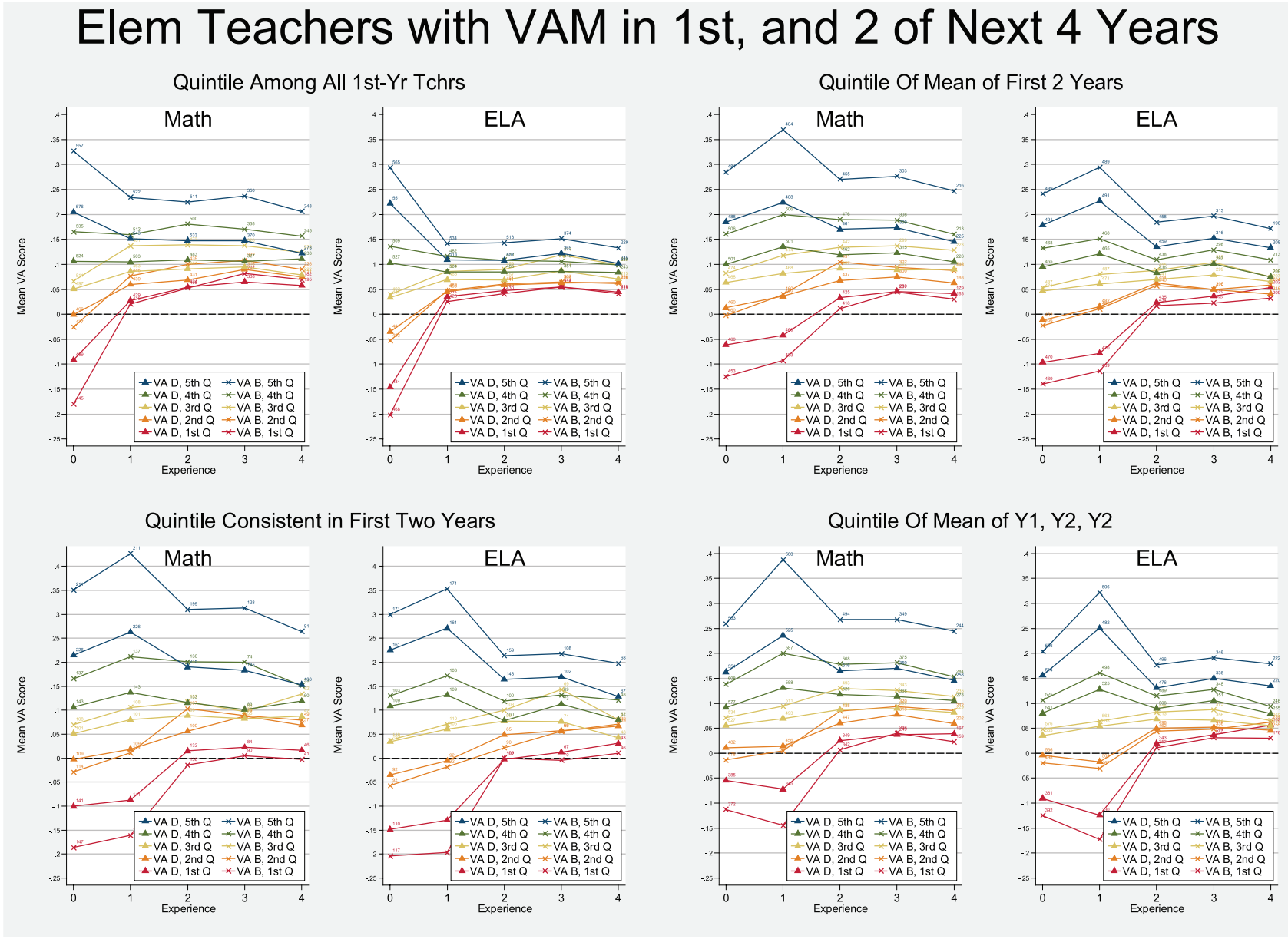
Quintile Of Mean of Y1, Y2, Y2





(continued)

FIGURE (CONTINUED)



Acknowledgments

We appreciate helpful comments from Matt Kraft, Eric Taylor, Tim Sass, and anonymous reviewers on previous versions of the article. We are grateful to the New York City Department of Education and the New York State Education Department for the data employed in this article. We appreciate financial support from the 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. The views expressed in the article are solely those of the authors. Any errors are attributable to the authors.

Notes

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

2. A one standard deviation increase in teacher effectiveness is typically 15% to 20% of a standard deviation of student achievement. See Hanushek, Rivkin, Figlio, and 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 (0.11–0.36, but with the average around 0.18 standard deviations) (Aaronson, Barrow, & Sander, 2007; Hanushek & Rivkin, 2010; Jacob & Lefgren, 2008; Kane, Rockoff, & Staiger, 2008; Kane & Staiger, 2008a, 2008b; Koedel & Betts, 2011; Nye, Konstantopoulos, & Hedges, 2004; Rivkin, Hanushek, & Kain, 2005; Rockoff, 2004; Rothstein, 2010).

3. We form quintiles of initial performance employing the average of the first 2 years of value added. See the appendixes 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.

4. These results are available from the authors upon request.

5. 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.

6. The value-added scores depicted in each distribution are each teacher's mean value-added score in years 3, 4, and 5. For brevity, we refer to these scores as "future" performance.

7. For example, typical policy discussions of top and bottom performers often reference 5% to 10% of teachers (see, e.g., Hanushek, 2011).

8. In this example, we categorize a teacher in the bottom third of performance as generally low performing. Any percentile of future performance could be employed to identify low performance. In our example, the most effective teacher in the bottom tercile performs at a level below that of the average new teacher. The selection of a percentile threshold for future low performance has implications for the precision of the estimates. If that percentile was the 25th rather than the 33rd, then we would be more restrictive in our definition of future performance and thus misidentify a higher proportion of teachers. The same implications hold for misidentification of high-performing teachers.

9. The inherent tensions associated with performance-based identification of teachers are present with identification of any segment of teachers for any purpose, although as we show, the consequences of identification likely influence how we weigh the costs of misidentification.

10. It is important to note that some of these "misidentified" teachers are still performing below average in the future. Therefore, even some of the misidentified teachers could benefit from some human capital intervention.

11. Although we are imagining an alternative policy scenario that we think is quite realistic—many districts today do not use value-added scores formally as part of evaluation—it is still the case that the mere availability of these measures could encourage some teachers to self-select out of the profession or for some teachers to be informally counseled out. Therefore, even if there is no formal policy that ties performance measures to teacher retention, there could be some informal sorting based on value added. It is also worth mentioning that most districts do tend to have systems for identifying teachers for dismissal, but given tenure policies, this tends to be a very small group of teachers.

12. In a separate analysis (not shown), we conduct a similar analysis examining the racial breakdown by initial performance, but we separate results across all five quintiles of the distribution of initial performance, rather than simply the top/bottom 5%, 10%, 15%, and 20%. 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. Results are available upon request.

13. Some argue that the gain score model is preferred because one does not place any prior achievement scores that 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, Sass, Lockwood, & Mihaly, 2009).

14. 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 (2008a, 2008b) suggests that student fixed effects estimates may be *more* biased than similar models using a limited number of student covariates.

References

Aaronson, D., Barrow, L., & Sander, W. (2007). Teachers and student achievement in the Chicago public high schools. *Journal of Labor Economics*, *25*(1), 95–135.

Atteberry, A. (2011). *Stacking up: Comparing teacher value added and expert assessment*. Working paper.

Atteberry, A., Loeb, S. L., & Wyckoff, J. (2013). *Teacher attrition from high stakes testing: Strategic behavior or the normal chaos?* Paper presented at the Association for Public

Policy Analysis and Management (APPAM) Fall Conference, Washington, DC.

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 from http://www.nber.org/papers/w14022.pdf.

Boyd, D. J., Grossman, P. L., Lankford, H., Loeb, S., & Wyckoff, J. (2008). *Who leaves? Teacher attrition and student achievement*. Washington, DC: National Bureau of Economic Research.

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.

Boyd, D. J., Lankford, H., Loeb, S., & Wyckoff, J. (2011). Teacher layoffs: An empirical illustration of seniority versus measures of effectiveness. *Education Finance and Policy*, *6*(3), 439–454.

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

Chetty, R., Friedman, J. N., & Rockoff, J. E. (2014). Measuring the impacts of teachers II: Teacher value-added and student outcomes in adulthood. *American Economic Review*, *104*(9), 2633–2679.

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.

Dee, T. S., & Wyckoff, J. (2015). Incentives, selection, and teacher performance: Evidence from IMPACT. *Journal of Policy Analysis and Management*, *34*(2), 267–297.

Donaldson, M. L., & Papay, J. P. (2015). Teacher evaluation for accountability and development. In H. F. Ladd & M. E. Goertz (Eds.), *Handbook of research in education finance and policy*, 174-193. New York: Routledge.

Glazerman, S., & Seifullah, A. (2012). *An evaluation of the Chicago Teacher Advancement Program (Chicago TAP) after four years: Final report*. Washington, DC: Mathematica Policy Research.

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. (2010a). *Is it just a bad class? Assessing the stability of measured teacher performance*. Seattle, WA: Center for Education Data and Research.

Goldhaber, D., & Hansen, M. (2010b). Using performance on the job to inform teacher tenure decisions. *The American Economic Review*, *100*(2), 250–255.

Goldhaber, D., & Theobald, R. (2013). Managing the teacher workforce in austere times: The determinants and implications of teacher layoffs. *Education*, *8*(4), 494–527.

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). Washington, DC: National Bureau of Economic Research.

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. (2011). The economic value of higher teacher quality. *Economics of Education Review*, *30*(3), 466–479.

Hanushek, E. A., Kain, J., O'Brien, D., & Rivkin, S. G. (2005). *The market for teacher quality* (NBER working paper). Washington, DC: National Bureau of Economic Research.

Hanushek, E. A., & Rivkin, S. G. (2010). *Constrained job matching: Does teacher job search harm disadvantaged urban schools?* Washington, DC: 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.

Hill, C. J., Bloom, H. S., Black, A. R., & Lipsey, M. W. (2008). Empirical benchmarks for interpreting effect sizes in research. *Child Development Perspectives*, *2*(3), 172–177.

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., McCaffrey, D. F., Miller, T., & Staiger, D. O. (2013). *Have we identified effective teachers? Validating measures of effective teaching using random assignment*. Seattle, WA: Bill & Melinda Gates Foundation.

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. (2008a). *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. (2008b). *Estimating teacher impacts on student achievement: An experimental evaluation*. Washington, DC: National Bureau of Economic Research.

Kane, T. J., & Staiger, D. O. (2012). *Gathering feedback for teaching, measures of effective teaching project*. Seattle, WA: 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* (Working Paper 708). Columbia: University of Missouri, Department of Economics.

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.

Lefgren, L., & Sims, D. (2012). Using subject test scores efficiently to predict teacher value-added. *Educational Evaluation and Policy Analysis*, *34*(1), 109–121.

Lockwood, J., Louis, T. A., & McCaffrey, D. F. (2002). Uncertainty in rank estimation: Implications for value-added modeling accountability systems. *Journal of Educational and Behavioral Statistics*, *27*(3), 255–270.

Loeb, S., Miller, L. C., & Wyckoff, J. (2014). *Performance screens for school improvement: The case of teacher tenure reform in New York City*. Washington, DC: American Institutes of Research.

Mattern, K. D., & Patterson, B. F. (2014). *Synthesis of recent SAT validity findings: Trend data over time and cohorts*. Retrieved from http://research.collegeboard.org/sites/default/files/info2go/2014/6/Synthesis-of-Recent-SAT-Validity-Findings.pdf

McCaffrey, D. F. (2012). *Do value-added methods level the playing field for teachers?* Stanford, CA: Carnegie Knowledge Network.

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? *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*. Retrieved from http://digitalcommons.ilr.cornell.edu/cgi/viewcontent.cgi?article=1134&context=workingpapers

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.

Raudenbush, S. W. (2013). *What do we know about using value-added to compare teachers who work in different schools?* Retrieved from http://www.carnegieknowledgenetwork.org/wp-content/uploads/2013/08/CKN_Raudenbush-Comparing-Teachers_FINAL_08-19-13.pdf

Rivkin, S. G., 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.

Rothstein, J. (in press). Teacher quality policy when supply matters. *American Economic Review*.

Schochet, P. Z., & Chiang, H. S. (2013). What are error rates for classifying teacher and school performance using value-added models? *Journal of Educational and Behavioral Statistics*, *38*(2), 142–171.

Staiger, D. O., & Rockoff, J. E. (2010). Searching for effective teachers with imperfect information. *The Journal of Economic Perspectives*, *24*(3), 97–117.

Sturman, M. C., Cheramie, R. A., & Cashen, L. H. (2005). The impact of job complexity and performance measurement on the temporal consistency, stability, and test-retest reliability of employee job performance ratings. *Journal of Applied Psychology*, *90*(2), 269.

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

Thomas, J. W., & Hofer, T. P. (1999). Accuracy of risk-adjusted mortality rate as a measure of hospital quality of care. *Medical Care*, *37*(1), 83–92.

Winters, M. A., & Cowen, J. M. (2013). Would a value-added system of retention improve the distribution of teacher quality? A simulation of alternative policies. *Journal of Policy Analysis and Management*, *32*(3), 634–654.

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

Authors

ALLISON ATTEBERRY, is an Assistant Professor at the School of Education, University of Colorado Boulder, 249 UCB, Boulder, CO 80309. She specializes in evaluating the effects of policies and interventions that are intended to provide effective teachers to historically under-served student populations.

SUSANNA LOEB, is a Professor at the Stanford School of Education, Stanford University, 2520 Galvez Mall #524, CERAS building, Stanford, CA 94305-3084. She specializes in the economics of education and the relationship between schools and federal, state and local policies.

JAMES WYCKOFF, is a Professor at the Curry School of Education, University of Virginia, Ruffner Hall 256, PO Box 400277, Charlottesville, VA 22904. His research focuses on teacher labor markets and policies intended to improve teacher quality and student outcomes.