VALIDATING VALUE-ADDED MEASURES OF TEACHER PERFORMANCE[1]

Steven Glazerman
Ali Protik

DRAFT: January 2, 2015

[1] This version prepared for the annual meeting of the American Economic Association, January 5, 2015. This is a preliminary draft for discussion and will be revised. Comments are welcome at sglazerman@mathematica-mpr.com. The authors thank Jonah Rockoff, Tom Cook, Brian Gill, Melissa Clark, and Hanley Chiang for helpful suggestions and Maureen Higgins for research assistance. The research was funded by the U.S. Department of Education's Institute of Education Sciences (IES), but the views do not represent IES, nor has this version been subject to IES review. Errors and views expressed are the authors'.

I. BACKGROUND/MOTIVATION

Measures of teacher performance that estimate a teacher's unique effect on student test scores, referred to as "value-added" measures, have been increasingly relied upon in practice. They are used to evaluate teachers in dozens of states and in some cases used as partial criteria for pay and promotion decisions.

But value-added measures have also received ample criticism. Several concerns have been raised,[2] but one of the central ones, which is the focus of this paper, is bias. True value added is the causal effect that a teacher has on his or her students' test scores independent of factors outside the teacher's control. It is usually expressed in terms of test score performance relative to the counterfactual, defined as the average performance among all other teachers in the reference group – usually the grade and subject peers in the same unit (school district or state). Yet the methods for estimating the value-added parameters amount to one of the weaker forms of nonexperimental analysis to identify causal effects: ordinary least squares regression with controls, typically including teacher fixed or random effects. There is no explicit attempt in value-added models to address selection bias, which is the systematic under- or over-estimation of teacher effectiveness resulting from a selection process through which unobserved factors might determine both student-teacher assignment and test score outcomes.

It is not obvious that one should expect value-added estimators to be unbiased. The observable factors that researchers are typically able to control for include indicators of disability, non-native English language, race/ethnicity, sex, eligibility for free or reduced price lunch (FRL, a crude proxy for family income), and most importantly, prior test scores. The estimators do not capture the range of social, behavioral, and economic attributes that might determine where students attend school, which teachers work in those schools, and which teachers they are assigned to within the school, attributes that also play a role in student learning and test score performance. For example, teachers who seek out highly motivated students might be unfairly given credit for those students' achievement gains while those who altruistically work with students facing profound but hard-to-measure disadvantages would be unfairly labeled as ineffective. In addition, the available measures of family income and student disability are binary measures that fail to capture the full range of income or the severity of the disability. Several studies have documented the tendency for teacher mobility as well as teacher-student assignment to be influenced by student characteristics (Hanushek et al. 2004; Clotfelter et al. 2006; Kalogrides et al. 2012) 1, citation 2. Therefore, it is possible that high value-added scores are less a reflection of teachers' productivity at raising test scores than they are a measure of teachers' tendency (ability or willingness) to teach students in less challenging circumstances.

What is needed to advance the debate in this area is empirical evidence. This paper synthesizes a newly emerging body of evidence that directly tests the hypothesis that value-added estimates are biased measures of true teacher productivity. As described in more detail below, this is done by generating a highly credible (plausibly unbiased) measure of causal effects and comparing the

[2] Critics point out that standardized tests may not capture all important dimensions of teachers' performance, or may not capture them well (Polikoff and Porter 2014). They also raise concerns about statistical precision (Schochet and Chiang 2013), although this issue has been addressed elsewhere (Kane and Staiger 2002; Glazerman et al. 2010).

results against routinely used nonexperimental value-added measures estimated from an earlier period.[3] We show that several different studies using different methods with different samples produced a similar result: value-added estimates of teacher performance were not significantly biased (Kane and Staiger 2008; Kane et al. 2013; Chetty et al. 2014; Bacher-Hicks et al. 2014).[4]

In addition to synthesizing the existing literature, we present results from a new empirical test we conducted of value added validity, this one being the first to exploit randomization to address the possibility of selection *between* schools. The previous literature focuses mainly on bias potentially caused by within-school sorting and selection. This paper seeks to address between-school sorting, which has been identified as a considerable source of inequality of educational opportunity in the U.S. (West and Woessman 2006).

We used data from a multi-site randomized field trial involving monetary incentives that were offered to high-value-added teachers to induce them to transfer to low-achieving schools (Glazerman et al. 2013). The only criterion for selecting teachers to offer them the transfer incentive was that they have high value-added scores.[5] The experiment targeted schools with very low achievement and randomly assigned those with a teaching vacancy to either a treatment group that could hire from the high-value-added pool or a control group where the school principal filled vacancies as they normally would. As a result, assignment to treatment serves as a good instrument variable (IV) for value added because it induced an exogenous increase in the value added scores of teachers in the receiving schools participating in the experiment.

Findings from this exercise suggest, consistent with previous studies, that value-added measures of elementary school teachers' performance are unbiased. We also present middle school results, although these findings are based on a smaller sample, have less statistical precision, and in some cases rely on weak instruments. Nevertheless, we note that the middle school findings are consistent with a story that value added estimates are biased. The experimentally-induced increases in teacher value added did not lead to higher test scores in middle schools. An important limitation for all of our results, especially for middle school, is the imprecision due to a limited sample size for generating IV estimates. This means that we may fail to reject the null hypothesis of no bias even when point estimates of bias are large enough to be potentially meaningful for policy. We return to this and other caveats in the final section of the paper.

Results from all of the empirical tests we review from the existing literature, in addition to the new results presented here, have strengths and limitations, but they tend to be off-setting. For example, our test has a modest-sized sample but has randomization between schools. The test conducted by Chetty et al. uses a much larger sample, but does not have randomization. The test by Kane et al. has somewhat larger sample size and randomization, but only addresses within-school sorting, not between-school sorting. Thus, any one study on its own may present an incomplete picture, but taken together, the findings present a consistent story, which is that researchers repeatedly fail to find bias in value-added measures of elementary school teacher

[3] We focus here on value added models that are used routinely in practice or are likely to be used. Some have also used this approach to search for specifications of value added models that produce the lowest bias.

[4] One additional paper, by Rothstein (2014), was used to present falsification tests arguing against the Chetty et al. test, but happened to replicate the Chetty et al. result using data from North Carolina.

[5] To be more precise, eligible teachers had to have value-added scores that placed them in the highest category of the district's distribution of teachers for whom value added scores could be computed, where the category was typically defined as the top 20 percent.

effectiveness, with forecast bias estimates ranging from 30 percent to less than 3 percent, none being significantly different from zero.

II. CONCEPTUAL FRAMEWORK

Economists have long tried to harness the power of randomized experiments to generate empirical evidence on the magnitude of bias associated with nonexperimental estimators. If we consider the definition of bias as the difference between the expected value of an estimator and the true value of the parameter being estimated ($B(\hat{\mu}) = E[\hat{\mu}] - \mu$), then the basic idea is to estimate the discrepancy between a credibly causal experimental estimate $\hat{\mu}_0$, which is believed to be unbiased and any number of nonexperimental estimates $\hat{\mu}_k$ that are likely to be used in real applications. If the experiment is well executed, then the estimated bias should itself be unbiased, as shown in equation (1), where the expected value of the bias estimate is the difference in expected values of the two estimators, which is the bias.

$$E[\hat{B}(\hat{\mu}_k)] = E[\hat{\mu}_k] - E[\hat{\mu}_0] = E[\hat{\mu}_k - \mu] = B(\hat{\mu}_k) \quad (1)$$

Early examples include comparisons by Lalonde (1986), Fraker and Maynard (1987), and Heckman and Hotz (1989) of nonexperimental and experimental estimates of the impact of the National Supported Work Demonstration. Glazerman et al. (2003) synthesized results from several studies of this type, called design replications studies, all of which were focused on labor market outcomes for job assistance programs (job training, job search assistance, supported work, etc.) The goal of these studies is to better understand the conditions under which nonexperimental or quasi-experimental methods can be used to generate reasonable causal estimates.

More recent applications of this method have focused on education. This is important because the selection process that drives student-school or student-teacher matches may be very different from the selection process by which workers access job training assistance. Moreover, the relationship between those selection factors and labor market outcomes like wages and employment may be different from the relationship between education selection factors and student test score performance.

To formalize this idea in education, Chetty et al. (2014) defined forecast bias as the difference between the *expected* effect of a one-unit change in value-added scores (which is one) and the *observed* effect of a one-unit change. To test for forecast bias, one needs an observation period and a pre-observation period. The pre-observation period is used to estimate the teacher's value added, mu-hat, which forms the expected effect or impact in the observation period. In order to employ the logic of equation (1) above, the observation period measure of teacher performance must be an unbiased measure of the causal effect of a teacher on student achievement. This can be achieved by using random assignment in the observation period, constructing a credible quasi-experiment (i.e. an identification strategy whose identifying assumption is credible), or in the case we discuss in this paper, instrumental variables with randomization as the instrument.

The nonexperimental value added estimator, whose bias we are interested in testing, is usually the teacher effect estimated from an ordinary least squares (OLS) regression of a test score on prior test scores and student characteristics. There are many variations on this type of model and a growing literature on model specification and estimation. We discuss the one we calculated in

Appendix A and below we discuss our specific approach to hypothesis testing as well. What the models share is that they belong to a class of estimators that Heckman and Hotz have referred to as linear control functions.

This paper focuses specifically on selection bias associated with measures of teacher effectiveness. There is a parallel literature that focuses on nonexperimental estimators of *school* value added. These design replication studies typically exploit non-binding lotteries that are used to assign spaces in over-subscribed schools. See Deutsch (2012), Tuttle et al. (2013), Furgeson et. al (2012), Fortson et al. (2012), Angrist et al. (2013), and Deming (2014). These studies have a common limitation of being restricted to students who choose over-subscribed schools, although recent work by Altonji and Mansfield (2014) offers a theoretical reason why controlling for group means can yield unbiased estimates of school (and teacher) effects. They argue that unobservable characteristics differ across schools only because school choosers value school or neighborhood amenities differently. When this is true, controlling for observables of group means can serve as a control for unobservables when estimating group impacts, such as school value added.

III. TESTS OF VALUE-ADDED BIAS FROM THE EMERGING LITERATURE

A series of recent papers, summarized in Table 1, examines the same question raised here. Two of those, one by Kane and Staiger (2008) using data from Los Angeles, and the other by Kane et al. (2013) using data from six districts, are based on randomized experiments. In both of the experiments, teacher pairs (or groups of teachers) within each school were randomly assigned to student rosters, ensuring that there were no systematic differences between the classrooms in any given teacher pair, save the characteristics of those teachers themselves. The authors then compared the difference in test scores between the groups during the experimental observation period to the differences predicted by their prior value added scores in the pre-observation period. If value-added measures are unbiased estimates of a persistent true effect of teachers on student test scores, then a one-unit difference in (nonexperimental) value added should translate into a one-unit difference in scores estimated after random assignment, on average.

Three other papers, by Chetty et al. (2014), Rothstein (2014), and Bacher-Hicks et al. (2014), use a similar logic, but instead of randomization, they use a quasi-experimental method. The quasi-experimental method treats teacher movements in and out of grade-teams as random with respect to future test score performance of the grade-team. By this logic, if a high-value-added teacher changes schools, the subsequent test score performance of the grade from which the teacher leaves should fall and the performance of the grade into which she transfers should rise, with the amounts being proportional to the number of students the teacher was responsible for and the size of her value-added score relative to the other teachers. The teacher-switching quasi-experiments rely on an assumption that, conditional on observable characteristics of students, teachers change grades and schools for reasons that are unrelated to student performance.

mpacts of a Unit of Teacher Value Added

r	Publication Type (as of December 2014)	Site(s) Included	Identification Method	Disaggregation
3	Working paper	Los Angeles	Randomization (within schools)	By subject
3	Foundation report	6 districts: New York (NY), Denver (CO), Dallas (TX), Memphis (TN), Hillsborough County (FL), Charlotte-Mecklenburg (NC)	Randomization (within schools)	By grade span, subject
4	Journal (*American Economic Review*)	Unnamed large school district	Quasi-experimental (teacher-switching)	None
4	Working paper	North Carolina	Quasi-experimental (teacher-switching)	Elementary grades only
4	Working paper	Los Angeles	Quasi-experimental (teacher-switching)	By grade span
5	Working paper	7 large, unnamed districts	Randomization (between schools)	By grade span, subject, and grade span x subject

Besides differing in their method of identification (experimental versus quasi-experimental), the first two (experimental) papers and next three (quasi-experimental) papers differ in the type of selection bias they address. The first two focus on non-random process by which students are assigned to teachers within schools. The next three, which are teacher-switching analyses, primarily address selection *between* schools, although they also include within-school moves. The student populations can differ between teachers in different schools for many reasons, including factors that also help determine students' families' residential location and teachers' and principals' job preferences. Thus the underlying selection for which we seek to control may be different.

The papers are drawn from a variety of contexts. The papers by Kane and Staiger, Bacher-Hicks et al., and Chetty et. al. focus on a single large urban school district (Los Angeles for the first two and an unnamed district for the third). Kane et al. use data from six districts around the country and Rothstein's data use the entire state of North Carolina. Sample sizes from the quasi-experimental studies are large, encompassing several years of data both for value-added estimation and for observing the switching behavior. These large datasets allow for a variety of flexible specifications to test bias. The experimental studies have more modest sample sizes, with the new evidence we present below being especially constrained by sample size.

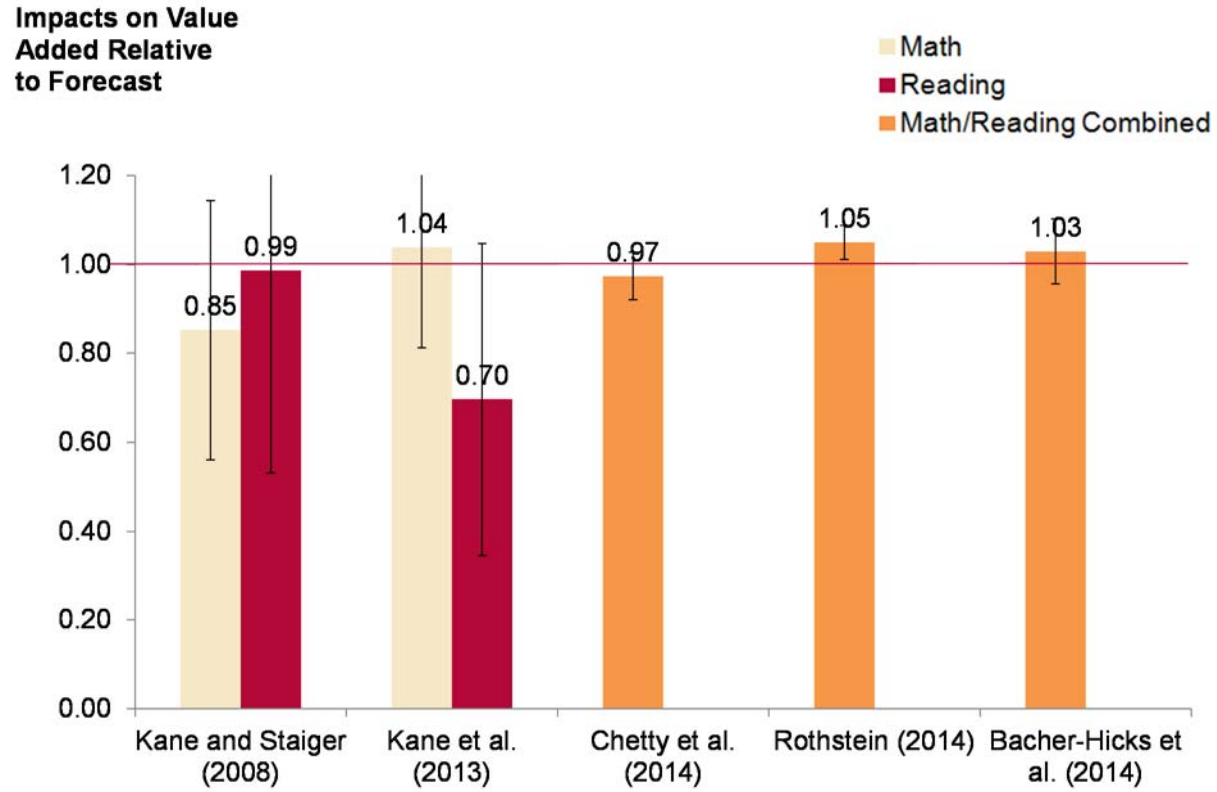
A. Main Findings

Despite the diversity of methods and contexts, the papers each showed that the forecast bias associated with value-added measures was small. (The goal of Rothstein's paper is to discredit Chetty et. al's method for testing for bias, but in the process the author also demonstrates that the different dataset he uses, encompassing the state of North Carolina, replicates Chetty et al.'s main result). Each of the studies conducts a variety of hypothesis tests and robustness checks, looking at different model specifications and sample definitions, but they can each be represented by one or two "main" estimates of the predicted score per unit change in the value added. Figure 1 shows the results for each of the studies mentioned above. As mentioned above, an observed score of 1.0 means that the actual score is exactly the same as the predicted score, implying zero bias.

The results in Figure 1 imply a bias that ranges between 30 percent and 1 percent, with none of the estimates rejecting the null hypothesis of no bias. For example, the Kane and Staiger (2008) study yielded an impact on math scores of a one-unit change in value added of 0.85 points. This implies a forecast bias of 15 percent. For reading[6], a one-point change in value added scores was associated with an experimental impact of 0.99 points, essentially no forecast bias. The 2013 study by Kane and colleagues, also known as the Measures of Effective Teaching, or MET study, showed point estimates with a reverse ordering by subject, with implied bias of 4 percent for math and 30 percent for reading. The other three studies only reported the findings from a "stacked" analysis that treated math and reading results as separate observations. The stacked model estimates are all within 5 percentage points of the benchmark value of 1.0.

[6] Throughout this paper we use "reading" to refer to a subject and teachers of the subject, although the subject is typically called English language arts in middle school grades.

Figure 1. Impact on Test Scores Relative to Value-Added Forecast, by Subject

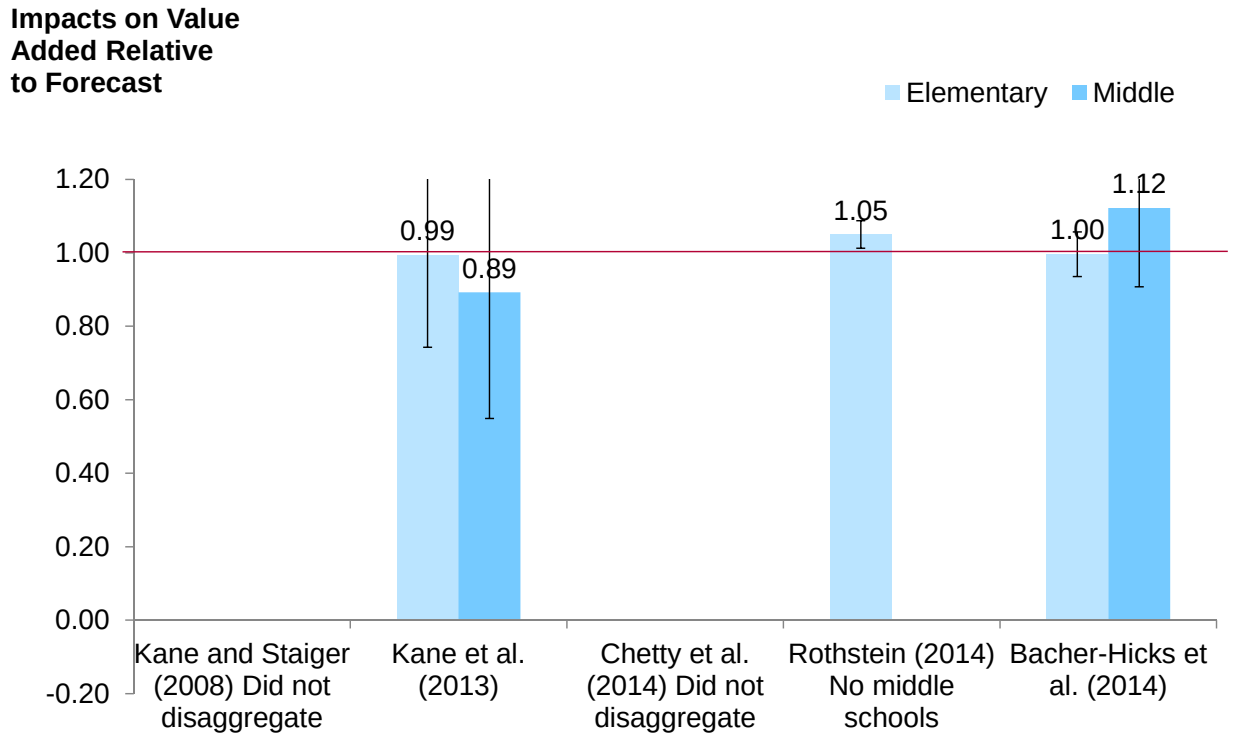


* Difference from 1.0 is statistically significant at the 0.05 level, two-sided test.
Error bars represent 90% confidence intervals.

B. Additional Findings

Figure 1 presents just one or two results from each paper, but the authors of each study presented several additional results. Some disaggregated the findings, as we did below, by grade span, reporting separate findings for elementary grades and middle school grades. There are many reasons why grade span might be important. In elementary school, a single teacher is often responsible for both math and reading, whereas in the middle school analyses typically the math and reading results are based on different teachers who specialize under departmentalized instruction. Also, the nature of sorting might be different as typically elementary schools are smaller and hence draw from smaller and presumably more homogeneous catchment areas than do middle schools. Additionally, the process by which administrators assign students to classrooms may be very different in elementary schools, where there is more often a desire to create balanced classrooms compared to middle schools, which more often offer ability-grouped classes. Figure 2 presents the findings for those studies that presented their results separately by grade span. For the two studies that did so, the degree of forecast bias was not especially different, but the point estimates suggest biases that were slightly larger for middle school than elementary school. We note that the Rothstein results are based on elementary grades only because of difficulty matching teachers to students in middle school grades using the North Carolina data.

Figure 2. Impact on Test Scores Relative to Value-Added Forecast, by Subject



* Difference from 1.0 is statistically significant at the 0.05 level, two-sided test

Error bars represent 90% confidence intervals.

The authors of each of the papers also presented results for a variety of different model specifications, including some that omitted covariates altogether or that included more controls. The findings will be reviewed in a future version of this paper that presents our own findings on different model specifications. The focus of this draft is on the main results.

IV. NEW EVIDENCE USING RANDOMIZATION AS AN INSTRUMENT: DATA AND METHODS

To complement the existing empirical tests of value-added bias, we focus on a recently completed randomized experiment that conveniently induced an exogenous change in the value-added scores of teachers in selected schools in several school districts around the country. First we describe the methods we used and next we describe the original study and the data we use from that study to estimate value-added bias.

A. The Selective Transfer Incentives Study

The experiment was designed to test the effectiveness of a selective transfer incentives intervention, known to participating districts as the Talent Transfer Initiative (TTI). TTI was motivated by the desire to get the most effective teachers to work with the most disadvantaged students. It was implemented in ten school districts in seven states and used a cash bonus to induce each district's highest performing teachers to transfer to selected schools with low test scores.

The TTI intervention worked as follows. The first step was to conduct the value-added analysis to identify the highest-performing teachers, defined as the top 20 percent based on a value-added measure of teachers in tested grades and subjects in each district. The second step was to identify the neediest or lowest-achieving schools and classify them as "potential receiving" schools eligible to receive a high value-added teacher. The rest were "potential sending" schools from which high value-added teachers would transfer to a potential receiving school. In the third step, the highest-performing teachers (identified in the first step) in potential sending schools were offered $20,000 to transfer into and remain in one of the receiving schools in their district. The bonus was paid in installments over two years. High value-added teachers who were already in potential receiving schools were offered $10,000 over two years as a retention incentive.

The experiment was designed around the randomization of receiving schools. At the same time that transfer candidates (those with high value-added scores who were in potential sending schools) were being identified and recruited, principals of potential receiving schools identified likely teaching vacancies in targeted grades (3-5 at the elementary level and 6-8 at the middle school level) and subjects (math and reading) and were then eligible for the experiment. The unit of random assignment was the teacher team—the group of teachers in the same school, grade, and subject—on which there was a vacancy or expected vacancy. The team randomly assigned teacher teams with vacancies to a treatment or control group in the following way. If they learned of multiple eligible teacher teams in the same district at approximately the same time, they matched schools with vacancies in the same grade (and subject, in the case of middle school teams) within the same district. When possible, they also matched schools with vacancies based on their student achievement ranking and the percentage of students eligible for FRL. These matched schools formed blocks; teacher teams were then randomly assigned within each block to either a treatment group (with the opportunity to fill the team's vacancy with a high value-added teacher identified by the study team) or a control group (in which vacancies were filled through whatever process the school would normally use).

Transfers from sending to receiving schools were voluntary, which means that high value-added teachers had to choose to apply, interview, and accept a position, while receiving school principals could decide whom to interview and whether to extend an offer. In practice, transfer-eligible teachers were selective and principals in potential receiving schools were not. Of the high value-added teachers who were invited to apply for a transfer, only 22 percent even submitted a brief application, despite being heavily recruited, and 5 percent ultimately transferred. See Protik et al. (forthcoming) for details.

The study followed students in potential receiving schools (treatment and control) for two years after random assignment. It captured test scores and teacher mobility, as well as survey data from teachers and principals. Surveys captured non-test score outcomes such as principal ratings of their teachers and principal and teacher attitudes and reports on collaboration and collegiality.

B. Data for this Analysis

We use a subsample drawn from the transfer incentives study (Glazerman et al. 2013). The main subsample that we present consists of data from seven of the ten school districts in the experiment, specifically those for which the districts provided the study team with the student-level data to conduct the value-added analysis. The three remaining districts provided the value-added scores that had already been calculated by their outside vendors and did not allow us to

estimate different models.[7] Four of the seven districts in this paper contributed both elementary and middle schools to the study. The other three districts contributed only elementary schools or only middle schools. This is important because the findings differed by grade span.

For one set of results we expanded the subset to include nine of the ten school districts in the original study. For this analysis (see Table 11) we were not able to manipulate the district's value-added measures, which did not control for student background characteristics. Also, the units in which the value-added estimates were given for these two districts did not necessarily correspond to the units in which we later measured test score performance in the experiment. Therefore, we consider this sample for a robustness test only.

Across the seven districts we primarily focus on in this paper, 68 of the potential receiving schools had teams (elementary school-grade combinations or middle school grade-subject combinations) that were randomly assigned to treatment or control status. The average school in this sample was 85 percent low-income (as measured by percent eligible for free or reduced price lunch [FRL]). Teacher teams in the sample ranged from 3rd grade to 8th grade. Some teams included more than one vacancy and some schools included more than one team. Fifty-five teams were assigned to the treatment group and 50 teams were assigned to the control group in the seven districts. Because the study randomly assigned teacher teams within blocks; some blocks had an odd number of teacher teams, random assignment generated an unequal number of treatment and control teams.

Student test scores were scaled as z-scores based on the distribution of test-takers statewide for each grade and subject. By construction, the value-added score was zero for the average teacher in the analysis sample for a given pool within a district. The value added by any given teacher is the amount of extra progress (if positive) that the teacher's students made with that teacher relative to the average teacher in terms of state-level student standard-deviation units.

Average value-added scores for the treatment teachers (the teachers who filled the vacancies in the treatment teams) are higher than those of the control teachers (those who filled the vacancies in the control teams) (Table 2). This is consistent with the fact that the treatment teams had the opportunity to fill their vacancies from the designated pool of high value-added teachers and the control teams did not. The mean value-added score for the treatment teachers of all grades together (grades 3 to 8) for reading was 0.12 standard deviations above the value-added score of the average teacher in the district; for math it was 0.19 standard deviations above the average (column 2 and 3). The mean value-added score for control teachers was lower: 0.03 standard deviations below the score of the average teacher for reading, and 0.07 standard deviations below that of the average teacher for math (column 4 and 5).

Missing Data. Value-added scores were unavailable for many teachers who filled the available control vacancies. Teachers were missing prior value-added scores if they had not taught a tested grade and subject for at least two years or if they could not be reliably matched to student records from the period during which value added was estimated. Only 41 percent of the control teachers in the reading sample and 36 percent of the control teachers in the math sample had prior value-added scores. Because of the possibility that the value-added scores are unavailable in a nonrandom manner (for example, if only the weakest teachers had value-added scores), the main

[7] We were able to use these district-provided value added scores for two of the districts but had to make additional assumptions to convert the teacher performance measures provided to us into units comparable to our value-added models. The analysis that included the two additional districts did not alter the qualitative conclusions.

analysis presented below treats all teachers other than the high-value-added transfer teachers as if they had the average value added of any teacher in the district. In other words, we imputed a value of zero even when we had data for some teachers that allowed us to use a specific value for their prior value added. The mean value-added score for control teachers after this imputation, shown in the last two columns of Table 2, is similar to the ones where we excluded those with missing value-added scores (column 4 and 5). Below, we repeat the analysis using the nonzero value-added scores for control teachers when available and repeat it again trying different arbitrary values of the imputed mean teacher value added, each time assuming that teachers filling control group vacancies were below-average performers instead of merely average. We tried several values, reporting values of -0.02 and -0.05, but our main conclusions were not affected.

Table 2. Average Value-Added Scores of Teachers who Filled the Vacancies in Treatment and Control Teams

	Treatment Teachers		Control Teachers		Control Teachers (with imputed scores) [a]	
	Mean	Sample Size	Mean	Sample Size	Mean	Sample Size
All Grades						
Reading	0.12	48	-0.03	35	-0.01	69
Math	0.19	47	-0.07	31	-0.03	63
Elementary (grades 3–5)						
Reading	0.14	35	-0.02	26	-0.01	48
Math	0.20	35	-0.07	25	-0.04	48
Middle School (grades 6–8)						
Reading	0.07	13	-0.04	9	-0.01	21
Math	0.17	12	-0.07	6	-0.03	15

Source: Estimation by study team from administrative data.

Notes: Value-added scores are in student-level standard-deviation units standardized at the state level.

[a] Value-added scores are imputed as zero for any teacher with a missing value-added score.

Experimental impacts. The transfer incentive study found statistically significant positive effects of the high value-added teachers at the elementary grades. In Table 3, we first compare the experimental effects for the subsample used for this paper with those for the full sample used in the transfer incentive study to convey the degree to which the sub-sample is representative of the larger study. Although the patterns of effects and standard errors in the seven-district sample of this paper are similar to the findings in the transfer incentive study, the specific findings differ in terms of statistical significance. But, similar to the findings for the full sample, there are no statistically significant effects of high-value-added teachers in middle school in the seven-district sub-sample. This previews the null finding that arises when we estimate the impact of value added as an instrument.

Table 3. Effect on Average Test Scores Resulting from Differential Treatment Assignment

	Subsample (7 districts)[a]		Full Sample (10 districts)[b]	
	Effect	Standard Error	Effect	Standard Error
Effects on Math Scores				
All grades	0.06	0.05	0.10*	0.04
Elementary	0.17*	0.06	0.18*	0.05
Middle	-0.01	0.10	0.04	0.09
Effects on Reading Scores				
All grades	0.08*	0.04	0.07	0.04
Elementary	0.07	0.04	0.10*	0.05
Middle	0.05	0.07	0.01	0.05

Notes:

[a] The seven districts in the subsample are the ones with available student-level data and Mathematica-estimated teacher value added. This is the sample used for the main analysis in this paper.

[b] Glazerman et al. (2013) used the full sample of 10 districts for the impact analysis of the transfer incentive study.

* Effect is significantly different from zero at the 0.05 level, two-sided test.

C. Two-Stage Least Squares Estimation Using Instrumental Variables for Estimating the Effects of High-Value-Added Teachers

The important step in assessing the bias in value added is to examine how well the value-added measures, which were estimated in the pre-observation period, related to student test scores in the observation period, nearly one year after teachers had the opportunity to transfer to new schools. If value added were a perfect measure of teacher performance, and if teacher performance were persistent over time and across school settings, then there would be a one-for-one relationship between value added and future test scores. The statistical model to estimate the effects of teacher value added on student test scores is the following:

$$(2) \qquad Y_{ij} = \mathbf{X}\beta + \delta VA_{jk} + \varepsilon_{ij}$$

where Y_{ij} is the test score after random assignment of student i with teacher j, VA_{jk} is the value added estimated before random assignment of teacher j currently in teaching team k, $\mathbf{X}$ is a vector of factors that may influence student test scores, including prior achievement and student background variables, and ε_{ij} is the error term, which captures unobserved determinants of test scores. The coefficient δ represents changes in student test scores for a one-unit change in estimated teacher value added. When estimating equation (2) using data from the low-achieving schools participating in the transfer incentive study, δ measures the effect of a one-unit increase in estimated teacher value added on student test scores, based on the movement of high-value-added teachers into low-achieving schools.

Under ordinary circumstances, one might worry that such movement of teachers into schools is related to the characteristics of the students, such that estimated effects of the mobile teachers would be confounded with the underlying characteristics of the students they moved to teach (or the unobservable characteristics of teachers who chose to make such a move). For instance, we

might expect that high value-added teachers usually move to schools with more motivated students. In this case, VA_{jk} and ε_{ij} in equation (2) would be correlated, and as a result, the estimated coefficient $\hat{\delta}^{OLS}$, using ordinary least squares, would be biased. To avoid this problem, we take advantage of the random assignment process in the transfer incentive study. Teams in low-achieving schools that were randomly assigned to the treatment group had an increased probability of receiving a high-value-added teacher. We use two-stage least squares estimation, where in the first stage we estimate the difference in estimated teacher value added as a result of random assignment to the treatment group using the following model:

(3a) $$VA_{jk} = \mathbf{X}\beta + \pi T_k + \tau_B + v_{jk}$$

where $\mathbf{X}$ is defined as in equation (2), T_k is the random assignment status of team *k*, 1 if treatment and zero if control; τ_B is a fixed effect for randomization block included to account for the block random assignment design of the study; and v_{ik} is the error term. The estimated coefficient $\hat{\pi}$ represents the average difference in estimated teacher value added between the treatment and control teams and is expressed in units of standard deviations of student test score because VA_{jk} is normally expressed in this same unit. The predicted estimated value-added from equation (3a), $\hat{V}A_{jk}$, is then used in the second stage to estimate the relationship between teacher value added and student achievement as follows:[8]

(3b) $$Y_{ij} = \mathbf{X}\beta + \delta\hat{V}A_{jk} + \tau_B + \varepsilon_{ij}$$

The estimation method above is essentially an instrumental variable (IV) estimation method, where we use treatment assignment, T_k, of the team to which student *i* belongs, as an instrument for the endogenous estimated value-added measure of teacher *j* in team *k*, VA_{jk}, assigned to student *i*.

Under the IV approach, the estimated coefficient, $\hat{\delta}^{IV}$, is consistent (approaches the true value of δ as the sample size becomes larger) if the two following conventional IV assumptions are satisfied.

(A1) $$E\left(T_k\varepsilon_{ij}\right) = 0$$

(A2) $$E\left(T_k VA_{jk}\right) \neq 0$$

Assumption A1 says that treatment status is not related to student test scores, except through its influence on the likelihood that a student is taught by a high value-added teacher. This assumption is likely to hold when we estimate the model for all teachers in the teaching teams—because teams were assigned to treatment status at random, treatment status should be uncorrelated with the characteristics of the students taught by the different teams. However, for the main

[8] In practice, equations (3a) and (3b) are estimated simultaneously to obtain correct standard errors in the second stage. We use the ivregress command in Stata with the 2sls option to implement the two-stage least square estimation using equations (3a) and (3b).

analysis we present below, we restrict the analysis to the classrooms within the teaching teams that had a vacancy at the start of the experiment. In this case, assumption A1 takes on a stronger meaning. It implies that the assignment of students and other resources to treatment classrooms within a team is uncorrelated with treatment status. However, in practice, schools in the treatment group may assign the lower ability students within a team to the high value-added teacher to take advantage of that teacher's effectiveness, while the opposite may occur in the control group because many of the teachers filling those vacancies are new to the district or the profession. In that case, treatment status will be correlated with student abilities violating assumption A1. We examine the validity of this assumption below in Part V.

Assumption A2 says that the instrument (treatment status) is correlated with the endogenous variable (estimates of teacher value added). This correlation is directly testable from the results of the regression in the first stage above, equation (3a). We discuss the first-stage regression results in Part V also.

D. Test of Bias in Value-Added Estimates

To test the bias in estimated teacher value added, VA_{jk}, we compared it to an unbiased measure of true teacher performance, which we obtain from the transfer incentive study. We compare the contribution of the teachers who filled the vacancies in the treatment teams to the contribution of the teachers who filled the vacancies in the control teams, and use the difference as the unbiased measure of true teacher performance. This difference represents the effect of the high-value-added teachers when the sample is the students of teachers who filled the vacancies in the randomly assigned teams, and is estimated using the following model:

$$(3c) \qquad Y_{ij} = \mathbf{X}\beta + \gamma T_k + \tau_B + \varepsilon_{ij}$$

Equation (3c) is the reduced-form model of the instrumental variables model presented above, and the estimated coefficient $\hat{\gamma}$ from equation (3c) is the effect of the high-value-added teachers on student achievement. $\hat{\gamma}$ is expressed in units of standard deviations of student test scores, as was the difference in estimated value added of teachers who filled the treatment and control team vacancies estimated in equation (3a), $\hat{\pi}$. We compare $\hat{\pi}$ with the unbiased measure of true teacher performance, $\hat{\gamma}$, to test for the bias in estimated teacher value added. Conceptually, if estimated teacher value added measures true teacher performance relatively well, then $\hat{\pi}$ and $\hat{\gamma}$ will be very close to each other. Their ratio will be close to one if they are similar, and exactly one if they are the same. Mathematically, this ratio is equal to the instrumental variable estimate $\hat{\delta}^{IV}$ obtained from equation (3b). Thus, the bias in estimated teacher value added in predicting student test scores can be measured by the deviation of $\hat{\delta}^{IV}$ from one. We test for the bias after estimating the two-stage least squares equations (3a) and (3b) using the following null hypothesis:

$$(H1) \qquad \hat{\delta}^{IV} = \frac{\hat{\gamma}}{\hat{\pi}} = 1$$

Rejection of the null hypothesis (H1) will imply that the estimated teacher value added is not equal to the unbiased measure of true teacher performance and is therefore a biased measure of

true teacher performance. This is the same as the "forecast bias" in Chetty et al. (2013) and is also the same test used to check for bias in estimated value added in Kane et al. (2013).

V. NEW EVIDENCE USING RANDOMIZATION AS AN INSTRUMENT: FINDINGS

This section presents the findings from the IV analysis of the TTI data. First we discuss the first stage results, which relate to the validity and strength of the instrument. Then we present findings from the second stage, which include the main object of interest, the effect of a one-unit change in value added on student performance. Additional results are presented to explore the robustness of the main results.

A. Validity of the Instrument

The validity of the instrument, the randomized treatment status, hinges on assumptions A1 and A2. Assumption A1 states that the instrument is not correlated with test scores except through the opportunity to hire high-value-added teachers. However, because the unit of random assignment was the team and we restricted our analysis to the teachers who filled the classrooms with vacancies—the focal teachers—this assumption could potentially be violated if student assignment within teams is correlated with treatment status. For example, if lower-ability students are assigned to the high-value-added teachers in the treatment teams but not to the newly hired teachers in the control teams, treatment status could affect student test scores through its correlation with the ability level of students assigned to the teachers in our sample, thus violating assumption A1. In other words, Assumption A1 implies that schools must have assigned students in a similar way to teachers within both the treatment and control teams.

We used administrative data to examine student assignment within teams between focal teachers and nonfocal teachers—teachers who were already teaching in the study teams—and then compared the focal/nonfocal difference between treatment and control teams to check for differential assignment between teams. To examine this relationship, we computed a difference-in-differences measure of student assignment, λ. For any student characteristic, Y, the measure of student assignment, λ, is calculated as follows:

$$\lambda = \left(Y_{focal} - Y_{nonfocal}\right)_{Treatment} - \left(Y_{focal} - Y_{nonfocal}\right)_{Control} \quad (4)$$

Because we were concerned that average difference-in-differences using equation (4) might mask the key phenomenon of interest (large positive or negative values, which could offset each other across teams in different schools), we created the first-differences (focal–nonfocal differences) as a categorical variable and conducted a chi-square test of the independence of first-differences from different treatment status (second-differences) across teacher teams.

We did not find any significant relationship between treatment status and characteristics of students assigned to the focal teachers relative to nonfocal teachers. As Table 4 shows, focal teachers on treatment and control teams were assigned students in a range of different ways, with some teaching students who were lower achieving than those of nonfocal teachers, some teaching similar students, and some teaching higher-achieving students. In terms of prior test scores, greater percentages of more disadvantaged students were being assigned to focal teachers in control teams than in treatment teams. For example, 40 percent of treatment focal teachers versus 27 percent of control focal teachers had students whose prior math scores were greater than 0.10. The

differences, however, were not statistically significant. We found a similar pattern for student disadvantage, measured by FRL, shown in the bottom panel of Table 4. We also examined the distribution of English language learners, students receiving special education services, and students belonging to certain race/ethnicity categories and found no evidence of differential student assignment on treatment and control teams for any of these characteristics.

Table 4. Team-Level Differences between Focal Teachers' and Non-Focal Teachers' Students

Type of Difference	Percentage of Treatment Teams	Percentage of Control Teams	Difference
Difference in Prior Math Scores			
Less than -0.25 (focal teachers assigned lower-scoring students)	20.0	23.5	-3.5
-0.25 to -0.10	24.4	29.4	-5.0
-0.10 to 0.10	15.6	20.6	-5.0
0.10 to 0.25	17.8	11.8	6.0
More than 0.25 (focal teachers assigned higher-scoring students)	22.2	14.7	7.5
Difference in Prior Reading Scores			
Less than -0.25 (focal teachers assigned lower-scoring students)	20.0	26.5	-6.5
-0.25 to -0.10	17.8	17.6	0.1
-0.10 to 0.10	24.4	29.4	-5.0
0.10 to 0.25	17.8	5.9	11.9
More than 0.25 (focal teachers assigned higher-scoring students)	20.0	20.6	-0.6
Difference in Percentage Low Income (FRL)			
More than 10 percent (focal teachers assigned more low-income students)	5.4	0.0	5.4
10 percent to 5 percent	5.4	7.4	-2.0
5 percent to -5 percent	73.0	81.5	-8.5
-5 percent to -10 percent	5.4	7.4	-2.0
Less than -10 percent (focal teachers assigned fewer low-income students)	10.8	3.7	7.1

Source: Administrative data.

Notes: There are 45 teams in the treatment group and 34 teams in the control group. FRL data are available for selected districts (37 treatment teams and 27 control teams). None of the relationships between treatment status and assigned-student difference are statistically significant based on Pearson's chi-square tests of independence.

We used other evidence to address this question as well. We surveyed principals and found no evidence of treatment-control differences in how principals said they assigned students to teachers. We also examined teacher survey data to verify that treatment versus control focal teachers were not reporting different rates at which they believed that they had been assigned academically or behaviorally more challenging students than their peers in the teaching team. These results provide suggestive evidence that assumption A1 was not violated. Nonetheless, if students assigned to treatment and control teachers differed in terms of their unobserved characteristics, assumption A1 would be violated, and our estimates of $\hat{\delta}^{IV}$ would be inconsistent.

Assumption A2, which states that the instrument is highly correlated with the endogenous teacher value-added measure, is directly testable. We report the first-stage coefficients on treatment status from equation (3a), π, and the associated R-squared values and F-statistics in Table 5 by subject and grade span. The estimated coefficient π reported in the first row of Table 5 represents the average differences in value added between the treatment and control focal teachers that resulted from the randomization. For example, in math, the average difference in value added between the treatment and control teachers was 0.2 standard deviation for all grades, including elementary and middle school. Coefficients on the randomized treatment status variable are statistically significant at the 5 percent level for both subjects and all grade spans. Different R-squared values for all of these samples are also high, which implies a strong relationship between the instrument and the endogenous estimated teacher value-added measures. The F-statistics are higher than the conventional threshold of 10 for both subjects in all grades and in elementary grades, suggesting weak instruments are not a concern (Stock et al. 2002). However, for the math sample at the middle school level, the F-statistics is 6.65, lower than the threshold value proposed by Stock and Yogo (2005) for conventional tests of a hypothesis at the 5 percent level of significance. This could result in bias in $\hat{\delta}^{IV}$ in the same direction as the bias in $\hat{\delta}^{OLS}$.

Table 5. Relationship between Randomized Treatment Status and Teacher Value Added (first-stage results)

	Math			Reading			Stacked (Math and Reading)		
	All Grades	Elementary	Middle	All Grades	Elementary	Middle	All Grades	Elementary	Middle
Treatment status	0.21*	0.17*	0.28*	0.10*	0.11*	0.07*	0.15*	0.14*	0.15*
	(0.05)	(0.02)	(0.10)	(0.01)	(0.01)	(0.02)	(0.02)	(0.01)	(0.04)
Student-level controls	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Block fixed effects	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
F-statistic	16.24	91.87	6.65	78.41	101.42	13.82	36.79	142.52	13.57
p-value (F-statistic)	0.000	0.000	0.017	0.000	0.000	0.001	0.000	0.000	0.001
Adjusted R-squared	0.65	0.75	0.66	0.78	0.84	0.75	0.55	0.65	0.53
Partial R-squared	0.40	0.40	0.41	0.57	0.59	0.46	0.36	0.45	0.31

Notes: All specifications include student-level control variables and randomization block fixed effects. Robust standard errors of the treatment status variable are in the parentheses.

* Coefficient is significantly different from zero at the 0.05 level, two-sided test.

B. Estimates of Value Added Bias

In this section, we discuss the instrumental variables estimates from the two-stage model in equations (7a) and (7b) and the proposed test of bias in estimated teacher value added in (H1). The instrumental variables estimates of the coefficients on teacher value added, $\hat{\delta}^{IV}$, are presented in Table 6. A one-unit increase in estimated teacher value added resulted in a 0.28-unit increase in student test scores in math across all grades and 0.78 in reading across all grades. These estimates should be close to 1.0 if estimated teacher value added measures true teacher performance. We formally tested this hypothesis (H1) and failed to reject for reading, but did reject for middle school math and for all grades math when we combined the results.

Table 6. Effect of Estimated Teacher Value Added on Student Test Scores, Using Randomization as an Instrument

	Math		Reading		Stacked (Math and Reading)	
	Effect	Standard Error	Effect	Standard Error	Effect	Standard Error
All grades	0.28#	0.23	0.78	0.41	0.47*#	0.22
Elementary	1.01*	0.33	0.66	0.34	1.01*	0.34
Middle	-0.05#	0.36	0.75	0.99	0.13#	0.27

Note: Effect can be interpreted as the effect of a one-unit increase in teacher value-added score on subsequent student test scores. Shaded cells pertain to results whose first stage F-statistic is less than 10.

* Effect is significantly different from zero at the 0.05 level, two-sided test.

Effect is significantly different from one at the 0.05 level, two-sided test.

The forecast bias itself can be calculated by subtracting each effect estimate in Table 6 from one. Thus the bias for all grades and both subjects is 53 percent, which masks a difference between elementary (1 percent bias, not statistically significant) and middle school (87 percent bias, statistically significant).

One concern about these findings, especially when compared to other estimates in the literature, is the lack of precision with which we can assess the bias. Standard errors in Table 6 range from 0.22 to 0.41 for most estimates, not including middle school reading, where the standard error is nearly one full unit, spanning the range from 0 (value added having no predictive validity) to 1 (value added having no forecast bias). Even the standard errors of 0.22 mean that we would not be able to reject the null hypothesis at the 10 percent level when the bias is as large as 36 percent. The sample sizes supporting these analyses are shown in Table 7. Other papers in the literature we reviewed above did not report comparable numbers of unique students and teachers, but the samples, based on numbers of observations in the data and the number of years involved, appear to be larger, often by orders of magnitude.

Table 7. Sample Sizes for Table 6

	Math		Reading		Stacked (Math and Reading)	
	Students	Teachers	Students	Teachers	Students	Teachers
All grades	3,896	110	4,021	117	8,198	129

Elementary	1,799	83	1,793	83	1,802	83
Middle	2,097	27	2,228	34	4,237	61

An important exercise that others researchers have done is to re-estimate the forecast bias using different specifications of the value added model. A future draft of this paper will report the results of these alternative specifications here for the seven districts where this is possible. We will look at models that include peer effects, different specifications of the pre-test variables, fewer student-level covariates, no adjustment for errors-in-variables induced by the lagged dependent variable, no Empirical Bayes shrinkage, and that mimic a popular approach known as the Colorado Growth Model, which uses pre-test in a highly nonlinear form, but uses no other information such as student background characteristics in estimating teacher effects. We will also test an instrument set that interacts treatment with school district.

C. Robustness Checks

1. Imputing Different Values for Missing Value Added Scores

The analysis presented above treated all teachers other than high value-added transfer teachers (teachers who filled the vacancies in the treatment teams through the transfer incentive study) as if they had the average value added of any teacher in the district. In other words, we imputed a value of zero, as explained in Section C. We re-estimated the instrumental variables model using estimated teacher value added whenever available and zero only when it was missing. The results in Table 8, which use the value-added information if available, lead to the same conclusions, in terms of the test of hypothesis H1, as those we drew from the main result reported in Table 6.

Table 8. Effect of Estimated Teacher Value Added on Student Test Scores, Using Alternative Imputation Rules for Teachers with Missing Value-Added Scores

	Math		Reading		Stacked (Math and Reading)	
	Effect	Standard Error	Effect	Standard Error	Effect	Standard Error
Impute VA = 0 (only if missing)						
All grades	0.23#	0.19	0.73	0.40	0.40#	0.19
Elementary	0.77*	0.26	0.58	0.30	0.82*	0.28
Middle	-0.04#	0.28	0.74	1.02	0.11#	0.25
Impute VA = -0.02 (all non-transfer teachers)						
All grades	0.27#	0.22	0.68	0.35	0.42#	0.20
Elementary	0.92*	0.30	0.57	0.29	0.80*	0.30
Middle	-0.05#	0.34	0.65	0.82	0.12#	0.25

Note: Default method (shown in Table 6) is to impute VA = 0 for all non-transfer teachers. Shaded cells pertain to results whose first stage F-statistic is less than 10.

* Effect is significantly different from zero at the 0.05 level, two-sided test.

Effect is significantly different from one at the 0.05 level, two-sided test.

2. Using Teacher Teams to Allow for Nonrandom Student Assignment

We noted above that the estimates in this paper focus on focal teachers: those who filled the vacancies on teams randomly assigned to the treatment or control group. However, the unit of random assignment was the team, and in order to interpret the main findings presented in Table 6, we required an exogeneity assumption that treatment status was uncorrelated with the error term in equation (3b) (assumption A1), even though principals could have assigned students with higher or lower unobserved ability to focal teachers in treatment teams than they did to teachers in control teams, which will violate the assumption.

However, to avoid having to make the strongest this assumption in its strongest form, we also conducted the same analysis using data from all students on study treatment and control teams including both focal and nonfocal teachers. In this analysis, assumption (A1) is much more plausible, but the average difference in estimated teacher value added between teachers in the treatment and control teams are now diluted and even insignificant for the middle-school math sample (Table 9). Also, the instrument, treatment status, explains little variation in estimated teacher value added

Table 9. Relationship Between Randomized Treatment Status and Teacher Value Added (first-stage results) for All Teachers in Treatment and Control Teams

	Math			Reading		
	All Grades	Elementary	Middle	All Grades	Elementary	Middle
Treatment status	0.05*	0.06*	0.04	0.03*	0.04*	0.02*
	(0.01)	(0.01)	(0.03)	(0.01)	(0.01)	(0.01)
Student-level controls	Yes	Yes	Yes	Yes	Yes	Yes
Block fixed effects	Yes	Yes	Yes	Yes	Yes	Yes
F-statistics	29.10	49.48	7.78	28.17	49.83	6.73
p-value (F-statistics)	0.000	0.000	0.011	0.000	0.000	0.016
R-squared	0.15	0.33	0.08	0.41	0.52	0.22
Adjusted R-squared	0.14	0.32	0.08	0.41	0.51	0.22
Partial R-squared	0.04	0.12	0.02	0.12	0.51	0.07

Notes: All specifications include student-level control variables and randomization block fixed effects. Robust standard errors of the treatment status variable are in the parentheses.

* Coefficient is significantly different from zero at the 0.05 level, two-sided test.

We re-estimated the instrumental variables model including all teachers, focal and nonfocal, in the teaching teams and present these estimates in Table 10. They are comparable to the main results using students of focal teachers, but are less precise. The standard errors exceed 0.67 for both subjects and grade spans, resulting in estimates that are not significantly different from either zero or one except for the all-grades effect on math, which is significantly different from one (and negative, but not different from zero).

Table 10. Effect of Estimated Teacher Value-Added on Student Test Scores for All Teachers in Treatment and Control Teams

	Math		Reading	
	Effect	Standard Error	Effect	Standard Error

All grades	-0.81[#]	0.91	0.83	0.67
Elementary	0.41	0.68	0.17	0.79
Middle	-1.18	1.81	1.18	1.57

Notes: Shaded cells pertain to results where the first stage F-statistic is less than 10.

* Effect is significantly different from zero at the 0.05 level, two-sided test.

[#] Effect is significantly different from one at the 0.05 level, two-sided test.

3. Including Districts with No Student-Level Data

The main analysis in this paper relies on data from seven districts for which we had student-level data to estimate value-added scores for teachers who participated in the study. However, the transfer incentive study included three additional districts, two of which provided us the value-added scores of the teachers in the districts directly. The value-added scores in these districts were reported in teacher standard deviation units. To make the scores comparable to those from the seven districts where we had student-level data and estimated value-added scores calculated in student standard deviation units, we converted the value-added scores in the two additional districts to student standard deviation units. To do this, we estimated value added in teacher standard deviation units in the seven districts and calculated the conversion factor between value-added scores in teacher and student standard deviation units. We then applied this conversion factor to the two additional districts to convert their teacher value-added scores from teacher standard deviation units to student standard deviation units. The instrumental variables estimates from this nine-district sample shown in Table 11 are similar to the main findings based on the original seven-district sample (Table 6). The sample sizes are shown in Table 12, which can be compared to Table 7. The number of students for this analysis is 53 percent greater. The number of teachers is 78 percent greater.

Table 11. Effect of Estimated Teacher Value Added on Student Test Scores, Using Data from Nine Districts

	Math		Reading		Stacked (Math and Reading)	
	Effect	Standard Error	Effect	Standard Error	Effect	Standard Error
All grades	0.54*#	0.19	0.74	0.46	0.57*	0.21
Elementary	1.18*	0.27	0.72	0.58	1.03*	0.33
Middle	0.16#	0.27	0.37	0.75	0.24#	0.24

Notes: Shaded cells pertain to results where the first stage F-statistic is less than 10.

* Effect is significantly different from zero at the 0.05 level, two-sided test.

Effect is significantly different from one at the 0.05 level, two-sided test.

Table 12. Sample Sizes for Table 11

	Math		Reading		Stacked (Math and Reading)	
	Students	Teachers	Students	Teachers	Students	Teachers
All grades	4,588	160	6,508	178	12,504	229
Elementary	2,297	126	3,348	132	6,516	149
Middle	2,475	34	3,160	46	5,988	80

4. Teacher Experience as an Alternative Explanation

In the transfer incentive study, teachers were required to have value-added scores for at least two years to be included in the rankings that identified the highest-performing teachers eligible for transfer incentives. Furthermore, the three-year period used to estimate value added did not include the school year immediately prior to random assignment, so transfer candidates – those with high-value added who were offered the transfer incentive—had at least three and usually more than four years of experience. There were no such requirements for the teachers who filled the vacancies in the control teams and these teachers could have had less experience than those who filled the vacancies in the treatment teams. In fact, 17 percent of the teachers who filled the vacancies in the control teams were new to teaching (Glazerman et al., 2013). Thus, it is possible that the treatment assignment resulted in differences in teacher experiences and the effects on student test scores reported in Table 13 are actually a reflection of the effects of teacher experience.

To test this possibility, we replaced teacher value added with teacher experience as the outcome on the left side of equation (3a). This makes it possible to examine if the random assignment process resulted in statistically significant differences in experience between the teachers who filled the vacancies in the treatment and control teams in the sample of seven districts used for the analysis in this paper. The first-stage results are presented in Table 13. There are indeed statistically significant differences between treatment and control teachers at the middle-school level for both the math and the reading samples, and at all grade levels for the reading sample. However, the partial R-squared values for all the samples are small indicating a weak relationship between the instrument, random assignment, and teacher experience. Also, F-statistics for all the samples are lower than 10, indicating a potentially weak instrument. As noted earlier, a weak instrument can result in bias in the estimated instrumental variables estimates in the same direction as the bias in the OLS estimates. We examined other differences too, such as race/ethnicity of the teacher and did not find meaningful results.

Table 13. Relationship Between Randomized Treatment Status and Teacher Experience (first-stage results)

	Math			Reading		
	All Grades	Elementary	Middle	All Grades	Elementary	Middle
Treatment status	0.77	-1.16	5.94*	3.32*	-1.21	4.30*
	(1.61)	(1.67)	(2.14)	(1.24)	(1.67)	(1.77)
Student-level controls	Yes	Yes	Yes	Yes	Yes	Yes
Block fixed effects	Yes	Yes	Yes	Yes	Yes	Yes
F-statistics	0.20	0.43	7.52	6.28	0.46	5.88
p-value (F-statistics)	0.659	0.517	0.013	0.014	0.502	0.027
R-squared	0.54	0.50	0.74	0.77	0.49	0.96
Adjusted R-squared	0.54	0.48	0.74	0.77	0.48	0.95
Partial R-squared	0.00	0.01	0.18	0.06	0.01	0.27

Notes: All specifications include student-level control variables and randomization block fixed effects. Robust standard errors of the treatment status variable are in the parentheses.

* Coefficient is significantly different from zero at the 0.05 level, two-sided test.

Despite the weak first-stage results, we re-estimated the IV model in equations (7a) and (7b) using teacher experiences as the endogenous variable instead of estimated teacher value added. As shown in Table 14, the IV estimates of the relationship between teacher experience and student test scores are not significant for either subject at any grade level. In other words, an increase in teacher experience by one year has no effect on student test scores. These results could be biased because random assignment is a weak instrument for teacher experience, as shown above, and are thus less reliable. However, the fact that treatment assignment is weakly correlated with teacher experience in the sample of seven districts used for the analysis in this paper mitigates the concern that the effects of high value-added teachers on student test scores are confounded with teacher experience.

Table 14. Effect of Teacher Experience on Student Test Scores, Using Random Assignment as an Instrument

	Math		Reading	
	Effect	Standard Error	Effect	Standard Error
All grades	0.08	0.19	0.02	0.02
Elementary	-0.14	0.20	-0.04	0.06
Middle	-0.01	0.03	0.02	0.02

Notes: Shaded cells pertain to results where the first stage F-statistic is less than 10. None of the effects are significantly different from zero at the 0.05 level based on a two-sided test.

VI. CONCLUSION

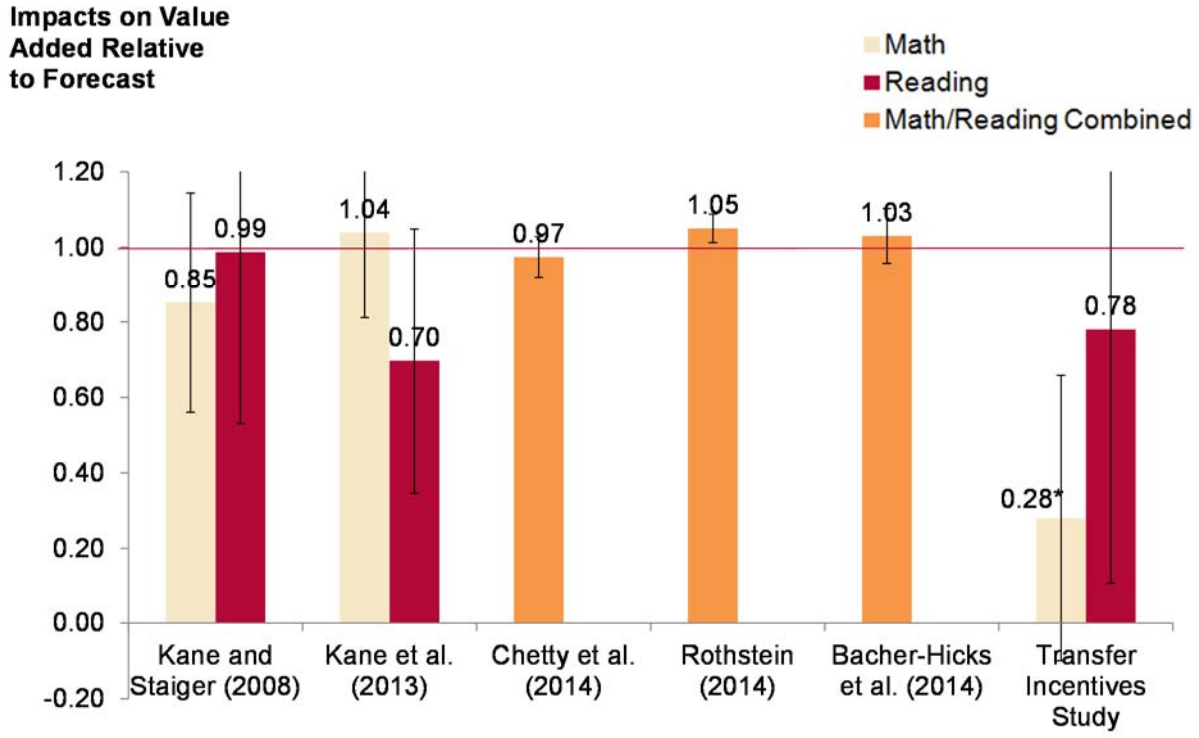
This paper started by documenting a pattern of consistent findings in the emerging literature that value-added estimates have very little forecast bias. That is, researchers who sought to estimate teacher value added in one period and use it to predict test score differences in another period were finding in different contexts with different data and methods that the predictions were accurate, in many cases to within a few percentage points.

We then used a recently completed randomized experiment to generate some additional evidence that complemented the existing evidence base by capturing potential bias in between-school comparisons of teachers. Our findings for elementary schools were consistent with all of the prior literature we reviewed in the sense that we could not reject the null hypothesis that the value added estimates had zero forecast bias. Our confidence intervals were wider than many of the other estimates in the literature, and our findings for middle school, which were especially imprecise, showed that value-added measures of teacher performance in the pre-observation period did *not* predict teacher performance estimated in the experimental observation period.

Figures 3 and 4 show the same results presented above in Figures 1 and 2, but we have added the new findings from the transfer incentive experiment so they can be viewed in context. The error bars help convey the relative precision associated with each test. The clearest result is the one that disaggregates by grade span (Figure 4). In that case, we can say with some confidence, given the agreement between our finding and the other findings in the literature, that value-added measures of *elementary* school teachers produce measures that credibly predict future performance. The latest middle school result, however is an outlier relative to previous studies.

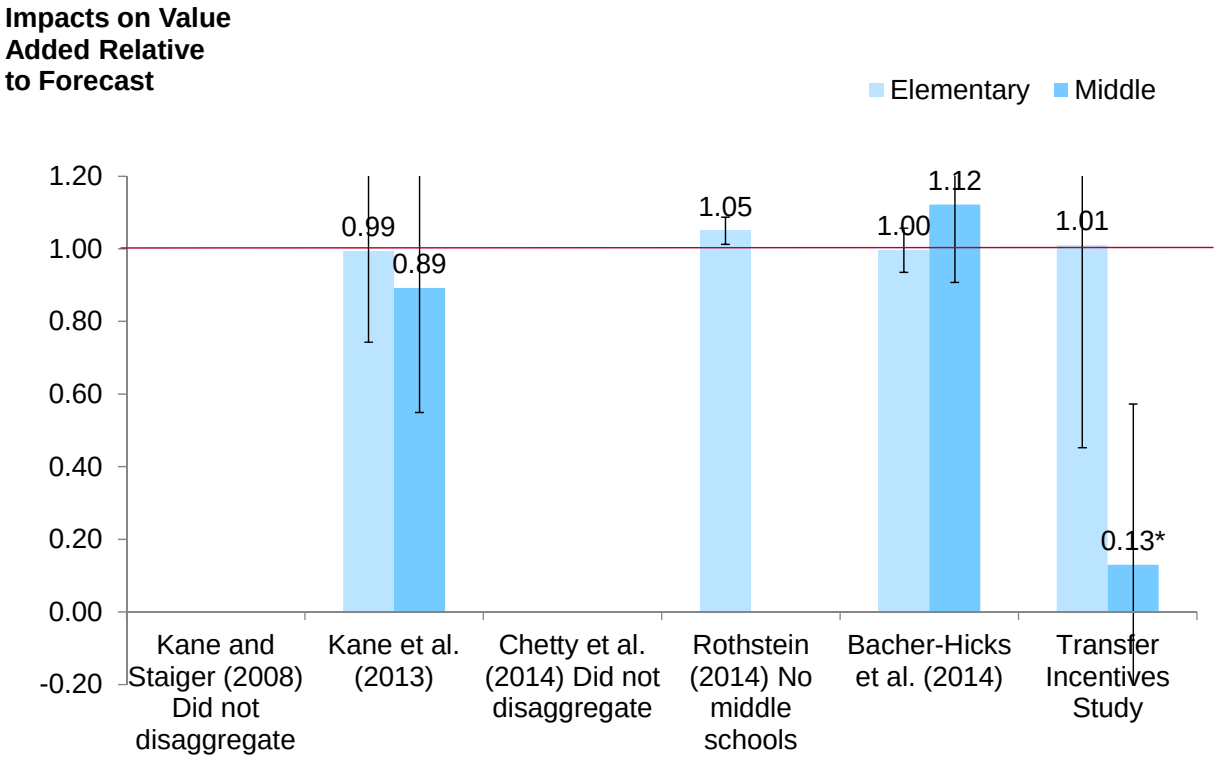
The most striking result is the one in Figure 4 that shows virtually no bias in the elementary schools (forecast = 1.01) and significant bias for middle schools (forecast = 0.13). It is clear that the results of the transfer incentive study reinforce an emerging consensus for elementary school. For middle school the result is far less clear. The sample size is smaller and the statistical precision concomitantly lower. Also the instrument is weaker, as judged by the first stage regression results, casting some doubt on the consistency of the IV estimate. Even if we were to believe the iV estimates of middle school effects, the finding that the pre-observation value added score was not reproduced in the experiment has an ambiguous interpretation. Because the effects were identified on the basis of teachers transferring from high-achieving to low-achieving schools, the inability to forecast post-transfer teacher effectiveness could be a consequence of teaching skills not being transferable between different settings. This possibility makes the high correspondence between value-added and experimental estimators for elementary school all the more remarkable.

Figure 3. Impacts on Value Added Relative to Forecast, by Subject, with Transfer Incentive Results



* Difference from 1.0 is statistically significant at the 0.05 level, two-sided test.
Error bars represent 90% confidence intervals.

Figure 4. Impacts on Value Added Relative to Forecast, by Grade Span, with Transfer Incentive Results



* Difference from 1.0 is statistically significant at the 0.05 level, two-sided test

Error bars represent 90% confidence intervals.

A natural question that arises is why the results differ by grade span. There are several factors that vary with grade span so we can only speculate on which of these might be influential. Elementary schools typically draw from a smaller and presumably more homogenous catchment area for their student populations than middle schools, which tend to be larger and which aggregate groups of feeder elementary schools. Yet this fact might lead one to expect more sorting on unobservables at the elementary school level, not less. On the other hand, elementary school teaching is a different job than middle school. Middle school teachers are specialized by subject, teach larger numbers of students for shorter periods of time, and may teach multiple grade levels and course sections with homogeneous ability groups of students. These factors may make it more difficult for value-added models to capture teacher effectiveness or may make it more difficult for teachers to be successful in very different settings and contexts.

The paper also explored several alternative hypotheses and included robustness tests. A future draft will weigh the estimated bias associated with different types of value-added estimators. Meanwhile, the weight of evidence from emerging studies attempting to validate value-added models suggests bias is not the central concern, at least at the elementary level. The middle school findings provide a qualifier to the otherwise clear consensus, suggesting that more research would be useful, including disaggregation of existing findings into grade spans, to help the field understand the conditions under which value-added measures are more or less likely to be unbiased predictors of future teacher impact on student test scores.

REFERENCES

Altonji, Joseph G., and Richard K. Mansfield. "Group-Average Observables as Controls for Sorting on Unobservables When Estimating Group Treatment Effects: the Case of School and Neighborhood Effects." NBER Working Paper No. 20781. Cambridge, MA: National Bureau of Economic Research, December, 2014.

Bacher-Hicks, Andrew, Thomas J. Kane, and Douglas O. Staiger. "Validating Teacher Effect Estimates Using Changes in Teacher Assignments in Los Angeles." October 2014.

Chetty, Raj, John N. Friedman, and Jonah E. Rockoff. "Measuring the Impacts of Teachers I: Evaluating Bias in Teacher Value-Added Estimates" *American Economic Review* vol. 104, no. 9, pp. 2593-2632. 2014.

Clotfelter, Charles T., Helen F. Ladd, and Jacob L. Vigdor. "Teacher-Student Matching and the Assessment of Teacher Effectiveness." *Journal of Human Resources.* Vol. 41, no. 4, pp. 778-820, Fall 2006.

Deutsch, Jonah. "Using Lotteries to Evaluate the Value-Added Model." Chicago, IL: University of Chicago, October 2012.

Fortson, Kenneth, Natalya Verbitzky-Savitz, Emma Koppa, and Philip Gleason. 2012. *Using an experimental evaluation of charter schools to test whether nonexperimental comparison group methods can replicate experimental impact estimates*. Report 2012–4019. Washington, DC: U.S. Department of Education NCEE.

Fraker, Thomas, and Rebecca Maynard. "The Adequacy of Comparison Group Designs for Evaluations of Employment-Related Programs." *Journal of Human Resources*, vol. 22, no. 2, Spring 1987, pp. 194-227.

Furgeson, Joshua, Brian Gill, Joshua Haimson, Alexandra Killewald, Moira McCullough, Ira Nichols-Barrer, Teh Bing-ru, 2012. *Charter school management organizations: Diverse strategies and diverse student impacts*. Cambridge, MA: Mathematica Policy Research.

Glazerman, Steven, Ali Protik, Bing-ru Teh, Julie Bruch, and Jeffrey Max. "Transfer Incentives for High-Performing Teachers: Final Results from a Multisite Randomized Experiment." (NCEE 2014-4004). Washington, DC: National Center for Education Evaluation and Regional Assistance, Institute of Education Sciences, U.S. Department of Education, 2013.

Glazerman, Steven, Susanna Loeb, Dan Goldhaber, Douglas Staiger, Steven Raudenbush, and Grover Whitehurst. "Evaluating Teachers: The Important Role of Value Added." Washington, DC: Brown Center on Education Policy at Brookings, November 2010.

Glazerman, Steven M., Daniel M. Levy, and David Myers. "Nonexperimental Versus Experimental Estimates of Earnings Impacts." *Annals of the American Academy of Political and Social Science*, vol. 589, September 2003.

Hanushek, Eric A., John F. Kain, and Steven G. Rivkin. "Why Public Schools Lose Teachers." *Journal of Human Resources*. Vol. 39, no. 2, pp. 326-354. 2004.

Heckman, James J., Hidehiko Ichimura, and Petra Todd. "Matching as an Econometric Evaluation Estimator: Evidence from Evaluating a Job Training Programme." *Review of Economic Studies,* vol. 64, 1997, pp. 605-654.

Heckman, James J., Hidehiko Ichimura, Jeffrey C. Smith, and Petra Todd. "Characterizing Selection Bias." *Econometrica,* vol. 66, no. 5, September 1998, pp. 1017-1098.

Isenberg, Eric, and Heinrich Hock. "Design of Value-Added Models for IMPACT and TEAM in DC Public Schools, 2010–2011 School Year." Washington, DC: Mathematica Policy Research, May 2011.

Kalogrides, Demetra, Susanna Loeb, and Tara Beteille. "Systematic Sorting: Teacher Characteristics and Class Assignments." *Sociology of Education*, vol. 86, No. 2, pp. 103-123, 2012.

Kane, Thomas J., Daniel F. McCaffrey, Trey Miller, and Douglas O. Staiger. "Have We Identified Effective Teachers? Validating Measures of Effective Teaching Using Random Assignment." Seattle, WA: Bill & Melinda Gates Foundation, 2013.

Kane, Thomas J., and Douglas O. Staiger. "Estimating Teacher Impacts on Student Achievement: An Experimental Evaluation." NBER Working Paper No. 14607. Cambridge, MA: National Bureau of Economic Research, 2008.

Kane, Thomas J. "Do Value-Added Estimates Identify Causal Effects of Teachers and Schools?" Washington, DC: Brookings Institute Brown Center of Education, October 2014.

Kane, Thomas J., and Douglas O. Staiger (2002). Volatility in school test scores: Implications for test-based accountability systems. In Diane Ravitch (Ed.), *Brookings papers on education policy, 2002* (pp. 235–260). Washington, DC: Brookings Institution

Lalonde, Robert. "Evaluating Econometric Evaluations of Training Programs with Experimental Data." *American Economic Review* vol. 76, no. 4, pp. 604-20. 1986.

Morris, Carl N. "Parametric Empirical Bayes Inference: Theory and Applications." *Journal of the American Statistical Association*, vol. 78, no. 381, 1983, pp. 47–55.

Polikoff, Morgan S., and Andrew C. Porter. "Instructional Alignment as a Measure of Teaching Quality." *Educational Evaluation and Policy Analysis.* Vol. 20, no. 10, May 2014. Pp. 1-18.

Protik, Ali, Steven Glazerman, Julie Bruch, and Bing-ru Teh. "Staffing a Low-Performing School: Behavioral Responses to Selective Teacher Transfer Incentives." Forthcoming in *Education Finance and Policy*.

Rothstein, Jesse. "Revisiting the Impacts of Teachers." University of California-Berkeley, October 2014.

Schochet, Peter Z., and Hanley S. Chiang. 2013. "What are Error Rates for Classifying Teacher and School Performance Using Value-Added Models?" *Journal of Educational and Behavioral Statistics*.vol. 38, no. 2, pp. 141-171.

Staiger, Douglas, and James H. Stock. "Instrumental Variables Regression with Weak Instruments." *Econometrica*, vol. 65, no. 3, 1997, pp. 557–586.

Stock, James H., Jonathan H. Wright, and Motohiro Yogo. "A Survey of Weak Instruments and Weak Identification in Generalized Method of Moments." *Journal of Business and Economic Statistics,* vol. 20, no. 4, 2002, pp. 518-529.

Stock, James H., and Motohiro Yogo. "Testing for Weak Instruments in Linear IV Regression." Chapter 5 in *Identification and Inference for Econometric Models: Essays in Honor of Thomas J. Rothenberg*, edited by J.H. Stock and D.H.K. Andrews, Cambridge, MA: Cambridge University Press, 2005.

Tuttle, Christina Clark, Brian Gill, Philip Gleason, Virginia Knechtel, Ira Nichols-Barrer, and Alexandra Resch. "KIPP Middle Schools: Impacts on Achievement and Other Outcomes." Washington, DC: Mathematica Policy Research, February 2013.

West, Martin R., and Ludger Woessman. "Which School Systems Sort Weaker Students into Smaller Classes? International Evidence." *European Journal of Political Economy.* Vol. 22, no. 4, pp. 944-968, December 2006.

APPENDIX A. VALUE-ADDED ESTIMATION

We estimated value-added measures for three pools of teachers: elementary school teachers, middle school math teachers, and middle school English/language arts (ELA) teachers. The study included elementary school teachers in grades 3 to 5 and middle school teachers in grades 6 to 8. The study used up to three waves of student achievement growth data to estimate teachers' value added.

We dropped all of a teacher's student observations for a particular year from the estimation sample if the teacher was linked to fewer than five students' test scores in that year. The study also excluded from a given teacher's estimation sample any students who spent less than 20 percent of the school year with that teacher.

Teacher value added was estimated using the following equation:

$$Y_{ijt} = \lambda_{t-1} * Y_{ij,t-1} + \alpha_1 * X_{ijt} + \alpha_2 * Z_{jt} + \beta_j * D_{ijt} + e_{ijt} \quad \text{(A.1)}$$

where Y_{ijt} is the post-test score, measured at the end of the school year, for student i who is taught by teacher j in year t; $Y_{ij,t-1}$ is the pre-test score, measured at the end of the previous school year, for that same student, and is assumed to capture previous inputs into student achievement; and e_{ijt} is the error term. X_{ijt} is a vector of control variables that includes the following student-level variables: indicators for gender, race/ethnicity, free or reduced-price lunch (FRL) status, English language learner status, special education status, disability type, whether the student had repeated a grade, and whether the student was old for his or her grade.[9] Z_{jt} includes the following teacher-level variables: percentage of a teacher's students who moved to or from a different class during the school year, percentage of a teacher's students who were were repeating their previous grade, and class size for the teacher's class in which student i was enrolled. Z_{jt} also includes grade-by-year dummies to eliminate any mean differences between grade levels and years. D_{ijt} is a vector of variables reflecting dosage, or the percentage of year t that student i was taught by teacher j (zero if student i was not taught by teacher j in year t), and includes separate values for each teacher-year. The coefficients λ_{t-1}, α_1, α_2, and β_j are parameters to be estimated. The vector β_j, the set of coefficients on the dosage variables, provides the value-added performance measures ("teacher effects") that are the focus of this analysis.

After initial estimation of the teacher effects, the study standardized the subject-specific performance measures (one for math and one for ELA, if applicable) within each grade level.[10] For the purpose of the study, any teachers with fewer than two years of subject-specific performance measures were excluded from the rankings to allow for a better estimate of teachers' "persistent performance" and reduce the influence of transitory performance. Although some

[9] Missing values in $Y_{ij,t-1}$, and X_{ijt} were imputed with predicted values from a regression model. See Glazerman et al. (2013) for details.

[10] Standardization of teacher effects within each grade results in the same mean (zero) and standard deviation (one) for the distribution of estimated teacher effects in each grade. This assumes that the distribution of teacher effectiveness is the same in each grade within a district, but has the benefit of removing any artificial differences associated, for example, with the properties of the assessment instrument and the ways such properties vary by grade.

elementary schools are departmentalized, with different subjects taught by different teachers, the majority of elementary school teachers in the study taught in self-contained classrooms. For these teachers, the study calculated performance measures by taking the average of their math and ELA performance measures. The top 20 to 25 percent of teachers in each of the three pools—elementary school teachers, middle school math teachers, and middle school ELA teachers—were identified as being the highest-performing teachers in their districts, and eligible to receive an incentive to transfer to a low-performing school.

1. Correcting for Measurement Error in Pre-Test Scores

In estimating teacher effects, the study used a two-stage procedure to correct for measurement error in pre-test scores. In the first stage, the study estimated equation (A.1) as an errors-in-variables model using the average published reliability of the test across grades and years to remove the bias caused by the measurement error in the pre-test:[11]

$$(A.2) \qquad Y_{ijt} = \lambda_{t-1} * Y_{ij,t-1} + \alpha_1 * X_{ijt} + \beta_j * D_{ijt} + e_{ijt}$$

Reliability statistics for each test, when available, were obtained from either the test publisher or the school district. The control variables for student background characteristics, X_{ijt}, in equation (A.2) are the same as those used in equation (A.1). Using $\hat{\lambda}_{t-1}$, the estimated value for the coefficient of the pre-test from equation (A.2), the estimated adjusted gain for each student in each year was calculated as follows:

$$(A.3) \qquad \hat{G}_{ijt} = Y_{ijt} - \hat{\lambda}_{t-1} * Y_{ij,t-1}$$

The study then estimated a second-stage regression model that pooled the data from all years and used the adjusted gain as the dependent variable:

$$(A.4) \qquad \hat{G}_{ijt} = \alpha_1 * X_{ijt} + \alpha_2 * Z_{jt} + \beta_j * D_{ijt} + e_{ijt}$$

In equation (A.4), robust standard errors are estimated to account for the correlation in outcomes for students who are in the dataset in more than one year. However, even with robust standard errors, the errors-in-variables correction method for measurement error underestimates the standard errors of β_j because it treats $\hat{\lambda}_{t-1}$ as identical to its true value, λ_{t-1}. If $\hat{\lambda}_{t-1}$ is estimated precisely, the underestimation of the standard errors will be negligible. By substituting equation (A.3) into (A.4), rearranging terms, and treating $\hat{\lambda}_{t-1}$ as λ_{t-1}, we arrive at equation (A.1).

2. Accounting for Imprecision in Estimated Performance Measures Using Shrinkage Estimators

After estimating equation (A.1) to obtain performance measures from the β_j coefficients, a shrinkage procedure outlined in Morris (1983) was applied to calculate empirical Bayes

[11] The errors-in-variables correction works by subtracting the reliability statistic from the diagonal terms of the regression cross-product matrix. The resulting parameters are consistent for the normal distribution. See Isenberg and Hock (2011) for a recent application. We estimated the model using the eivreg command in Stata.

performance measures and standard errors. Using this procedure, the empirical Bayes estimate of each performance measure is approximately the precision-weighted average of the original performance measure (an individual element of the β_j vector) and the mean of all the point estimates (all the elements of β_j), as shown in equation (A.5):

$$\text{(A.5)} \qquad \beta_j^{EB} \approx \left(\frac{\frac{1}{\sigma_j^2}}{\frac{1}{\sigma_j^2} + \frac{1}{\sigma_\beta^2}} \right) \beta_j + \left(\frac{\frac{1}{\sigma_\beta^2}}{\frac{1}{\sigma_j^2} + \frac{1}{\sigma_\beta^2}} \right) \mu_\beta$$

where β_j^{EB} is the empirical Bayes estimate of an element of the β_j vector, β_j is the original point estimate, σ_j is the standard error of the original point estimate, μ_β is the mean of all the point estimates, and σ_β is the standard deviation of all the point estimates.

Due to the precision weighting of the original estimate and the mean of all the point estimates, the empirical Bayes performance measure is designed to place relatively more weight on the mean when the original estimate has a high standard error. This is especially important for a transfer incentive intervention as in Glazerman et al. (2013) because the focus is on the upper tail of the teacher-performance distribution. Random estimation error will vary across teachers when we try to estimate their value added, because they have different numbers of students and their students can be more or less homogeneous, with characteristics that can be more or less similar to the population average. Each of these factors influences the precision of the individual teacher's value-added estimate. Most important, if that precision does vary, the most imprecisely estimated teacher effects will be overrepresented in both tails of the distribution (because the variance in the effect estimates will contain true variation in teacher quality plus a larger error variance). As a result, an intervention like the one in Glazerman et al. (2013) would identify an artificially high number of teachers with small classes or outlier students unless the estimates were corrected. The empirical Bayes shrinkage adjusts the estimates to account for this phenomenon.