NBER WORKING PAPER SERIES

AN EVALUATION OF BIAS IN THREE MEASURES OF TEACHER QUALITY: VALUE-ADDED, CLASSROOM OBSERVATIONS, AND STUDENT SURVEYS

Andrew Bacher-Hicks
Mark J. Chin
Thomas J. Kane
Douglas O. Staiger

Working Paper 23478
http://www.nber.org/papers/w23478

NATIONAL BUREAU OF ECONOMIC RESEARCH
1050 Massachusetts Avenue
Cambridge, MA 02138
June 2017

The research reported here was supported in part by the Institute of Education Sciences, U.S. Department of Education, through Grant R305C090023 to the President and Fellows of Harvard College to support the National Center for Teacher Effectiveness. The opinions expressed are those of the authors and do not represent views of the Institute or the U.S. Department of Education. The views expressed herein are those of the authors and do not necessarily reflect the views of the National Bureau of Economic Research.

At least one co-author has disclosed a financial relationship of potential relevance for this research. Further information is available online at http://www.nber.org/papers/w23478.ack

NBER working papers are circulated for discussion and comment purposes. They have not been peer-reviewed or been subject to the review by the NBER Board of Directors that accompanies official NBER publications.

An Evaluation of Bias in Three Measures of Teacher Quality: Value-Added, Classroom Observations, and Student Surveys

Andrew Bacher-Hicks, Mark J. Chin, Thomas J. Kane, and Douglas O. Staiger

NBER Working Paper No. 23478
June 2017
JEL No. I21,J24

ABSTRACT

There are three primary measures of teaching performance: student test-based measures (i.e., value added), classroom observations, and student surveys. Although all three types of measures could be biased by unmeasured traits of the students in teachers' classrooms, prior research has largely focused on the validity of value-added measures. We conduct an experiment involving 66 mathematics teachers in four school districts and test the validity of all three types of measures. Specifically, we test whether a teacher's performance on each measure under naturally occurring (i.e., non-experimental) settings predicts performance following random assignment of that teacher to a class of students. Combining our results with those from two previous experiments, we provide further evidence that value-added measures are unbiased predictors of teacher performance. In addition, we provide the first evidence that classroom observation scores are unbiased predictors of teacher performance on a rubric measuring the quality of mathematics instruction. Unfortunately, we lack the statistical power to reach any similar conclusions regarding the predictive validity of a teacher's student survey responses.

Andrew Bacher-Hicks
Harvard Kennedy School
79 JFK St.
Cambridge, MA 02138
abacherhicks@g.harvard.edu

Mark J. Chin
Harvard Graduate School of Education
Center for Education Policy Research
50 Church St., 4th Floor
Cambridge, MA 02138
mark_chin@g.harvard.edu

Thomas J. Kane
Harvard Graduate School of Education
Center for Education Policy Research
50 Church St., 4th Floor
Cambridge, MA 02138
and NBER
kaneto@gse.harvard.edu

Douglas O. Staiger
Dartmouth College
Department of Economics
HB6106, 301 Rockefeller Hall
Hanover, NH 03755-3514
and NBER
douglas.staiger@dartmouth.edu

I. Introduction

For decades, researchers have documented heterogeneity in student achievement gains across teachers' classrooms (e.g., Gordon, Kane, & Staiger, 2006; Jacob & Lefgren, 2005; Kane, Rockoff, & Staiger, 2008; McCaffrey, Lockwood, Koretz, Louis, & Hamilton, 2004; Rivkin, Hanushek, & Kain, 2005; Rockoff, 2004). Such findings have inspired recent efforts to measure and reward teacher performance based primarily on three types of measures: student test achievement gains, classroom observations, and student surveys. In fact, between 2008 and 2014, nearly every state reformed their teacher evaluation policies to include one or more of such measures (Minnici, 2014). Yet, legitimate questions have been raised about whether the measures are valid reflections of teachers' performance, or are instead driven by unmeasured characteristics of the students they teach. Although the test-based measures (i.e., value-added estimates) have been most controversial in this regard, the same questions could be raised regarding classroom observations and student surveys.

In this study, we test whether a teacher's performance on each measure, when collected under naturally occurring settings (i.e., non-experimentally), is a valid predictor of that teacher's performance on the same measure following random assignment. The study was conducted over three academic years. In the first two years, we observed a sample of fourth- and fifth-grade mathematics teachers and collected a series of measures—end-of-year student standardized test achievement, classroom observations conducted by external raters, and student surveys. In the third year of the study, we randomly assigned participating teachers to classrooms within their schools and then again collected all three measures. Using those data, we ask two questions regarding the predictive validity of these measures:

1. Do teachers with higher scores on these performance measures during the first two years—when teachers were not randomly assigned to students—score higher in the third year following random assignment?
2. To what degree does the magnitude of teachers' scores on these performance measures from the first two years predict their scores under random assignment?

If the distinctions drawn between teachers during the first two years are largely driven by unmeasured characteristics of their students, we would not expect performance from the prior period to predict performance following random assignment. Moreover, the magnitude of the relationship will illustrate the degree of forecast bias (henceforth referred to simply as "bias") in estimates of teaching performance measured under naturally occurring settings (i.e., when classrooms are not randomly assigned).

To date, predictive validity studies have focused largely on value added.[1] Specifically, two experimental studies randomly assigned teachers to classrooms within schools to assess the predictive validity of these test-based measures of teacher effectiveness. Kane and Staiger (2008) randomly assigned teachers to 156 classrooms within Los Angeles schools and found that prior-year value-added estimates were unbiased predictors of average student test score growth in their randomly assigned classrooms. As part of the Measures of Effective Teaching Project, Kane, McCaffrey, Miller, and Staiger (2013) randomly assigned over 1,100 teachers to classrooms across six school districts. Kane et al. (2013) used scores on state and project-developed tests, classroom observations, and student surveys to form a composite measure of effectiveness, and

[1] One notable exception exploits a nearly random student teacher assignment mechanism for kindergarten students in Ecuador to test the relationship between teacher observation scores and student achievement (Araujo, Carneiro, Cruz-Aguayo, & Schady, 2016). Although they do not explicitly test for forecast bias in performance on teacher effects based on classroom observations, they find that differences in teacher observation scores are predictive of differences in student achievement gains under random assignment.

found that this combined measure was an unbiased predictor of average state test score growth.[2] Glazerman and Protik (2014) conducted the only study so far to use random assignment of teachers across schools to test the validity of teacher value-added estimates in predicting student performance across schools. For elementary teachers, they could not reject the hypothesis that teachers' value-added measures from one school were unbiased predictors of teachers' students' performance after transferring to another school, but they lacked precision to reach any meaningful conclusion for middle school teachers.

In addition to the results from random assignment experiments, three recent studies used a quasi-experimental design to test the predictive validity of teacher effectiveness measures derived from student test performance. This quasi-experimental method, introduced by Chetty, Friedman, and Rockoff (2014a), predicts changes in student test scores using naturally occurring variation in teacher assignments as teachers move from school to school and from grade to grade. Using this method, Chetty et al. (2014a) found that teachers' value-added scores were unbiased estimators of changes in student achievement when there were changes in the specific teachers working in a given grade and subject. Two replication studies applied the same quasi-experiment in different samples and found similar results (Bacher-Hicks, Kane, & Staiger, 2014; Rothstein, 2014).[3]

[2] Kane et al. (2013) also explored the validity of a non-experimental composite measure in predicting student scores on their project-developed supplemental test and on student survey responses following random assignment. However, this composite measure was estimated specifically to predict student state test performance, and not supplemental test performance or student survey responses. With a third year of data in our study, we can create non-experimental estimates tailored specifically to predict supplemental test performance or student survey responses following random assignment.

[3] Rothstein (2014) finds little evidence of forecast bias in value added when replicating the preferred specification in Chetty et al. (2014a). However, Rothstein argues that the quasi-experiment itself is not a valid test, since it fails a placebo test correlating changes in value added with changes in prior test scores. Chetty, Friedman, and Rockoff (2014b) respond to this criticism arguing that the placebo test is influenced by a mechanical effect rather than selection bias.

Because the quasi-experiments utilize large, pre-existing administrative data, which include many teacher transitions over multiple years, these studies generate substantially more precise estimates than the experimental evaluations. However, questions about the additional assumptions of these quasi-experiments—specifically that annual changes in teacher composition do not correspond with changes in student baseline characteristics—have fueled an ongoing debate on the validity of the quasi-experimental test itself (Chetty et al., 2014b; Goldhaber & Chaplin, 2015; Rothstein, 2010, 2014). Because of this, the current study employs random assignment to provide additional experimental evidence on the predictive validity of teacher quality measures.

We make three primary contributions.

First, we provide additional experimental evidence on the predictive validity of value added in a setting where there were high rates of compliance with randomized teacher assignments. Although some level of non-compliance caused by student and teacher movement is unavoidable, previous experiments (i.e., Kane et al., 2013) experienced higher levels of non-compliance, as teachers and students subsequently switched classrooms. In the current study, 71% of students and teachers comply with their randomized assignments. Accordingly, our results are less susceptible to the questions about the generalizability of effects to non-compliers. Our results are consistent with previous evidence: test-based value-added measures are unbiased predictors of teachers' impacts on student achievement following random assignment.

Second, we present the first evidence on the predictive validity of classroom observations and student surveys. Because students are typically not randomly assigned to teachers, such measures are potentially susceptible to the same selection biases as the test-based measures. (It is puzzling that potential bias in classroom observation scores has received so little attention

relative to test-based measures, given that classroom observations typically receive more weight in teacher evaluation systems.) We find evidence that a teacher's score on a classroom observation conducted when students are assigned naturally is an unbiased predictor of the teacher's score on the same rubric when students are assigned randomly. Unfortunately, we lack the statistical power to draw any conclusions about the predictive validity of student surveys collected non-experimentally.

Finally, we use meta-analytic methods to combine the results from the current study with those from the two existing within-school random assignment experiments. By doing so, we provide a more precise, pooled experimental estimate of the predictive validity of value added. The pooled coefficient indicates that value added is a valid predictor of students' average test scores following random assignment, and is more precise than existing experimental evidence.

II. Research Design

We use data collected for the National Center for Teacher Effectiveness (NCTE) study, which was funded by IES to develop and test the validity of multiple measures of teacher effectiveness. The study comprised four large east coast school districts and spanned three school years, from 2010-11 through 2012-13. During all three school years, the study collected data on teachers and students. In the third year, eligible and participating teachers were additionally assigned randomly to classroom rosters. Across the four school districts, 316 fourth- and fifth-grade teachers were eligible for and agreed to participate in at least one of the three years of the study, and 66 teachers were eligible for and agreed to participate in the random assignment portion of the study in the third year.[4]

[4] Several factors contributed to the smaller sample of teachers in the random assignment portion of this study. The most important was teacher movement during the course of the study. Of the 316 teachers who participated in any of

NCTE collected pre-existing administrative data, including classroom rosters, demographic information, and state test scores for all fourth- and fifth-grade students in the four participating districts. In addition, the study collected the following information from students and teachers who were in classrooms participating in the study each year: student test performance on a project-developed low-stakes mathematics test; student responses to a survey probing perceptions of their classroom; digitally-recorded mathematics lessons, used as classroom observations; and teacher responses to a questionnaire about teaching preparation and background.

A. Measures

We used the pre-existing administrative records and the additional data collected by NCTE to generate estimates of teacher performance on five measures: (a) students' scores on state standardized mathematics tests; (b) students' scores on the project-developed mathematics test (Hickman, Fu, & Hill, 2012); (c) teachers' performance on the Mathematical Quality of Instruction (MQI; Hill et al., 2008) classroom observation instrument; (d) teachers' performance on the Classroom Assessment Scoring System (CLASS; La Paro, Pianta, & Hamre, 2012) observation instrument; and (e) students' responses to a Tripod-based perception survey (Ferguson, 2009).[5]

---

the three years in the study, only 132 remained teaching in participating schools in the third year of the study. Such high levels of teacher movement are not atypical, especially in urban districts (Papay, Bacher-Hicks, Page, & Marinell, 2015). Further, leadership changed in some schools, resulting in a smaller number of principals who remained interested in participating in the random assignment part of the study. Of the 132 teachers, 78 remained interested in participating and were in schools where leadership remained interested. Among these 78 teachers, 66 satisfied all other conditions for eligibility (e.g., must teach a class of no fewer than five students with current and baseline state standardized test scores, not be the only remaining teacher in the random assignment block, etc.).

[5] We present reliability estimates for the four project-administered (i.e., non-state test) measures in Appendix Table A1. Cronbach's alpha ranges from 0.78 to 0.91 across these four measures.

Students' scores on state standardized tests came from three different tests, as two of the four participating districts were situated in the same state. There was considerable variability in the format and cognitive demand of items across tests. For example, students in one district took assessments that were completely composed of multiple-choice items, whereas students in another district took assessments with open-ended items that were markedly more difficult (Lynch, Chin, & Blazar, in press). To account for these differences between tests, we rescaled student test scores by district, grade, and academic year using van der Waerden rank-based standardization methods (Conover, 1999).

In conjunction with the Educational Testing Service, NCTE developed a fourth-grade and a fifth-grade mathematics test, designed to align with the Common Core State Standards for Mathematics (National Governors Association Center for Best Practices & Council of Chief State School Officers, 2010) and with the other project-developed measures of teacher quality. The tests contained gridded-response and open-ended items in addition to traditional multiple-choice items. Similar to the state test scores, we standardized these test scores to have a mean of zero and a standard deviation of one, within grade and school year. Because the test was the same across districts (unlike the state test scores), we preserved between-district differences in means and variances.

The MQI and CLASS teacher observation measures were based on up to three video and audio recordings of mathematics lessons per teacher per year. Teachers selected which lessons to record, under the condition that they choose lessons typical of their teaching; on average, recorded lessons lasted approximately one hour.[6] The videos were scored on two established observation instruments: MQI (Hill et al., 2008) and CLASS (La Paro et al., 2012). The MQI

[6] No sanctions or rewards were associated with the study, so teachers had no explicit incentive to strategically select when to record their lessons. Ho and Kane (2013) find that teacher rankings, based on classroom observation, were stable between teachers' self-selected lessons and other lessons.

instrument was designed to assess teacher proficiency in delivering rich, error-free, reform-based mathematics instruction. It comprised 14 codes, which capture practices during instruction such as: using multiple approaches to solve problems; (in)correctly using mathematical language or notation; and remediating student mathematical difficulties (Hill et al., 2008).

The CLASS, which measures general pedagogical practice and is agnostic to the subject-matter of instruction, comprised 12 codes and assesses teachers on more general classroom practices and features such as: the positive or negative climate of the classroom; the quality of the feedback provided by the teacher; the behavior management skills of the teacher; and the level of engagement of students during instruction (La Paro et al., 2012). To generate a teacher's annual overall MQI observation score, we first averaged scores across codes within a lesson, and then across lessons within a year for each teacher. Because the instrument was the same across fourth and fifth grade and across districts, we standardized these teacher-level scores to have a mean of zero and a standard deviation of one within school year. We followed the same procedure to estimate a teacher's CLASS observation score.[7]

Finally, we derived a measure from student responses to a perception survey, comprising 26 Likert-scale items based on the Tripod survey (Ferguson, 2009). These items covered a range of topics, such as whether students felt that: the mathematics presented by his/her teacher was engaging; his/her teacher cared about the students; his/her teacher challenged students to engage with the mathematics; his/her teacher presented mathematical content clearly; his/her teacher regularly assessed the understanding of material presented during lessons; his/her teacher provided useful feedback; or the classroom stayed productive during mathematics lessons. We

[7] MQI and CLASS observation measures are standardized at the teacher level since student-level data do not exist; the measures derived from students' state test scores, project-developed test scores, and survey responses are standardized at the student level.

calculated the simple average across responses to all 26 items and standardized these student-level scores to have a mean of zero and standard deviation of one within school year.

For the three measures with student-level data (i.e., state test scores, project test scores, and Tripod responses), we generated an estimate of teacher quality as the teacher-year level average residuals from the following OLS regression equation:

$$A_{i,k,t} = A_{i,t-1}\alpha + S_{i,t}\beta + P_{k,t}\delta + \varphi_{g,t} + \eta_d + \varepsilon_{i,k,t}, \quad (1)$$

where $A_{i,k,t}$ is the standardized state test score for student *i* taught by teacher *k* during school year *t*. In addition to grade-by-year fixed effects, $\varphi_{g,t}$, and district fixed effects, η_d, we included the following control variables: $A_{i,t-1}$, a cubic polynomial of student *i*'s baseline achievement; $S_{i,t}$, a vector of indicators for gender, race and ethnicity, free-or-reduced lunch eligibility (FRPL-eligibility), limited English proficiency, and special education status; and $P_{k,t}$, a vector of average characteristics of student *i*'s peers in the same class, including average baseline test scores and classroom-level averages of $S_{i,t}$. The student-level idiosyncratic error is ($\varepsilon_{i,k,t}$). As is typical in value-added models, we estimated teacher-year residuals ($\bar{\varepsilon}_{k,t}^{State\_Test}$) by averaging $\varepsilon_{i,k,t}$ across students of teacher *k* in year *t*, which provides an estimate of the performance of teacher *k* in year *t*. We followed an analogous process to generate teacher-year residuals for the project-developed test ($\bar{\varepsilon}_{k,t}^{Project\_Test}$) and for the student survey results ($\bar{\varepsilon}_{k,t}^{Survey}$) using Equation 1 by changing the dependent variable to standardized test scores on the project test or standardized scores on the survey.

Because student-level data did not exist for the two measures based on classroom observations, we generated teacher-year residuals for those measures by fitting the following OLS regression equation:

$$M_{k,t} = P_{k,t}\delta + \eta_d + \varepsilon_{k,t}, \quad (2)$$

where $M_{k,t}$ is a measure of a teacher's classroom observation score and $P_{k,t}$ is the same vector of average characteristics of student *i*'s peers used in Equation 1, which included average baseline test scores and averages of the student characteristics, $S_{i,t}$. We estimated this model separately for the standardized MQI and the CLASS classroom observations scores in order to generate two teacher-year residuals, $\bar{\varepsilon}_{k,t}^{MQI}$ and $\bar{\varepsilon}_{k,t}^{CLASS}$. As we later describe in the Empirical Strategy, we used these teacher-year residuals to generate predictions of teacher performance.

B. Description of the random assignment experiment

During the second year of data collection (2011-12), the NCTE project team worked with staff at participating schools to identify the teachers who met the necessary requirements to be part of the random assignment sample in the third year of the study (2012-13). The primary eligibility conditions to participate in the random assignment part of the study were that (a) teachers had to be part of a group of two or more NCTE project teachers who were scheduled to teach the same grade in that school, and (b) principals had to view the teachers within this group as being capable of teaching any of the classroom rosters designated for the group of teachers without any major adjustments.[8]

Teachers satisfying these two conditions were placed into a randomization block of either two or three teachers. School administrators generated one classroom roster per teacher in each randomization block. For example, if a randomization block had three teachers, school administrators would construct three rosters of students. After school administrators created

[8] In some cases, certain students were required to be paired with specific teachers within a randomization block (e.g., if only one teacher was certified to instruct students with limited English proficiency). In these cases, NCTE staff allowed these students to be paired non-randomly with the appropriate teacher, and then filled the remaining seats in the classroom randomly. We exclude these non-randomly-placed students from all analyses that are restricted to the random assignment sample, but we include them when generating aggregate peer control variables. Approximately 7% of the students in the random assignment classrooms were paired non-randomly.

these rosters, they were submitted to the NCTE study team, who randomly matched eligible teachers with classrooms and then returned the matched rosters to school administrators. Of the 29 total randomization blocks, 21 contained two teachers, and 8 contained three teachers.

In an ideal setting, every randomly assigned student would have been taught by the teacher to whom they were randomly assigned. However, since these rosters were constructed before the start of the random assignment year, a certain amount of movement was unavoidable. In Table 1, we document the disposition of the 1,177 students in the random assignment sample. Notably, 71% of these students remained in their randomly assigned classroom for the entire school year. This is a significant improvement from prior randomization studies. For example, across the six sites described in Kane et al. (2013), between 27% and 66% of students remained in their randomly assigned classroom.

C. Descriptive statistics

In Tables 2 and 3, we present a series of descriptive statistics to examine whether the students and teachers who participated in this study were representative of those from the four NCTE districts. We use data from the first two years of the study, which allows us to examine the characteristics of students naturally assigned by schools to the teachers in our sample, rather than the characteristics of those who were subsequently randomly assigned.

In Table 2, we explore these student characteristics among three distinct subsamples: students assigned to teachers who subsequently participated in random assignment (column 1); students assigned to teachers who participated in the project in any of the three years, but did not participate in random assignment (column 2); students assigned to all other fourth- and fifth-

grade mathematics teachers in these four districts (column 3).[9] Compared to teachers who did not participate in the project (column 3), teachers who participated in random assignment (column 1) were more likely to have been assigned white students and less likely to have been assigned special education and FRPL-eligible students. In addition, the average baseline test score of students assigned to teachers who participated in random assignment was 0.12 student-level standard deviations higher than other project non-random assignment project teachers (column 2) and 0.07 standard deviations higher than non-project teachers (column 3). These results imply that teachers included in the random assignment experiment tended to be assigned—in the years prior to random assignment—classrooms with somewhat more advanced students. This is likely an artifact of our sample design: eligibility required that teachers be capable of teaching any classroom within their randomization block, which generally excluded specialized teachers (e.g., special education teachers). As such, we caution against over-interpreting our results as extending to these teachers.

In Table 3, we present means and standard deviations for our five measures of teacher quality. We again report these statistics for three distinct groups of teachers and use data only from the two years prior to random assignment. In the first row, we report teacher-level residuals based on students' state test performance (i.e., $\bar{\varepsilon}_{k,t}^{State\_Test}$ from Equation 1). Although the average teacher residuals are nearly identical for the project teachers who were not part of the randomized experiment and all other fourth and fifth grade teachers, they were somewhat lower for the teachers who participated in the randomized experiment. Since the other measures were only collected for teachers who participated in the NCTE project, we can only compare them across the random assignment sample and the non-random assignment project sample. Unlike for

[9] Although we determine the three subsamples based on whether teachers participated in the experiment in the third year of the study, all of the student data used in Table 2 comes only from the first two years of the study.

the state test residuals, the random assignment teachers had somewhat higher mean residuals on the project-developed test, but lower mean residuals for the two observation metrics and the student perception surveys.

In the middle panel of Table 3, we present standard deviations for the five measures across the three subsamples of teachers. Although the random assignment group exhibits less variation in the MQI observation measure and the student survey measure, they exhibit similar variation to the other project teachers in the other three measures. Finally, in the lower panel of Table 3, we estimate the persistent component of a teacher's effectiveness estimate (i.e., the "signal" standard deviation) for each of the five measures, by calculating the square root of the covariance in the first two years of the study for each of the five measures. The signal is virtually identical across the three samples for the state test measure. However, for the other four measures the signal is considerably less in the random assignment sample. This suggests that although the random assignment sample of teachers had a similar distribution of "true" teacher effects based on the state test, they were a more homogeneous sample on the other four measures. Unfortunately, this reduced variation in true effects on the other four measures reduces the precision for our validation test, especially for the CLASS observation measure and the surveys.

Finally, in Table 4, we explore whether there were systematic differences in student compliance. We present characteristics of the students who did not remain in their randomly assigned classrooms ("switchers") and compare them to the students who remained in their randomly assigned classrooms for the duration of the 2012-13 school year ("compliers"). Although the switchers and compliers were not statistically significantly different, on average, across seven of the eight observable characteristics in Table 4, the switchers were less likely to

be classified as having limited English proficiency. However, because this result may be driven by multiple hypothesis testing, we also conduct a test of joint significance.[10] The p-value for this joint hypothesis test was 0.40, meaning we could not reject the null hypothesis that these eight characteristics are jointly unrelated to compliance. We also note that most of the student movement occurred early in the school year, before students had much of a chance to experience a teacher's effectiveness (79% of non-compliers moved before the first time NCTE verified classroom rosters in the fall semester) and many students who did not comply with their randomly assigned classroom left the school or the district, which is unlikely to be driven by teacher assignments. We, therefore, conclude that there is little evidence that compliers are substantially different from non-compliers.

III. Empirical Strategy

Our empirical strategy involves two steps. First, we generate estimates of teacher performance in the years prior to the experiment. Second, we evaluate whether these estimates accurately predict teacher performance following random assignment.[11] We describe this empirical strategy in detail below.

A. Predicting teachers' expected performance

[10] To test for joint significance, we use a randomization omnibus test, which is preferred to a conventional F-test with a small number of clusters (Young, 2015). We first draw 1000 block-bootstrapped samples (clustered at the random assignment block level) and within each draw, we reassign each student's compliance status randomly. For each bootstrapped sample, we calculate the conventional F-statistic of joint significance. Then, we generate our p-value as the fraction of bootstrapped null F-statistics that are greater than the actual F-statistic. We use this general approach for joint hypothesis tests throughout the paper.

[11] This strategy assumes that the teachers who participate in the random assignment experiment were not already—in the years prior to random assignment—being assigned classrooms under a process that was random. If the participating teachers were limited to those who were already effectively assigned classrooms at random, then the results of this test would not be generalizable to settings in which classrooms are not randomly assigned. However, similar to Kane et al. (2013), we find strong evidence that the teachers who participated in the random experiment were—in the years prior to the experiment—not already being assigned students under a random process. Please see Appendix B for details.

To generate predictions of teacher performance, we use the two years of data prior to the experiment to identify the best linear combination of teacher measures from one year to predict teacher performance in an adjacent year. Specifically, we use the five teacher-year level measures from the second year of the study to predict each of these five measures in the first year using a separate OLS equation of the following form for each of the five teacher measures:

$$\bar{\varepsilon}^{m}_{k,Y1} = \beta^{m}_{1}\bar{\varepsilon}^{State\_Test}_{k,Y2} + \beta^{m}_{2}\bar{\varepsilon}^{Project\_Test}_{k,Y2} + \beta^{m}_{3}\bar{\varepsilon}^{Survey}_{k,Y2} +$$

$$\beta^{m}_{4}\bar{\varepsilon}^{MQI}_{k,Y2} + \beta^{m}_{5}\bar{\varepsilon}^{CLASS}_{k,Y2} + \beta^{m}_{6}Novice_{k,Y2} + \beta^{m}_{7}Masters_{k,Y2} + \eta^{m}_{d} + \nu^{m}_{k}. \quad (3)$$

The outcome variable, $\bar{\varepsilon}^{m}_{k,Y1}$, is the mean residual for teacher *k* on measure *m* in the first year and the predictor variables, $\bar{\varepsilon}^{State\_Test}_{k,Y2}$, $\bar{\varepsilon}^{Project\_Test}_{k,Y2}$, $\bar{\varepsilon}^{Survey}_{k,Y2}$, $\bar{\varepsilon}^{MQI}_{k,Y2}$, and $\bar{\varepsilon}^{CLASS}_{k,Y2}$, are the average residuals for teacher *k* in the second year. In addition to these predictor variables, we include indicators for novice teachers ($Novice_{k,Y2}$) and teachers with a master's degree ($Masters_{k,Y2}$), and district fixed effects (η^{m}_{d}). The error term is ν^{m}_{k}.

Then, we apply these coefficients from Equation 3 to the teacher-level measures from the second year to generate predictions of teacher performance in the third year—the random assignment year. We generate five different predictions—one for each of the five outcome measures, *m*—and denote them as $\tau^{m}_{k,Y3}$.[12] Note that as measurement error increases, the coefficients will tend to zero. Thus, by generating a linear combination of measures from one year to predict another year, we not only combine information across several types of teacher

[12] For teachers who do not have scores on all five predictor measures, we estimate separate models only including the predictors that were not missing for those teachers. Using this algorithm, we predict a teacher's performance using all of the available information. For example, although all teachers have value-added estimates, teachers entering the project in 2012-13 do not have observation or survey scores in the years prior to random assignment. To generate predictions for these teachers we only include value-added as a predictor and fit the model using all teachers who also had value-added estimates (including those who also may have had scores on other predictors). First-year teachers in 2012-13 do not have any of the five measures of effectiveness from 2011-12 or 2012-13, but we can still generate coarse predictions of their third year effectiveness using indicators for having a master's degree and indicators for the school district in which they teach.

performance measures, we also produce a "shrunken" estimate that accounts for the measurement error (Mihaly, McCaffrey, Staiger, & Lockwood, 2013).

B. Comparing expected quality to actual outcomes following random assignment

To estimate the relationship between our predictions of teacher performance and the actual outcomes following random assignment, we use an instrumental variables (IV) estimator. While we are confident that the effectiveness of a student's *randomly assigned* teacher is not correlated with observable or unobservable student characteristics, we cannot be sure that the effectiveness of the *actual* teacher is not. For example, as we documented in Table 1, although our compliance was higher than in past random assignment studies, there was some reshuffling of students to different classrooms within the school, which may not have been random (e.g., it could have followed the same sorting that occurs in a typical year). Therefore, we instrument for the effectiveness of a student's actual teacher with the effectiveness of their randomly assigned teacher.

We fit our IV model using two-stage least squares. In the first-stage, we estimate the effect of the randomly assigned teacher on the actual teacher's effectiveness using the following equation:

$$\tau_{i,k,Y3}^{m,actual} = \beta\tau_{i,k,Y3}^{m,assigned} + \emptyset_b^m + \nu_i^m, \quad (4)$$

where $\tau_{i,k,Y3}^{m,actual}$ is the predicted performance on outcome measure *m* for the actual teacher of student *i* and $\tau_{i,k,Y3}^{m,assigned}$ is the predicted performance on outcome measure *m* for the teacher randomly assigned to student *i.* Randomization block fixed effects are $\emptyset_b^m$ and the error term is ν_i^m. We perform this procedure for the three measures of teacher performance based on student data. If we had observed perfect compliance, the β coefficient would be one. However, we found

that a one standard deviation unit increase in assigned teacher performance was associated with between a 0.82 and 0.91 standard deviation unit increase in actual teacher performance, across all three measures.

In the second stage, we use the fitted values from Equation 4, $\hat{\tau}_{i,k,Y3}^{m,actual}$, to predict actual student outcomes, $A_{i,Y3}^{m}$, following random assignment:

$$A_{i,Y3}^{m} = \gamma_{IV}^{m}\hat{\tau}_{i,k,Y3}^{m,actual} + \emptyset_{b}^{m} + \nu_{i}^{m}, \quad (5)$$

where $\hat{\gamma}_{IV}^{m}$ is the IV estimate for the three student-level outcomes: state standardized test scores, project-based test scores, and student survey responses.

The IV model described in Equations 4 and 5 requires student-level data, so it cannot be used to estimate the relationship between teachers' predicted and actual observation scores, which vary at the teacher level. Instead, we use the following teacher-level OLS model to estimate the relationship between of teachers' predicted observation scores and their actual observation scores, following random assignment:

$$\bar{\varepsilon}_{k,Y3}^{m} = \gamma_{OLS}^{m}\tau_{k,Y3}^{m} + \emptyset_{b}^{m} + \nu_{k}^{m}. \quad (6)$$

The outcome, $\bar{\varepsilon}_{k,Y3}^{m}$, is the teacher-level residual for measure *m* in the third year described in Equation 2 and the predictor, $\tau_{k,Y3}^{m}$, is the teacher-level prediction for measure *m* from Equation 3. The coefficient $\hat{\gamma}_{OLS}^{m}$is the OLS estimate for the two teacher-level outcomes: MQI and CLASS observations. Although the IV estimates in Equation 5 are not biased by students who do not comply with their random teacher assignments, the OLS estimates in Equation 6 could be biased if students selectively move from their randomly assigned classrooms. As we documented in Table 4, we find little evidence that non-compliance is related to students' observable characteristics. Of course, non-compliance could be related to unobserved differences (e.g., motivation, parental involvement), but the balance on observable characteristics and the

relatively high compliance rates suggest that these OLS estimates are unlikely to be substantially biased.

IV. Results

A. Predicting teachers' expected quality

In Table 5, we report estimates of Equation 3 for the sample of teachers who have data on all five predictor measures. The five columns present the coefficients used to predict each of the five teacher measures, using the best linear combination of all five measures from an adjacent year. By fitting different models to predict each of the five measures, we highlight how these weights differ depending on the measure.

Although the weights differ substantially across the five outcomes, one clear pattern emerges: the most weight is always assigned to the measure that is the same as the outcome. For example, in the first column, we present coefficients from the regression of 2010-11 state test value-added on the five 2011-12 measures. Among the five predictor measures, we find that the 2011-12 state test value-added receives the most weight (0.504) and is statistically significant. The coefficient of 0.504 implies that for each student-level standard deviation that a teacher generated this year, we estimate approximately half a student-level standard deviation next year, controlling for other teacher measures. The other four measures follow a similar pattern: the most weight is placed on the predictor that comes from the same measure as the outcome and the coefficients on these predictors are always statistically significant. However, the magnitude is lower than for the state test, ranging from 0.242 to 0.391. The attenuation in these weights indicates more year-to-year volatility in these four teacher residuals, relative the those derived from the state test.

Although the predictor that comes from the same measure as the outcome is given the most weight, in some cases other predictor variables are also statistically significant. For example, although the 2011-12 project-developed test is the strongest predictor of the 2010-11 project-developed test (0.274), the 2011-12 state test is also a statistically significant predictor (0.184). This highlights the benefit of using additional information from multiple measures to generate predictions.[13]

B. Comparing expected quality to actual outcomes following random assignment

In Table 6, we present the results from Equations 5 and 6, which represent the effect of our prediction of a teacher's effectiveness on the actual outcome following random assignment. We follow Kane et al. (2013) and include controls for students' baseline test scores and background characteristics, but not peer characteristics.[14] Recall from our discussion of our empirical strategy, the predicted teacher outcomes have been "shrunk" to account for measurement error. Thus, we would expect—if there were no bias—that the coefficients on the shrunken predictions of teacher performance to be equal to one.

In the first three columns of Table 6, we report the IV estimates of $\hat{\gamma}_{IV}^{m}$ for the state test, the project-developed test, and student perception scores, based on Equation 5. The most precisely estimated coefficient is in column 1, where we constructed a prediction of teacher

[13] We note that the coefficient on the project-based test is negative (-0.184) when predicting CLASS. Since all five predictor variables are measures of classroom quality, we expect a certain amount of multicollinearity to impact the estimates in Table 5. This would be concerning if we were interested in identifying the causal impacts of the individual predictor variables. However, since this step of the analysis simply solves a prediction problem, the coefficients on each predictor are not a major focus. We caution against the over-interpretation of individual parameter estimates, especially the negative relationship between CLASS and project-based test.

[14] In theory, the coefficients should remain unchanged by the introduction of students' baseline test scores and background characteristics and including additional student-level controls will increase precision. However, we choose not to include peer controls because peers were not randomly assigned and any subsequent non-random student movement of actual peers could introduce bias if we controlled for peers. However, we provide results from a taxonomy of models in Table 7 and the results are robust to each specification.

impacts as it relates to students' state test performance. Using this measure of teacher effects, the coefficient on predicted teacher effectiveness on student achievement is 0.847 with a block-bootstrapped standard error of 0.228.[15] Thus, we are able to reject the hypothesis that this coefficient is zero and fail to reject the hypothesis that this coefficient is one (i.e., the 95% confidence interval of this estimate contains one, but not zero). In other words, we cannot reject the hypothesis that a one-unit increase in a teacher's predicted effectiveness, on average, produces a one-unit change in student state test outcomes after random assignment. This finding is consistent with the two previous within-school random assignment studies (Kane et al., 2013; Kane & Staiger, 2008) and is the third piece of experimental evidence that teacher effect estimates, based on students' state standardized test scores, are a forecast unbiased estimator of student achievement, on average.

In the second column of Table 6, we present the IV estimate using the project-developed test instead of the state standardized test scores.[16] The coefficient on predicted teacher effectiveness on student achievement in the project-developed test is 1.486 with a standard error of 0.267. Similar to the state test outcomes, we reject the hypothesis that this coefficient was zero and fail to reject the hypothesis that this coefficient was one (i.e., the 95% confidence interval of this estimate contains one, but not zero). In the third column, we present the IV estimate using the student perception survey to construct the non-experimental prediction of a teacher's effectiveness and as the post-randomization student outcome. Unfortunately, the point estimate for this measure (-0.228) is imprecise, with a standard error of 0.778. The 95% confidence

[15] Since we have a small number of randomization blocks, we present block-bootstrapped standard errors, which are preferred to cluster-robust standard errors when there are a small number of clusters (Cameron, Gelbach, & Miller, 2008). We use 1000 bootstrap draws, clustered at the random assignment block level. This approach is used throughout the paper to generate standard errors when we cluster at the random assignment block level. The bootstrapped standard errors are slightly larger than those obtained using asymptotic theory.

[16] Note that both the outcome of the IV model changes from the state to the project-developed test and that the prediction of teacher effects is also based on the best linear combination of measures that predict the project-developed test score residuals (see Table 5).

interval of this estimate contains both zero and one and the imprecision of this estimate prevents us from reaching any meaningful conclusions on the validity of this measure. This imprecision is likely due in part to the fact that we are unable to control for students' baseline survey responses in the model represented by Equation 5. While we control for the same vector of variables as we did in the first two columns (including, notably, students' baseline state test scores) these controls fail to explain a substantial portion of variation in students' survey responses. We comment on this limitation and provide suggestions for future research in the conclusion.

In the last two columns of Table 6, we present the OLS estimates of $\hat{\gamma}_{OLS}^{m}$ (from Equation 6) for MQI and CLASS. For MQI, we find that the non-experimental scores have an estimated impact of 0.926 on MQI scores following random assignment, with a standard error of 0.284. For CLASS, the coefficient is 0.877, with a substantially larger standard error of 0.543. Thus, similar to the findings for state and project test scores, we find no evidence that the non-experimental classroom observation predictions are biased for MQI. However, the standard error on the estimate for the CLASS observation metric is large and the 95% confidence interval contains both zero and one. As a result, we are unable to reach any meaningful conclusions on possible presence of bias in those non-experimental estimates based on teacher performance on the CLASS observation metric. As noted above, this is likely driven by the lower "signal" variation in CLASS, relative to the other measures (see Table 3).

In Table 7, we explore the impact of including different student and peer controls. We focus this analysis on our most precisely estimated outcome: the state mathematics test performance. In theory, since students were randomly assigned, the inclusion (or exclusion) of student-level controls should not substantially change our estimates. However, to the extent that additional control variables explain a significant portion of residual variation in the outcome, we

expect precision to improve. In the first column, we only include controls for randomization block, which resulted in a coefficient of 0.755. In the second column, we include controls for students' baseline achievement on state mathematics tests, which produced a coefficient of 0.848. The third column presents our preferred specification (also reported Table 6), which additionally includes controls for students' demographics characteristics and generated a coefficient of 0.847. Finally, in the fourth column, we include additional controls for the average characteristics of the students' in each student's classroom; the coefficient was 0.715.[17] Although all four models generate substantively similar point estimates, the inclusion of baseline student achievement controls (column 2) reduces the standard errors substantially from the model with only fixed effects for random assignment block (column 1), highlighting the importance of including baseline controls for this type of analysis.

C. Combining our evidence with prior studies

In order to facilitate comparison between our study and the existing studies exploring the predictive validity of test-based teacher quality, in Figure 1 we present our main results alongside the main results from the five previous validation studies. Each bar represents the coefficient on teachers' value-added in predicting outcomes following random or quasi-random assignment. A value of one indicates that the non-experimental estimates are unbiased predictors. In addition to point estimates, we plot the 95% confidence intervals around each estimate.

The first three bars in Figure 1 report the findings from previous large-scale quasi-experimental studies that applied the teacher-switching identification strategy proposed by

[17] Although students were randomly assigned, because of student movement following random assignment, the peer group that remained in the classroom is not guaranteed to be random. We explore this possibility in the Threats to Validity section and find that the actual peer characteristics are unrelated to assigned teacher quality. Moreover, in Table 7 we find that our main estimates are similar across all four specifications.

Chetty et al. (2014a) to data on mathematics and English test scores from grades four through eight from three different datasets. Each of these studies provides a precise estimate with the confidence interval including one, implying that, in each independent study, they could not reject the null hypothesis that the non-experimental estimates are unbiased predictors of teacher effects. The next three bars report the findings from the two previous within-school experimental studies and from the current study, each of which are based on randomly assigning mathematics teachers to classrooms. The confidence intervals for each of these estimates also include one, but are substantially wider than the corresponding intervals for the large-scale quasi-experimental studies.

In the last bar in Figure 1, we use meta-analytic methods to combine the results from the three random assignment studies. Using a chi-squared test, we find no evidence of heterogeneity in the coefficient across the three studies (Higgins & Thompson, 2002), suggesting that our results are consistent with the two existing within-school random assignment studies. Because there is no heterogeneity across studies, we generate the pooled estimate reported in the last bar simply as a precision-weighted average of the estimates from the three random assignment studies. The pooled estimate (0.946) is more precise (the standard error is 0.098) and the confidence interval continues to include one. Together, these random assignment studies yield a pooled estimate with precision much closer to the quasi-experimental studies.

V. Threats to Validity

A. Baseline equivalence

In a traditional random experiment, it is common to test for baseline differences in the treatment and control group. Since our analysis relies on many blocks of randomized groups, we

test the degree to which the baseline characteristics of the randomly assigned students were balanced across teachers within a randomization block. Specifically, we estimate the relationship between assigned teachers' predicted effectiveness in 2012-13 and students' characteristics in 2011-12 using an OLS regression of the following form for each of the eight student baseline characteristics, X_i^l, for student *i* and baseline characteristic *l*:

$$X_i^l = \pi_l \tau_{k,Y3}^{state\_test,assigned} + \emptyset_b^l + \varepsilon_i, \quad (7)$$

where $\tau_{k,Y3}^{state\_test,assigned}$ is the predicted effectiveness in 2012-13 of the randomly assigned teacher (based on state test scores) and $\emptyset_b^l$ are fixed effects for randomization blocks. We report the coefficients, $\hat{\pi}_l$, and standard errors for each of eight baseline characteristics in Table 8. We find that none of the eight student characteristics were statistically significantly related to assigned teacher effectiveness.

B. Peer equivalence

The characteristics of the actual peers in a classroom are not randomly assigned in all cases, since non-randomly assigned students enter classrooms throughout the school year and randomly assigned students exit. Thus, our random assignment process does not guarantee that assigned teacher effectiveness is unrelated to actual peer characteristics. To test for the balance of predicted teacher effectiveness by classroom-level averages of peer characteristics, we use the same regression model presented in Equation 7 to examine the relationship between assigned teachers predicted effectiveness and average actual peer characteristics. In Table 9, we present the coefficients on teacher effectiveness. We find that seven of the eight student characteristics are not statistically significantly related to assigned teacher effectiveness, but lower preforming teachers are assigned classrooms with higher concentrations of students with limited English

proficiency. However, a joint hypothesis test that all eight coefficients are zero has a p-value of 0.442, meaning we cannot reject the null hypothesis that these eight characteristics are jointly unrelated to teacher performance.

*C. **Attrition***

Because the random assignment rosters were generated prior to the beginning of the 2012-13 school year, before new students may have signed up and before teachers may have had to be reshuffled to other schools, some amount of student attrition was unavoidable. Although our study maintained a relatively low level of attrition, we examine whether student movement is related to the predicted performance of the randomly assigned teacher. To do so, we estimate a model similar to that in Equation 7, but the dependent variable instead indicates whether the student remained in the sample in 2012-13 and had student outcomes in that year. In Table 10, we present the coefficients from this regression along with the percentage of students who remained in the sample and had achievement scores. As shown in the first column, approximately 88% of students from the random assignment sample have state standardized test scores, approximately 74% of students have scores on the project-developed test and responded to the student survey. In the second column, we present the coefficients on assigned teacher effectiveness from three separate regressions, one for each of these three outcomes. In all cases, the coefficients are not statistically significantly different from zero at the 5% level.

VI. Conclusion

Reforms of teacher evaluation systems have inevitably raised questions about the validity of the performance measures being used. Our findings are the latest in a series of studies

suggesting that the most controversial measure—the test-based value-added measure—is a valid predictor of teacher impacts on student achievement following random assignment.

Until now, much of the public controversy—and all of the predictive validity studies—have focused on the test-based value-added estimates. However, more than two-thirds of teachers are in non-tested grades and subjects, where such value-added measures do not apply (Papay, 2012). As a result, most teachers are evaluated on measures other than test-based value-added measures—such as classroom observations and student surveys. Above, we provide the first evidence testing for bias in classroom observations and student surveys, measuring a teacher's performance before and after students were randomly assigned. Although our estimates of the CLASS observation instrument were too imprecise to draw meaningful conclusions on the validity of that measure, our evidence suggests that the MQI classroom observation measure is, like the test-based value-added measure, an unbiased predictor of teachers' classroom observation following random assignment. In other words, the MQI measure seems to be identifying variation in teaching practice, and does not seem to be biased by the unmeasured characteristics of students the teacher typically teaches. Unfortunately, like the CLASS measure, our evidence on the validity of student surveys was not conclusive.

One of the limitations of this study lies in the imprecision in our validity estimates, particularly for the measure based on student survey responses. As mentioned above, this was due—at least in part—to our inability to control for students' baseline survey responses. Since we did not follow students longitudinally, we were unable to control for any baseline student-level variables that were not contained in administrative records. Based on this, future experiments may consider following students longitudinally in order to control for students' baseline responses on all outcome variables. In addition, the precision on all outcomes will

improve as the sample increases. Future studies may consider testing for bias in classroom observations and student survey responses by applying the quasi-experimental identification strategy of Chetty et al. (2014a) to large administrative databases. Finally, we note that the properties of any measure of teaching performance could change as those measures are used for increasingly high stakes purposes, and so we hope that future work explores the validity of teacher effects under different environments.

References

Araujo, M. C., Carneiro, P., Cruz-Aguayo, Y., & Schady, N. (2016). Teacher quality and learning outcomes in kindergarten. *The Quarterly Journal of Economics*, *131*(3), 1415-1453.

Bacher-Hicks, A., Kane, T. J., & Staiger, D. O. (2014) *Validating teacher effect estimates using changes in teacher assignments in Los Angeles* (No. w20657). Cambridge, MA: National Bureau of Economic Research.

Cameron, A. C., Gelbach, J. B., & Miller, D. L. (2008). Bootstrap-based improvements for inference with clustered errors. *The Review of Economics and Statistics*, *90*(3), 414-427.

Chetty, R., Friedman, J. N., & Rockoff, J. E. (2014a). Measuring the impacts of teachers I: Evaluating bias in teacher value-added estimates. *American Economic Review,* 104(9), 2593-2632.

Chetty, R., Friedman, J. N., & Rockoff, J. E. (2014b). *Response to Rothstein (2014) on "Revisiting the impacts of teachers."* http://obs.rc.fas.harvard.edu/chetty/Rothstein_response.pdf

Conover, W. (1999). *Practical Nonparametric Statistics (Volume 3)*. Hoboken, NJ: Wiley.

Ferguson, R. (2009). *Tripod student survey, MET project upper elementary and MET project secondary versions*. Westwood, MA: Cambridge Education.

Glazerman, S. & Protik, A., (2014). *Validating value-added measures of teacher performance.* https://www.aeaweb.org/aea/2015conference/program/retrieve.php?pdfid=1241

Goldhaber, D. & Chaplin, D. D. (2015). Assessing the "Rothstein Falsification Test": Does it really show teacher value-added models are biased? *Journal of Research on Educational Effectiveness 8*(1), 8-34.

Gordon, R., Kane, T. J., & Staiger, D. O. (2006). *Identifying effective teachers using performance on the job* (The Hamilton Project Policy Brief No. 2006-01). Washington, DC: Brookings Institution.

Hickman, J. J., Fu, J., & Hill, H. C. (2012). *Technical report: Creation and dissemination of upper-elementary mathematics assessment modules.* Princeton, NJ: Educational Testing Service.

Hill, H. C., Blunk, M. L., Charalambous, C. Y., Lewis, J. M., Phelps, G. C., Sleep, L., & Ball, D. L. (2008). Mathematical knowledge for teaching and the mathematical quality of instruction: An exploratory study. *Cognition and Instruction, 26*(4), 430-511.

Higgins, J. P., & Thompson, S. G. (2002). Quantifying heterogeneity in a meta-analysis. *Statistics in Medicine*, *21*(11), 1539-1558.

Ho, A. D., & Kane, T. J. (2013). *The reliability of classroom observations by school personnel. Research paper. MET Project.* Seattle, WA: Bill & Melinda Gates Foundation.

Jacob, B. A., & Lefgren, L. (2005). *Principals as agents: Subjective performance measurement in education* (No. w11463). Cambridge, MA: National Bureau of Economic Research.

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. Research paper. MET Project*. Seattle, WA: Bill & Melinda Gates Foundation.

Kane, T. J., & Staiger, D. O. (2008). *Estimating teacher impacts on student achievement: An experimental evaluation* (No. w14607). Cambridge, MA: National Bureau of Economic Research.

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.

La Paro, K. M., Hamre, B. K., & Pianta, R. C. (2012). *Classroom Assessment Scoring System (CLASS) manual*. Baltimore, MD: Brookes.

Lynch, K., Chin, M., & Blazar, D. (in press). *Relationships between observations of elementary teacher mathematics instruction and student achievement: Exploring variability across districts.* http://scholar.harvard.edu/files/david_blazar/files/lynch_chin_and_blazar_classroom_observations_and_achievement_across_districts_working_paper.pdf

McCaffrey, D. F., Lockwood, J. R., Koretz, D., Louis, T. A., & Hamilton, L. (2004). Models for value-added modeling of teacher effects. *Journal of Educational and Behavioral Statistics*, *29*(1), 67-101.

Mihaly, K., McCaffrey, D. F., Staiger, D. O., & Lockwood, J. R. (2013). *A composite estimator of effective teaching*. Seattle, WA: Bill & Melinda Gates Foundation.

Minnici, A. (2014). The mind shift in teacher evaluation: Where we stand—and where we need to go. *American Educator*, *38*(1), 22-26.

National Governors Association Center for Best Practices & Council of Chief State School Officers. (2010). *Common Core State Standards for Mathematics.* Washington, DC: Authors.

Papay, J. P. (2012). Refocusing the Debate: Assessing the Purposes and Tools of Teacher Evaluation. *Harvard Educational Review*, 81(2), 123-141.

Papay, J. P., Bacher-Hicks, A., Page, L. C., Marinell, W. H. (2015). *The challenge of teacher*

retention in urban schools: Evidence of variation from a cross-site analysis. Retrieved from http://www.appam.org/assets/1/7/Marinell.pdf

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

Rockoff, J. E. (2004). The Impact of individual teachers on student achievement: Evidence from panel data. *American Economic Review*, *92*(2), 247-252.

Rothstein, J. (2014). *Revisiting the impacts of teachers*. Retrieved from http://eml.berkeley.edu/~jrothst/workingpapers/rothstein_cfr.pdf

Rothstein, J. (2010). Teacher quality in educational production: Tracking, decay, and student achievement. *Quarterly Journal of Economics*, *125*(1), 175-214.

Young, A. (2015). *Channelling fisher: Randomization tests and the statistical insignificance of seemingly significant experimental result*s. http://personal.lse.ac.uk/YoungA/ChannellingFisher.pdf

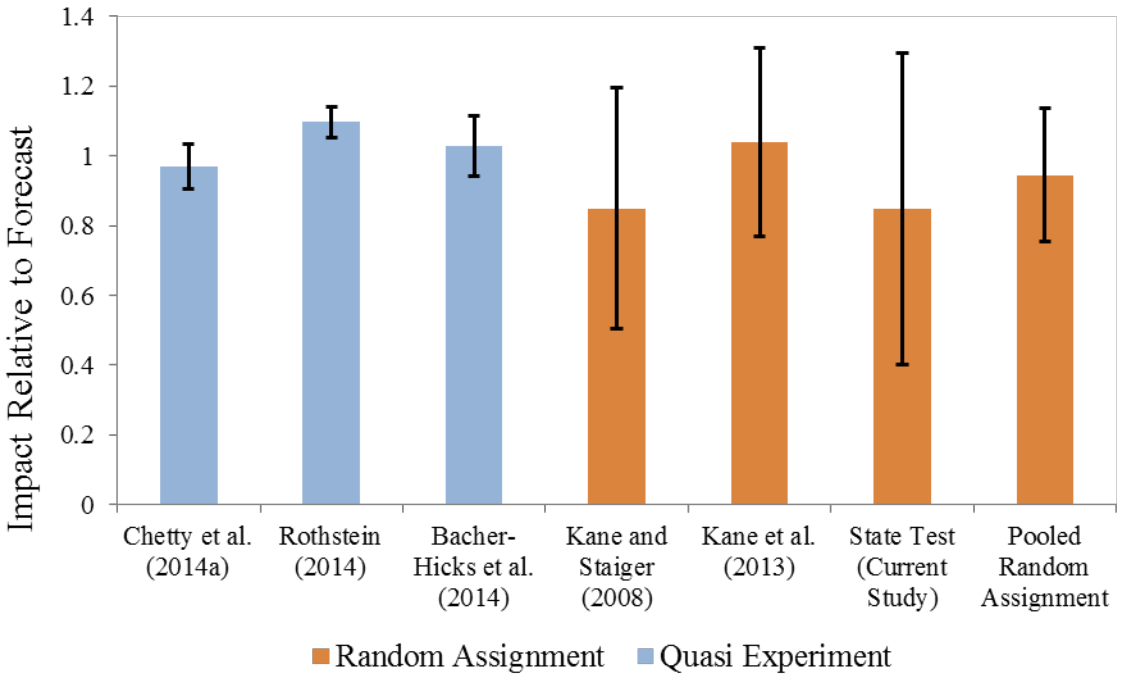

Figure 1
Comparison across studies of the predictive validity of value-added

Notes: This figure presents a visual summary of the key findings from the five prior studies validating teacher effects. In addition, we present the results of the current study using state mathematics test scores and a pooled estimate using precision-weighted average of the estimates from the current study and the two existing random assignment studies. Each column plots the coefficient from a regression of the (quasi-) experimental student test scores outcomes on non-experimental teacher value-added and the associated 95% confidence intervals. A coefficient of one indicates that the non-experimental estimates contain zero bias, on average, in predicting outcomes following random assignment. Estimates and standard errors are collected from reported values in each paper. Since the two random assignment studies report findings separately by mathematics and ELA, we report only the mathematics coefficients. The three quasi-experimental studies only present combined results for mathematics and ELA, so we present the combined estimates across both subjects for those three studies.

Table 1
Summary of students' random assignment compliance

	Number of Students	Percent of Total
Remained with randomly assigned teacher	838	71.2%
Switched teacher within random assignment block	50	4.2%
Switched teacher within school	148	12.6%
Left school or district	141	12.0%
Number of students	1,177	100%

Notes: Sample consists of fourth- and fifth-grade students from the four districts in our sample who were randomly assigned to a classroom in 2012-13.

Table 2

Summary of student characteristics using pre-random assignment data

	Project Teachers in Randomized Sample	Project Teachers not in Randomized Sample	Non-Project Teachers
% Male	51.0%	50.3%	50.2%
% White	31.2%	18.5%	24.5%
% Black	34.3%	46.7%	35.5%
% Hispanic	20.1%	24.9%	27.5%
% FRPL-eligible	58.6%	70.3%	63.0%
% Special education	11.7%	13.3%	14.8%
% Limited English proficiency	16.8%	22.3%	20.4%
Baseline state mathematics test scores	0.122	-0.002	0.053
Number of students	2,433	7,914	50,636

Notes: Sample consists of three mutually exclusive subgroups from the four districts in our sample. The first group (N=2,433) includes all fourth- and fifth-grade students in 2010-11 and 2011-12 who were taught by teachers who participated in the random assignment study in 2012-13. The second group (N=7,914) includes all other fourth- and fifth-grade students in 2010-11 and 2011-12 who were taught by teachers who participated in our study in any of the three years, but did not participate in the random assignment part of the study. The third group (N=50,636) includes all other fourth- and fifth-grade students in 2010-11 and 2011-12 who were taught by teachers in these four districts who did not participate in any of the three years of the study. Although we use teachers' random assignment status in the third year of data to identify these three subgroups, students' demographic and baseline test score data come from the first two years of the study (2010-11 and 2011-12) to explore differences in the type of students assigned to these teachers in a non-experimental setting.

Table 3

Summary statistics for teacher measures using pre-random assignment data

	Project Teachers in Randomized Sample	Project Teachers not in Randomized Sample	Non-Project Teachers
Means			
State test score residuals	-0.014	0.005	0.006
Project test score residuals	0.023	-0.006	
MQI residuals	-0.048	0.014	
CLASS residuals	-0.064	0.018	
Tripod residuals	-0.041	0.015	
Standard Deviation			
State test score residuals	0.262	0.271	0.282
Project test score residuals	0.221	0.259	
MQI residuals	0.683	0.843	
CLASS residuals	0.877	0.876	
Tripod residuals	0.355	0.437	
Signal Standard Deviation (square root of year-to-year covariance)			
State test score residuals	0.198	0.197	0.197
Project test score residuals	0.120	0.168	
MQI residuals	0.270	0.566	
CLASS residuals	0.151	0.486	
Tripod residuals	0.075	0.263	
Number of teachers	61	243	1,763

Notes: Sample consists of three groups of fourth- and fifth-grade teachers. The first group (N=61) includes all teachers who participated in our random assignment study (in the 2012-13 school year) for whom we have data in at least one of the two years before random assignment. Note that five of the 66 teachers randomized in 2012-13 were new teachers in 2012-13, so they are not included in this table. The second group (N=243) includes all other fourth- and fifth-grade teachers who participated in our study in any of the three years, but did not participate in the random assignment part of the study. The third group (N=1,763) includes all other fourth- and fifth-grade teachers in these four districts who did not participate in any of the three years of the study. We use the years prior to random assignment (2010-11 and 2011-12) to explore differences in teacher quality using non-experimental data. The state test, project test, and survey residuals are in student-level standard deviation units, while the two classroom observation measures are in teacher-level standard deviation units.

Table 4

Comparison of random assignment compliers vs. switchers

	Compliers		Switchers		
	Mean	% Non-Missing	Mean	% Non-missing	Mean Difference
% Male	49.3%	99.5%	49.6%	98.8%	-0.30%
% White	22.6%	99.5%	21.8%	98.8%	0.80%
% Black	40.7%	99.5%	42.5%	98.8%	-1.80%
% Hispanic	22.6%	99.5%	18.6%	98.8%	4.00%
% FRPL-eligible	67.8%	99.5%	62.1%	98.8%	5.7%*
% Special education	6.3%	99.5%	7.5%	98.8%	-1.20%
% Limited English proficiency	18.3%	99.5%	10.7%	98.8%	7.6%***
Baseline state math test scores	0.111	92.2%	0.160	89.4%	-0.049
p-value on joint hypothesis test					0.401
Number of students	838	838	339	339	

Notes: Sample consists of two groups of fourth- and fifth-grade students from the four districts in our sample. The first group (N=838) includes students in 2012-13 who remained in classrooms to which they were randomly assigned for the duration of the school year. The second group (N=339) includes students in 2012-13 who did not remain in the classroom to which they were randomly assigned. Statistical significance of differences is based on block bootstrapped standard errors with 1000 draws, clustered at the random assignment block level. The p-value on the joint null hypothesis test is generated using a Fisher-style permutation test, where we draw 1000 block bootstrapped (clustered at the random assignment block level) and we re-assign each student's compliance status randomly. We calculate the F-statistic in each 'null' bootstrapped sample, and the p-value presented in this table is the fraction of null F-statistics that are greater than the actual F-statistic. *** $p<0.01$, ** $p<0.05$, * $p<0.1$

Table 5

Using teacher performance measures from 2011-12 to predict teacher performance measures from 2010-11

	2010-11 State Test Value-Added	2010-11 Project Test Value-Added	2010-11 Tripod Survey	2010-11 MQI Observation	2010-11 CLASS Observation
2011-12 State Test Value-Added	0.504***	0.184**	0.095	0.033	0.051
	(0.085)	(0.092)	(0.097)	(0.093)	(0.102)
2011-12 Project Test Value-Added	0.172*	0.274***	0.010	0.081	-0.184**
	(0.092)	(0.093)	(0.109)	(0.114)	(0.091)
2011-12 Tripod Survey	-0.138*	0.004	0.391***	0.096	0.069
	(0.081)	(0.048)	(0.103)	(0.092)	(0.093)
2011-12 MQI Observation	0.211	-0.007	-0.112	0.277***	0.113
	(0.291)	(0.023)	(0.095)	(0.083)	(0.081)
2011-12 CLASS Observation	0.090	0.031	0.005	0.048	0.242***
	(0.056)	(0.027)	(0.086)	(0.081)	(0.075)
Indicator for Master's Degree	0.056	0.064	-0.316*	-0.058	-0.145
	(0.046)	(0.054)	(0.183)	(0.212)	(0.213)
Indicator for Novice Teacher	-0.061	0.067	-0.172	-0.114	-0.224
	(0.081)	(0.060)	(0.260)	(0.367)	(0.245)
Count of teachers	175	151	151	151	151
R-squared	0.358	0.248	0.166	0.126	0.131

Notes: Sample consists of all teachers who relevant outcome variable from 2010-11 and all five teacher performance measures in 2011-12. Because state test value-added is collected for all teachers (even those not participating in the study), more teachers are included in the first column than in the subsequent columns. Block bootstrapped standard errors are presented in parentheses, which are generated by drawing 1000 bootstrapped samples, clustered at the school-by-grade level. *** p<.01, ** p<.05, * p<.1

Table 6

Estimates of teacher effects on student achievement, student survey responses, and classroom observation scores

	State Test VA	Project Test VA	Tripod Survey	MQI Observations	CLASS Observations
Expected outcome, based on teacher effectiveness	0.847***	1.486***	-0.228	0.926***	0.877
	(0.228)	(0.267)	(0.778)	(0.284)	(0.543)
Type of estimation	IV	IV	IV	OLS	OLS
Number of observations (students)	888	859	858		
Number of observations (teachers)				61	61
R-squared	0.576	0.566	0.013	0.207	0.075

Notes: In the first three columns, the sample includes all students with the relevant outcome variable and were assigned to and taught by teachers who had predicted teacher effects in 2012-13. In the last two columns, the sample includes the 61 teachers who participated in our random assignment study who had predicted and actual observation scores in 2012-13. The first three columns are based on the IV model described in Equation 4 and Equation 5 using students' state test scores (column 1), project test scores (column 2), and survey responses (column 3) as outcome variables. We include controls for students' prior achievement on state tests and demographics, and fixed effects for random assignment block. The OLS models for MQI and CLASS are described in Equation 6 and control for random assignment block fixed effects. For all five models, block bootstrapped standard errors are presented in parentheses, which are generated by drawing 1000 bootstrapped samples, clustered at the random assignment block level. *** p<.01, ** p<.05, * p<.1

Table 7

Specification checks for estimates of teacher effects on student state mathematics test achievement

	(1)	(2)	(3)	(4)
Expected student achievement, based on teacher effectiveness	0.755*	0.848***	0.847***	0.715**
	(0.387)	(0.222)	(0.228)	(0.297)
Controls for student's prior achievement?	No	Yes	Yes	Yes
Controls for student's demographics?	No	No	Yes	Yes
Controls for actual peer characteristics?	No	No	No	Yes
Type of estimation	IV	IV	IV	IV
Number of students	888	888	888	888
R-squared	0.004	0.563	0.576	0.589

Notes: The sample includes all students with the state test scores in 2012-13 and were assigned to and taught by teachers who had predicted teacher effects in 2012-13. All columns are based on the IV model described in Equation 4 and Equation 5 using students' state test outcomes. In addition to the control variables specified in the table, we include fixed effects for random assignment block in all columns. Block bootstrapped standard errors are presented in parentheses, which are generated by drawing 1000 bootstrapped samples, clustered at the random assignment block level. *** p<.01, ** p<.05, * p<.1

Table 8
Balance of randomly assigned classrooms

	Sample Mean	Coefficient on Assigned Teacher Effectiveness
Baseline math state test score	0.113	-0.016
		(0.229)
Baseline ELA state test score	0.126	-0.278
		(0.288)
Male	50.0%	0.079
		(0.094)
Black	40.3%	0.121
		(0.178)
Hispanic	21.7%	-0.053
		(0.150)
Limited English proficiency	17.9%	-0.088
		(0.116)
FRPL-eligible	67.2%	0.056
		(0.097)
Special education	6.0%	0.045
		(0.066)
p-value on joint null hypothesis test		0.965
Number of students	888	

Notes: The sample includes all students with the state test scores in 2012-13 and were assigned to and taught by teachers who had predicted teacher effects in 2012-13. Block bootstrapped standard errors are presented in parentheses, which are generated by drawing 1000 bootstrapped samples, clustered at the random assignment block level. The p-value on the joint null hypothesis test is generated using a Fisher-style permutation test, where we draw 1000 block bootstrapped (clustered at the random assignment block level) and we re-assign teachers' value-added randomly from the pool of value-added estimates within randomization blocks. We calculate the F-statistic in each 'null' bootstrapped sample, and the p-value presented in this table is the fraction of null F-statistics that are greater than the actual F-statistic. *** p<.01, ** p<.05, * p<.1

Table 9
Peer balance

	Sample Mean	Coefficient on Assigned Teacher Effectiveness
Mean baseline math state test score of actual peers	0.085	-0.083
		(0.186)
Mean baseline ELA state test score of actual peers	0.094	0.111
		(0.169)
Percent of actual peers identifying as male	49.5%	-0.095
		(0.072)
Percent of actual peers identifying as Black	39.0%	0.099
		(0.111)
Percent of actual peers identifying as Hispanic	22.8%	-0.075
		(0.123)
Percent of actual peers with limited English proficiency	19.5%	-0.250**
		(0.122)
Percent of actual peers who are FRPL-eligible	69.0%	-0.074
		(0.085)
Percent of actual peers classified as special education	8.0%	-0.062
		(0.072)
p-value on joint null hypothesis test		0.442
Number of students	888	

Notes: The sample includes all students with the state test scores in 2012-13 and were assigned to and taught by teachers who had predicted teacher effects in 2012-13. Block bootstrapped standard errors are presented in parentheses, which are generated by drawing 1000 bootstrapped samples, clustered at the random assignment block level. The p-value on the joint null hypothesis test is generated using a Fisher-style permutation test, where we draw 1000 block bootstrapped (clustered at the random assignment block level) and we re-assign teachers' value-added randomly from the pool of value-added estimates within randomization blocks. We calculate the F-statistic in each 'null' bootstrapped sample, and the p-value presented in this table is the fraction of null F-statistics that are greater than the actual F-statistic. *** p<.01, ** p<.05, * p<.1

Table 10
Attrition

	Percentage of Sample	Coefficient on Assigned Teacher Effectiveness
Student has state test outcomes	88.4%	0.189*
		(0.097)
Student has project test outcomes	73.7%	-0.027
		(0.146)
Student has Tripod outcomes	73.7%	-0.026
		(0.146)
Number of students	1177	

Notes: Sample consists of fourth- and fifth-grade students from the four districts in our sample who were randomly assigned to a classroom in 2012-13. Block bootstrapped standard errors are presented in parentheses, which are generated by drawing 1000 bootstrapped samples, clustered at the random assignment block level. *** p<.01, ** p<.05, * p<.1

Appendix A: Estimates of Reliability

Table A1

Reliability of project-developed teacher effectiveness measures

	# Items	Cronbach's Alpha
MQI	14	0.78
CLASS	12	0.89
Tripod	26	0.91
Project test	46	0.82 - 0.89

Notes: Cronbach's Alpha is reported across scores on items at the lesson-level for MQI and CLASS, and reported at the student-level for the Tripod and Project test. The range in Cronbach's Alpha represents the range in internal consistencies across different test forms. The reliability estimates for the state tests used in our analysis have a range of reliability estimates from 0.90 to 0.93, based on the states' technical reports in relevant years.

Appendix B: Evidence of Non-Random Sorting Prior to Randomization

In this appendix, we present information on the extent to which students were sorted non-randomly to teachers in the years prior to randomization. If the teachers who participated in the random assignment portion of our study were already randomly assigned classroom rosters, our validation test would not generalize to other settings where the students are systematically sorted to teachers. However, we find that our experimental sample of teachers appears subject to such sorting in the years before randomization.

In the two years prior to the random assignment, the between-teacher standard deviation in average test scores was 0.412 for the sample of randomized teachers, compared to 0.550 and 0.662 for the non-randomized project teachers and for the non-project teachers, respectively.[18] This indicates that there was a considerable amount of sorting of students to teachers based on prior achievement in all three groups of classrooms, but there was somewhat less sorting for teachers who agreed to be randomized in the third year of our study.

Sorting in a single year does not necessarily lead to bias in teacher effectiveness measures. For example, imagine that students are systematically sorted into classrooms by ability, but then randomly assigned to teachers. In this scenario, end-of-year achievement, in expectation, would still be an unbiased measure of a teacher's effect on student achievement since no teacher would be more likely to receive the most- or least-able students. Only when sorting persists across years does failing to control for baseline achievement lead to bias. To estimate the extent of persistent student-teacher sorting in the years prior to the random

[18] Under normality assumptions, sorting students to teachers perfectly on prior test scores would produce a between-teacher standard deviation in teacher-level average baseline test scores that is similar to the student-level standard deviation (i.e., a standard deviation of one). Alternatively, if there were no sorting (i.e., assignments were random), the coefficient would not be zero, but $1/\sqrt{n}$, where n is the number of students per classroom. Assuming 25 students per classroom, this would be approximately 0.2. Therefore, our estimates suggest some amount of sorting on baseline test scores in all three samples in the years before random assignment.

assignment experiment, we calculate the square root of the covariance across the first two years of the study in the average baseline test scores for students in a teacher's classroom (the "signal" standard deviation in the baseline test scores). We also estimate the within-school signal by adjusting the baseline scores for school-by-year fixed effects.[19] If there were no sorting the signal would be zero, but we find that signal standard deviation in baseline sorting for the randomized teachers is 0.346 and the within-school signal standard deviation is 0.306, indicating the presence of persistent within- and between-school sorting of students to teachers across the two pre-randomization years. We observe even higher persistent sorting for the non-randomized project teachers (0.495 overall and 0.408 within school) and for the non-project teachers (0.652 and 0.544 within-school). We also estimate within-random assignment block signal in baseline sorting by adjusting for random-assignment block fixed effects. This estimation yields a standard deviation (0.150) smaller than the within-school or overall standard deviations, but still indicates within non-random sorting of students to teachers in the years before our random assignment study.[20]

[19] We remove the school-by-year fixed effects from the full sample, as opposed to the school-by-year fixed effects within each of the three sub-samples.

[20] Since there are at most three teachers within a random assignment block, we correct the within-block estimates of variance and signal variance in baseline scores to account for the degrees of freedom. Specifically, we first remove the randomization block-by-year fixed effect from each baseline score and then calculate the square root of the variance and the square root of the covariance in these demeaned baseline scores across the first two years in our sample of teachers. To correct for the loss of degrees of freedom, we multiply the square root of the variance by $\sqrt{(n-1)/(n-k)}$, where n is the number of observations across the two years and k is the number of randomization block-by-year fixed effects. We use the same equation to adjust the square root of the covariance, using the number of unique teachers across the two years for n and the number of unique randomization blocks across the two years for k.

Table B1

Summary of student-to-teacher sorting using pre-random assignment data (2010-11 and 2011-12)

	Project Teachers in Randomized Sample	Project Teachers not in Randomized Sample	Non-Project Teachers
S.D. in baseline test scores	0.412	0.550	0.662
Signal S.D. in baseline sorting	0.346	0.495	0.652
Within-school S.D. in baseline scores	0.406	0.470	0.543
Within-school signal S.D. in baseline sorting	0.306	0.408	0.543
Within-school S.D. in baseline scores	0.277		
Within-random assignment block signal S.D. in baseline sorting	0.150		
Number of students	2,433	7,914	50,636

Notes: Sample consists of three groups of fourth- and fifth-grade students from the four districts in our sample. The first group (N=2,433) includes all students in 2010-11 and 2011-12 who were taught by teachers who later participated in our random assignment study (i.e., the 2012-13 school year). The second group (N=7,914) includes all other fourth- and fifth-grade students in 2010-11 and 2011-12 who were taught by teachers who participated in our study in any of the three years, but did not participate in the random assignment part of the study. The third group (N=50,636) includes all other fourth- and fifth-grade students in 2010-11 and 2011-12 who were taught by teachers in these four districts who were never a part of the study. We use the year prior to random assignment (2011-12) to explore differences in the type of students assigned to these teachers in a non-experimental setting. To estimate the extent of persistent student-teacher sorting in the years prior to the random assignment experiment, we calculate the square root of the covariance across the first two years of the study in the average baseline test scores for students in a teacher's classroom (the "signal" standard deviation in the baseline test scores). We also estimate the within-school and within-random assignment block signal by adjusting the baseline scores for school-by-year fixed effects or block-by-year fixed effects. If there were no sorting the signal would be zero.