Abstract Title Page

Not included in page count.

Title:

Validating Components of Teacher Effectiveness: A Random Assignment Study of Value-Added, Observation, and Survey Scores

Authors and Affiliations:

Andrew Bacher-Hicks, Harvard Kennedy School of Government
Mark Chin, Harvard Graduate School of Education
Thomas J. Kane, Harvard Graduate School of Education
Douglas O. Staiger, Dartmouth College

Abstract Body

Background / Context / Purpose:

Policy changes from the past decade have resulted in a growing interest in identifying effective teachers and their characteristics. After the introduction of No Child Left Behind, practitioners and researchers focused on measuring teachers' contributions to student test score growth, or 'value-added' estimates. In response to more recent grants such as the Teacher Incentive Fund, however, many school districts have adopted a more holistic approach to evaluate teachers, using multiple measures of effectiveness including value-added estimates, observation scores, and ratings from student surveys. Despite this emerging consensus around the importance of evaluating teachers using multiple measures, few experiments have investigated the validity of using these measures as predictors of effectiveness.

Numerous non-experimental studies have documented considerable heterogeneity in the distribution of teacher effectiveness estimates (e.g., Gordon, Kane, & Staiger, 2006, Jacob & Lefgren, 2005, Kane, Rockoff, & Staiger, 2008, McCaffrey et. al., 2004, Rivkin, Hanushek, & Kain, 2005, Rockoff, 2004). Ongoing debates in the economics of education literature, however, question the validity of non-experimental estimates because of potential unaccounted for biases, such as the student-teacher sorting bias (e.g., Kane & Staiger, 2008, Rothstein, 2010, Koedel & Betts, 2011).

Though new quasi-experimental techniques have found minimal bias in estimates of teacher effectiveness (Chetty, Friedman, & Rockoff, 2014)[1], true measure validation and assessments of estimate bias are contingent on results from random assignment studies. Few studies, however, have been able to implement a study design where students are randomly assigned to teachers, despite practical and academic interest. To our knowledge, only two studies in extant literature have used a randomized experiment for the purpose of measure validation. The first randomized experiment used data from 156 classrooms in Los Angeles and found that non-experimental value-added estimates were unbiased measures of teacher effectiveness following random assignment (Kane & Staiger, 2008).[2] A second random assignment study was conducted as part of the Measures of Effective Teaching (MET) Project in which students were randomly assigned in over 1,100 classrooms across six school districts (Kane et. al, 2013). In this study, non-experimental measures of value-added, observations, and student surveys were used to form an estimate of effectiveness. Kane and colleagues concluded that the combined measure was an unbiased measure of student test score growth.

Our study is the third study to use data from a randomized experiment to test the validity of measures of teacher effectiveness. Our study is an important contribution to this literature in two ways. First, it is only the second study to combine data from student test score growth, classroom video observations, and student surveys to test the validity of a combined effectiveness measure.[3] As more and more districts use a combination of these three measures, it will be increasingly important to validate them. Second, we improve on one key limitation from the MET report; we maintain a higher compliance rate of 85% for students within our random

[1] This finding assumes that the non-experimental value-added model controls, at least, for prior-year state standardized test scores.

[2] Again, this finding assumes that the non-experimental value-added model controls, at least, for prior-year state standardized test scores.

[3] As mentioned earlier, the MET Project was the first to do this. More information, including several reports, are available at http://www.metproject.org/

assignment sample, improving our ability to generalize our findings to the entire randomized sample.

We collected effectiveness measures across three school years from three broad areas: value-added, classroom observation, and student surveys. In the first two years, we collected non-experimental estimates of these measures and, in the third year, we designed a randomized experiment to test the validity of these estimates. Using these data, we answer two questions:

1. Does a combination of these three distinct non-experimental measures identify teachers who, on average, produce higher student achievement gains following random assignment?
2. Does the magnitude of the gains correspond with what we would have predicted based on their non-experimental estimates of effectiveness?

In line with the previous literature, we find that our non-experimental estimates of effectiveness are unbiased estimates of the causal impact.

Data:

Our analysis sample consists of 66 fourth- and fifth-grade teachers from four large East coast school districts in the 2010-2011 through 2012-2013 school years. From our sample, we collected videos of classroom practice and student surveys. We also collected student demographic and achievement data for all fourth- and fifth-grade students in these districts. We used these data to generate scores for three measures of teacher effectiveness measures: (1) observation scores, using the MQI (Hill et al., 2008) and CLASS (Pianta, LaParo, & Hamre, 2007) observation instruments; (2) aggregate student perception scores, using a 26 item subset of the Tripod survey (Ferguson, 2009), and; (3) value-added scores from state standardized test scores.

(Please insert Table 1 here). In Table 1, we provide a summary of the characteristics of two different groups: (1) the full sample of fourth- and fifth- grade students in these four districts in the year prior to random assignment and (2) the students taught by teachers in our study in the year prior to random assignment.[4] Because the 66 teachers in our experiment consented to participate, our sample is not a *random* sample of teachers in these districts. Nevertheless, we find that the characteristics of the students assigned to the teachers participating in our study and non-participating teachers are fairly balanced, with similar percentages of male, African-American, Asian, and special education students. Teachers in our randomized sample were assigned a lower percentage of Hispanic, English language learners, and subsidized lunch-eligible students. The 66 teachers in our study have similar overall value-added as non-participating teachers, although we find that they are typically assigned students with higher baseline test scores.

Analysis / Findings:

To answer our research questions, we first constructed the best linear combination of non-experimental student test score, survey, and classroom observation data from the first two years of the study (2010-11 and 2011-12) to predict teachers' average contribution to student

[4] We use the year prior to random assignment to explore differences in the type of students assigned to these teachers in a non-experimental setting.

growth on state standardized math tests another year.[5] We used these predicted outcomes as our non-experimental estimates of teachers' contributions to student test score growth in 2012-13. Then, we examined actual student growth in 2012-13 (the third year of the study in which re randomly assigned students to teachers) and compared our non-experimental prediction of growth to actual growth. To the extent that the predicted growth matches the actual observed student growth following random assignment, the non-experimental estimates will contain no bias. Such a finding implies that the potential threat of unmeasured biases in non-experimental measures (e.g., the teacher-student sorting bias) do not produce a biased causal estimate.

Step 1: Predicting teachers' expected contribution to student growth using non-experimental data from the first two years (2010-11 and 2011-12).

First, we create teacher-level average residuals of student test scores, survey responses, and observations. To do so, we use the non-experimental data from 2010-11 and 2011-12 to fit the following OLS regression equation:

$$A_{i,k,t} = A_{i,t-1}\alpha + S_{i,t}\beta + P_{k,t}\delta + \eta + \varepsilon_{i,k,t}, \qquad (1)$$

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 and district fixed effects, η, we include the following control variables: $A_{i,t-1}$, a cubic polynomial of student i's prior achievement; $S_{i,t}$, a vector of indicators for gender, race and ethnicity, subsidized-priced lunch eligibility, English language learner status, and special education status; and $P_{k,t}$, a vector of average characteristics of student i's peers in the same class and school, including average prior-year test scores and averages of $S_{i,t}$. We generate teacher-level average test residuals ($\hat{\tau}_k$) by averaging the residuals across a teacher's students.

To generate teacher-level average survey residuals ($\widehat{Survey}_k$), we estimate a model identical to equation (1) but use the student-survey score in place of the standardized test score as the dependent variable. Finally, since student-level data does not exist for classroom observations, to generate teacher-level residuals from classroom observations ($\widehat{Observe}_k$), we use the average residuals from the following equation:

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

where $M_{k,t}$ is a measure of a teacher's classroom observation score and $P_{k,t}$ is a vector of average characteristics of student i's peers in the same class and school, including average prior-year test scores and averages of $S_{i,t}$.[6]

After generating these teacher-level residuals, we combine them to estimate the best linear combination for predicting a teacher's impact in another year:

$$\hat{\tau}_{k,2009-10} = \beta_0 + \beta_1\hat{\tau}_{k,2010-12} + \beta_2\widehat{Survey}_{k,2010-12} + \beta_3\widehat{Observe}_{k,2010-12} + \eta + \varepsilon_{i,k,t}, \qquad (3)$$

Where the outcome variable is teacher's estimated impact on test score outcomes in the year before our study began (2009-10). $\hat{\tau}_{k,2010-12}$, $\widehat{Survey}_{k,2010-12}$, and $\widehat{Observe}_{k,2010-12}$ are the teacher-level average residuals from equations 1 and 2 above from the first two years of the study (2010-11 and 2011-12). This approach takes into account the measurement error in each

[5] We use data from the first two years in the study (2010-11 and 2011-12) to 'predict' student test growth in 2009-10. We choose 2009-10 as the outcome because we do not want to use 2012-13 (the random assignment year), since data from that year will be used later to the validity of the non-experimental predictions.

[6] We use four different observation measures. Two are from the MQI (Ambitious Instruction and Major Errors) and two are from the CLASS (Classroom Organization and Classroom Climate).

of the predictors (Mihaly et al., 2013). (Please insert Table 2 here). We report the coefficients from this model in Table 2.

Step 2: Comparing expected contribution to student growth (from 2010-11 and 2011-12) to actual growth following random assignment (2012-13).

To estimate the relationship between our non-experimental estimates of contribution to student growth and actual student achievement growth following random assignment, we use an instrumental variables technique to detect the effect of a student's actual teacher indirectly. We employ this technique because, while we can be confident that the effectiveness of a student's randomly assigned teacher is not correlated with measurable or unmeasurable student characteristics, we cannot be sure that the effectiveness of the actual teacher is not. Despite a compliance rate of 85%, we use the instrumental variables technique to account for the small number of non-compliers.

In Table 3, we present the results of both the 'first-stage' and the 'second-stage' or instrumental variable (IV) estimates. In the first-stage model, we estimate the effect of the randomly assigned teacher on the actual teacher's effectiveness. If we observed perfect compliance, this coefficient would be one. However, we find that a one standard deviation unit increase in assigned teacher effectiveness is associated with approximately a 0.95 standard deviation unit increase in actual teacher effectiveness. Next, we estimate the 'reduced-form' or 'intent-to-treat' estimate. This is the effect of the randomly assigned teacher on student achievement. Finally, we compute the IV estimates by dividing the intent-to-treat estimate by the first-stage estimate:

$$\hat{\gamma}_k^{IV} = \frac{ITT\text{: Effect of Randomly Assigned Teacher Rating on Actual Teacher Rating}}{\text{First} - \text{Stage: Effect of Randomly Assigned Teacher Rating on Student Achievement}}$$

(Please insert Table 3 here). We report our IV estimate of teacher effectiveness in the second column of Table 3. We find that the coefficient of the teacher's expected contribution to student growth on actual student achievement (following random assignment) was 1.08 with a standard error of 0.242. Thus, we reject the hypothesis that this coefficient is equal to zero and we fail to reject the hypothesis that it is equal to one. In other words, we find no evidence that the non-experimental predictions are biased.

Conclusions:

Our findings align closely with those of other experimental and quasi-experimental studies that aim to validate measures of teacher effectiveness. Our study is unique due to its high compliance rate and ability to validate a composite measure including teacher scores of value-added, student surveys, and classroom observation. Despite a weaker identification strategy, large-scale quasi-experimental studies, such as Chetty, Friedman, and Rockoff (2014), have the benefit of more precise estimates than those observed in our results. Our findings, however, provide experimental confirmation of their findings.

We plan to continue this work in the next several months by investigating the validity of individual components of our composite measure. For example, we are particularly interested in exploring the validity of non-experimental observation scores for predicting student test scores and classroom observation scores following random assignment.

Appendices

Not included in page count.

Appendix A. References

Chetty, R., Friedman, J. N., Rockoff, J. E. (2014). Measuring the Impacts of Teachers I: Evaluating Bias in Teacher Value-Added Estimates. *American Economic Review* 104(9), 2593-2632

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. *Brookings Institution*.

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.

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

Jacob, B. A., & Lefgren, L. (2005). *Principals as agents: Subjective performance measurement in education* (No. w11463). 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. *Bill & Melinda Gates Foundation*.

Kane, T. J., & Staiger, D. O. (2008). *Estimating teacher impacts on student achievement: An experimental evaluation* (No. w14607). 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.

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

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

Rothstein, J. (2009). Student Sorting and Bias in Value-Added Estimation: Selection on Observables and Unobservables. *Education Finance and Policy,* 4(4), 537-571.

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

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.

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

Rockoff, Jonah E. (2004). The Impact of Individual Teachers on Student Achievement: Evidence from Panel Data. *American Economic Review*, 92(2), 247-252.

Appendix B. Tables and Figures

Not included in page count.

Table 1. *Comparing Randomized Teachers to Other Teachers, Before Randomization (2010-11 and 2011-12)*

	Randomized Sample	**Non-Randomized Sample**
Student Demographics:		
% Male	50%	49%
% African American	34%	35%
% Asian	11%	9%
% Hispanic	20%	28%
% Other Race	4%	4%
% FRPL	56%	62%
% SPED	12%	10%
% LEP	16%	20%
Teacher Value-Added:		
Mean VA	-0.03	0.01
SD in VA	0.27	0.29
Signal SD in VA	0.19	0.20
Classroom Mean Characteristics:		
Mean student baseline scores	0.14	0.06
SD in baseline scores	0.44	0.59
Signal SD in baseline sorting	0.34	0.54
Within-school SD in baseline scores	0.26	0.48
Within-school signal SD in baseline sorting	0.15	0.42
N (Students)	1,215	30,333

Note: Sample consists of two groups of fourth and fifth grade students from the four districts in our sample. The first group (N=1,215) consists of all students in 2011-12 who were taught by teachers who later participated in our random assignment study (in the 2012-13 school year). The second group is the sample of all other fourth- and fifth-grade students in 2011-12 from the four districts participating in our 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 calculate the signal SD in value-added and baseline scores, we used the square root of the covariance at the teacher level between 2010-11 and 2011-12.

Table 2. *Using Teacher Performance Measures from Other Years to 'Predict' Student Achievement Growth*

Adjusted State Test Score	0.502***
Adjusted Survey Score	-0.051
Adjusted MQI Ambitious Instruction Score	0.114
Adjusted MQI Error Score	-0.210
Adjusted CLASS Classroom Climate	0.041
Adjusted CLASS Organized Instruction	0.036
Teacher Has Master's Degree	-0.035
Teacher Experience == 2	-0.042
Teacher Experience == 3	0.154
Teacher Experience == 4	-0.031
Teacher Experience == 5	0.012
Teacher Experience == 6	0.029
Teacher Experience == 7	0.055
Teacher Experience == 8	0.119
Teacher Experience == 9	0.073
Teacher Experience == 11	-0.030
Teacher Experience == 12	-0.007
Teacher Experience == 13	-0.101
Teacher Experience == 14	-0.022
Teacher Experience == 15	0.002
District == 12	0.027
District == 13	0.013
District == 14	-0.003
Missing Masters and/or Experience Information	-0.136*
Missing Observation and/or Survey Scores	0.122~
Constant	0.020
N (Teachers)	1,283
R^2	0.226

*** p < 0.001, ** p < 0.01, * p < 0.05, ~ , < 0.10

Note: Sample consists of all teachers who taught grades 4 or 5 and had average residuals for at least the state test score data in 2009-10 and (2010-11 or 2011-12). The dependent variable is student achievement on the state math test in 2009-10 (the year before our study began). All predictor variables are averages across the 2010-11 and 2011-12 (the first two years of our study). Teachers who were not part of our study will be missing observation and survey scores, but not adjusted state test scores.

Table 3. *First-Stage and Instrumental Variable Estimates of Teacher Effects on Student Achievement*

	First-Stage	**Second-Stage (IV)**
Predicted (Non-Experimental) Value-Added	0.948***	1.077***
Students' Baseline (Prior-Year) Test Score	0.001	0.784***
Constant	-0.000	-0.399**
Fixed Effects for Random-Assignment Block	Yes	Yes
N (Students)	828	828
R^2	0.956	0.628

*** p < 0.001, ** p < 0.01, * p < 0.05, ~ , < 0.10

Note: Sample consists of all randomized students in grades 4 or 5 in the analysis sample. The dependent variable for the first stage model is each student's actual teacher's value-added. This column reports the coefficient from a regression of actual teacher value-added on assigned teacher value-added. The dependent variable for the second-stage model is student achievement on the state math test. This column reports the Instrumental variable estimates of the effect of actual teacher effectiveness using the effectiveness of the randomly assigned teacher as an instrument.