NEPC
NATIONAL EDUCATION
POLICY CENTER

Review of *Learning about Teaching*

Reviewed By

Jesse Rothstein
University of California at Berkeley

January 2011

Summary of Review

The Bill & Melinda Gates Foundation's "Measures of Effective Teaching" (MET) Project seeks to validate the use of a teacher's estimated "value-added"—computed from the year-on-year test score gains of her students—as a measure of teaching effectiveness. Using data from six school districts, the initial report examines correlations between student survey responses and value-added scores computed both from state tests and from higher-order tests of conceptual understanding. The study finds that the measures are related, but only modestly. The report interprets this as support for the use of value-added as the basis for teacher evaluations. This conclusion is unsupported, as the data in fact indicate that a teacher's value-added for the state test is not strongly related to her effectiveness in a broader sense. Most notably, value-added for state assessments is correlated 0.5 or less with that for the alternative assessments, meaning that many teachers whose value-added for one test is low are in fact quite effective when judged by the other. As there is every reason to think that the problems with value-added measures apparent in the MET data would be worse in a high-stakes environment, the MET results are sobering about the value of student achievement data as a significant component of teacher evaluations.

Kevin Welner
Editor

William Mathis
Managing Director

Erik Gunn
Managing Editor

National Education Policy Center

School of Education, University of Colorado
Boulder, CO 80309-0249
Telephone: 303-735-5290
Fax: 303-492-7090

Email: NEPC@colorado.edu
http://nepc.colorado.edu

Publishing Director: Alex Molnar

GREAT LAKES CENTER
FOR EDUCATION RESEARCH & PRACTICE

This is one of a series of Think Twice think tank reviews made possible in part by funding from the Great Lakes Center for Education Research and Practice. It is also available at http://greatlakescenter.org.

Review of *Learning about Teaching*

Jesse Rothstein, University of California at Berkeley

I. Introduction

The Bill & Melinda Gates Foundation's "Measures of Effective Teaching" (MET) Project aims to "improve the quality of information about teaching effectiveness available to education professionals within states and districts"[1] by collecting and analyzing data from six major urban school districts. Two activities form MET's core. First, lessons from over 3,000 teachers were videotaped and are being scored according to five leading practice guidelines, including Charlotte Danielson's "Framework for Teaching."[2] Second, during the 2010-11 school year, students have been randomly assigned to participating teachers. This is intended to permit unbiased estimation of teachers' causal effects on student test scores by ensuring that teacher assignments are uncorrelated with other determinants of student achievement.

The first MET report, *Learning about Teaching: Initial Findings from the Measures of Effective Teaching Project*, was released in December 2010.[3] As neither the videotaped lesson scores nor the random assignment data are yet available, the report provides only preliminary analyses of teachers' "value-added" scores in 2009-10 (before the MET random assignment) and of surveys of student perceptions. The value-added scores are computed both for the regularly administered state tests and for supplemental assessments of "higher-order conceptual understanding" (p. 6). Students are asked to rate their teachers and classrooms on seven dimensions, such as control of the classroom and care shown for students.[4]

The MET study is an unprecedented opportunity to learn about what makes an effective teacher. However, there are troubling indications that the Project's conclusions were predetermined. The Project has two stated premises: "First, a teacher's evaluation should depend to a significant extent on his/her students' achievement gains; second, any additional components of the evaluation (e.g., classroom observations) should be valid predictors of student achievement gains" (pp. 4-5). These premises rule out conclusions that test score gains are too redundant with other available information, too loosely related with good teacher practice, or too poor a measure of the wide spectrum of material that students should learn in school to be useful components of teacher evaluations. Yet these are precisely the sorts of conclusions that the MET data can be used to support or reject.

In fact, the preliminary MET results contain important warning signs about the use of value-added scores for high-stakes teacher evaluations. These warnings, however, are not heeded in the preliminary report, which interprets all of the results as support for the use of value-added models in teacher evaluation. Moreover, the report's key conclusions prejudge the results of the unfinished components of the MET study. This limits the report's value and undermines the MET Project's credibility.

II. Findings and Conclusions of the Report

The report lists four findings (p. 9):

- "In every grade and subject, a teacher's past track record of value-added is among the strongest predictors of their students' achievement gains in other classes and academic years."
- "Teachers with high value-added on state tests tend to promote deeper conceptual understanding as well."
- "Teachers have larger effects on math achievement than on achievement in reading or English Language Arts, at least as measured on state assessments."
- "Student perceptions of a given teacher's strengths and weaknesses are consistent across the different groups of students they teach. Students seem to know effective teaching when they experience it."

III. The Report's Rationale for Its Findings and Conclusions

The results presented in the report do not support the conclusions drawn from them. This is especially troubling because the Gates Foundation has widely circulated a stand-alone policy brief (with the same title as the research report) that omits the full analysis, so even careful readers will be unaware of the weak evidentiary basis for its conclusions.[5]

The report first examines the magnitude of "stable" differences—the component that persists across different classes taught by the same teacher—in teacher value-added and student perceptions. The stable component of value-added on state assessments is found to have a wider distribution for math than for English language arts (ELA). On alternative assessments designed to test higher-order skills, however, the distribution is wider for ELA than for math. (A wider distribution is taken to mean that teacher effectiveness is a more important determinant of student outcomes, though as I discuss below this relies on an unsupported assumption that value-added scores can be interpreted as reflecting teachers' causal effects.)

Student evaluations are more stable than is value-added; teachers tend to draw similar ratings across their different classes. Stability need not indicate consistency in the teacher's behavior, however. Instead, it may simply indicate that evaluations are contaminated by stable but irrelevant external factors. For example, a teacher who teaches only honors classes may elicit more agreement with the statement "My classmates behave the way my teacher wants them to" (part of the "Challenge" domain) than her colleague teaching the regular track, even if she is in fact worse at classroom management, simply because her students are better behaved to begin with. This problem is well known in the area of school accountability, where the most stable measures (e.g., average test scores) consistently favor schools with students from more advantaged backgrounds while measures that better isolate school contributions (e.g., average gain scores) are much less stable.[6] The report does not attempt to adjust the survey responses to isolate a component more plausibly attributable to the teacher.

Next, the report examines correlations among the measures. The correlation between the stable components of a teacher's estimated value-added on the state test and on the alternative assessment is 0.54 for math and 0.37 for ELA. A 0.54 correlation means that a teacher whose stable value-added for the state assessment places her at the 80th percentile—better than four out of five teachers in the MET sample—has about a 30% chance of being below average on the alternative assessment (and vice versa). A correlation of 0.37 means that the 80th percentile teacher is below average on the alternate test more than one-third of the time.

Hence, while the report's conclusion that teachers who perform well on one measure "tend to" do well on the other is technically correct, the tendency is shockingly weak. As discussed below (and in contrast to many media summaries of the MET study), this important result casts substantial doubt on the utility of student test score gains as a measure of teacher effectiveness. Moreover the focus on the stable components—which cannot be observed directly but whose properties are inferred by researchers based on comparisons between classes taught by the same teacher—inflates the correlations among measures. Around 45% of teachers who appear based on the actually-observed scores to be at the 80th percentile on one measure are in fact below average on the other. Although this problem would decrease if information from multiple years (or multiple courses in the same year) were averaged, in realistic settings misclassification rates would remain much higher than the already high rates inferred for the stable components.

While the report's conclusion that teachers who perform well on one measure "tend to" do well on the other is technically correct, the tendency is shockingly weak.

Finally, the report uses a teacher's estimated value-added and student perceptions in one class to predict the teacher's value-added score in another class. The data (presented in Table 9 of the report) indicate that estimated value-added in one class is far more useful than are student perceptions for predicting value-added in another class, and indeed the student surveys add little or nothing to the predictions obtained from value-added scores alone.

Unfortunately, the results of this exercise are presented in an unusual way that makes it difficult to assess the strength of the predictions. Some simple calculations, described in the appendix to this review, suggest that the prediction models are only modestly successful. For example, even in the model for value-added on the state math test—the easiest to predict of the measures considered—a teacher whose predicted value-added is at the 25th percentile (that is, lower than 75% of her colleagues) has only about a one-third chance of actually being that far below average and about the same chance of in fact being *above* average. High-stakes decisions made based on predicted value-added will inevitably penalize a large number of teachers who are above average even when judged solely by the narrow metric of value-added for state tests.

IV. The Report's Use of Research Literature

The report focuses on the new MET Project data and does not review the literature. The estimates of the stable component of teacher value-added resemble those found in earlier studies.[7] Those earlier studies also identify some unresolved concerns about the value-added methodology that, as I discuss below, bear importantly on the MET Project's agenda. Unfortunately, the report does not grapple with these issues, even when the data could be used to shed light on them.

V. Review of the Report's Methods

The report relies primarily on simple, descriptive analyses. (The calculation of the value-added scores themselves is somewhat complex, though the method is well established.) While this is commendable, the methods have a number of shortcomings that are not clearly explained.

First, value-added models can only partially control for differences in the kinds of students that different teachers are assigned.[8] A teacher assigned unusually bright students may receive a high value-added score even though she is quite ineffective. As noted earlier, student perception measures may be sensitive in the same way, if students who tend to give their teachers high marks are disproportionately assigned to particular teachers. The random assignment component of the MET project, not yet complete, is designed to avoid these sorts of biases. Until those results are available, even the modest correlations found here cannot be confidently interpreted as reflecting common components of teachers' causal effects. Moreover, while the available data could have been investigated for signs of potential bias, this was apparently not done.[9]

Second, the report focuses on the components of the value-added and student survey measures that are stable across a teacher's different classrooms. This discards "transitory" components that account for two-thirds of the variation in value-added scores and one-third of the variation in student perceptions.[10] The stable components cannot be measured directly; instead, the report uses statistical analysis to infer their properties. The correlations among the observed measures are much weaker than those inferred for the unobserved stable components. For example, while the stable components of student perceptions and value-added are modestly correlated (around 0.4), the correlation between actually-measured perceptions and value-added scores is closer to 0.2. It is this weak correlation that is the relevant statistic for assessing whether a policy that evaluates teachers based on a classroom-level measure of one of these is adequately identifying the teachers who do well or poorly by the other. That is, a real-world policy will have to use actual measures, not the hypothetical stable constructs.

A third important limitation lies in the specific value-added model (VAM) used. This VAM, which has been used in past research by MET Project director Thomas J. Kane and his coauthors, is not intended to identify differences in effectiveness between the teachers assigned to high- and low-achieving classrooms.[11] By design, its estimates indicate no such differences. It thus cannot be used for many of the purposes for which VAMs are intended; a district concerned about inequity in the distribution of effective teachers could not use this VAM to assess the problem. The correlations reported here might be higher or lower with the VAMs used for teacher assessments in many states and districts, which typically do permit comparisons between high- and low-scoring classrooms.[12]

Finally, the report does not investigate one question central to the MET Project agenda. The project's rationale for considering non-test-based measures of teacher effectiveness is that these measures can be constructive in a way that value-added scores are not, suggesting to teachers why they were judged to be effective or ineffective and what they can do to improve their performance.[13] But the report includes no analyses of teachers' *relative* performance on different student perception domains, and as a result we learn nothing about whether a teacher would be wise to use the student surveys to provide "feedback on specific strengths and weaknesses in [his/her] practice" (p. 31).

VI. Review of the Validity of the Findings and Conclusions

The report's main conclusion, that "a teacher's past track record of value-added is among the strongest predictors of their students' achievement gains in other classes and academic years" (p. 6), is not supported by the underlying analysis. To evaluate this claim, one would have to compare the strength of *several* plausible predictors (including, for example, the classroom practice scores still being assigned). Yet this study examines only the student perception surveys, and few would expect these to be among the strongest measures of teacher effectiveness.

Other conclusions are better supported but of unclear relevance. Yes, teachers with high value-added scores on state tests "tend to" have somewhat above-average scores when value-added is computed from alternative assessments. It is true as well that student perceptions in one class "are related to" student achievement in others. But the report has nothing to say about whether these relationships, even if causal—although the report uses causal language, this is unsupported pending the random assignment study—are large enough to be useful. They are not.

In particular, the correlations between value-added scores on state and alternative assessments are so small that they cast serious doubt on the entire value-added enterprise. The data suggest that more than 20% of teachers in the bottom quarter of the state test math distribution (and more than 30% of those in the bottom quarter for ELA) are in the top half of the alternative assessment distribution. Furthermore, these are "disattenuated" estimates that assume away the impact of measurement error. More than 40% of those whose actually available state exam scores place them in the bottom quarter are in the top half on the alternative assessment.

In other words, teacher evaluations based on observed state test outcomes are only slightly better than coin tosses at identifying teachers whose students perform unusually well or badly on assessments of conceptual understanding. This result, underplayed in the MET report, reinforces a number of serious concerns that have been raised about the use of VAMs for teacher evaluations.

One such concern is that teachers facing accountability pressure will emphasize topics, skills and lessons that raise scores on the tests and de-emphasize those that are not tested. For example, if the exam covers multiplication tables but not fractions or knowledge of history, a teacher accountable for her students' scores will face an incentive to find time for more drilling on multiplication, even at the expense of those other valuable areas.

This concern is evidently well founded. The MET Project teachers whose students excel on state tests get only moderately good outcomes on more conceptually demanding tests in the same subject. A teacher who focuses on important, demanding skills and knowledge that are not tested may be misidentified as ineffective, while a fairly weak teacher who narrows her focus to the state test may be erroneously praised as effective. At a minimum, any test used for teacher evaluations would need to map perfectly to the curriculum that teachers are meant to cover. Existing state tests do not come close to this, and it is not clear that it is realistic. Even supposing higher-order conceptual understanding could be tested economically on a mass scale, the tests would also need to cover citizenship, ethics, creativity, physical fitness, impulse control, and the other "soft" and "noncognitive" skills that are undeniably important parts of what students learn in school.

The MET Project teachers whose students excel on state tests get only moderately good outcomes on more conceptually demanding tests in the same subject.

A second, related concern has to do with the advisability of targeting short-term achievement gains. An extensive literature makes clear that students assigned to high-value-added teachers see higher test scores in the year of that assignment, but that this benefit evaporates very quickly.[14] As just noted, one interpretation that is consistent with the MET Project data is that value-added scores capture in part the degree to which teachers are teaching to the test. Until "fade-out" is better understood, policy makers should be wary of creating even greater incentives than already exist for teachers to aim their efforts at short-term achievement. And the MET Project premise that student test score gains—which will only realistically be available as short-term measures—should be the core of teacher assessments is unsupported and perhaps deeply misguided.

VII. Usefulness of the Report for Guidance of Policy and Practice

The MET Project is assembling an unprecedented database of teacher practice measures that promises to greatly improve our understanding of teacher performance. Even the preliminary analyses in this report expand the boundaries of our knowledge, pointing, for example, to student perceptions as a potentially valuable source of information. Unfortunately, however, the analyses do not support the report's conclusions. Interpreted correctly, they undermine rather than validate value-added-based approaches to teacher evaluation.

The design of the MET study—in particular, its focus on data collected in settings where those data are not being used to guide decisions—places sharp limits on its ability to inform policy. Many of the most pressing concerns about high-stakes, test-based teacher accountability are closely tied to what is known as Campbell's Law: "The more any quantitative social indicator is used for social decision-making, the more subject it will be to corruption pressures and the more apt it will be to distort and corrupt the social processes it is intended to monitor."[15] The properties of the MET measures would likely be different if they were being used for high-stakes decisions, as teachers, students, and administrators would target the measures in order to achieve desired outcomes.

The MET study cannot reveal the extent or severity of the distortions that this would produce, though the preliminary data do suggest that they would be a serious problem. The specific design of student assessments is evidently an important determinant of which teachers are identified as effective, even in an environment where teachers are not held individually accountable for their student's scores. If the stakes are raised, teachers will have strong incentives to redirect their efforts toward activities that raise their measured value-added, further widening the already large wedge between the measurement and the reality.

Campbell's Law applies to other indicators as well, certainly including student surveys. Mischievous adolescents given the opportunity to influence their teachers' compensation and careers via their survey responses may not answer honestly. Teachers facing high stakes may feel compelled to alter their practice to cater to student demands, whether or not the changes are compatible with more effective instruction. Studies of zero-stakes student surveys can tell us little about how the students would respond if their teachers' careers were on the line.

A successful teacher evaluation policy must include a balanced set of measures that are relatively unsusceptible to manipulation and gaming. We have a great deal yet to learn and the complete MET study will contribute to our knowledge base. But the MET study's inability to examine how teachers, students, and administrators respond to the use of the MET measures for high-stakes decisions limits what the MET Project can tell us about the utility of the measures for real-world use.

Future research should evaluate alternative teacher evaluation *policies* rather than measures.[16] For example, does the use of high-stakes teacher evaluation lead to better student outcomes? Do students do better when their teachers are evaluated based on classroom observations or based on their value-added? Pilot programs that use value-added scores for teacher ratings and compensation have not, in general, shown the hoped-for impacts.[17] Perhaps policies based on the MET Project measures will be more successful. Until that is known, we cannot judge whether the MET Project's premises are wise or misguided. The Project's initial report is thus not very useful in guiding policy, but the guidance it does provide points in the opposite direction from that indicated by its poorly-supported conclusions.

Appendix

This appendix describes the calculations used to assess the strength of the models used to predict teacher value-added in one class given value-added and student assessments in another class.

Table 9 of the report shows the average realized value-added of teachers in the bottom quartile and top quartile of predicted value-added. For example, when math-score value-added and student perceptions are used to predict value-added on the state math test in another section, the quarter of teachers with the "least evidence of effectiveness" had average value-added of -0.074 in that other section, while the quarter with the "most evidence of effectiveness" had average value-added of 0.134. The difference between these is 0.208, as reported in column 3 of the table.

My calculations assume that both student perceptions and estimated value-added are normally distributed. This is likely not exactly right but good enough to support a reasonable approximation. If so, the range between the top-quartile and bottom-quartile mean can be used to infer the standard deviation of predicted value-added. A normal variable with a standard deviation of one has a top-quartile to bottom-quartile range of 2.54, so a prediction with a range of 0.208 must have a standard deviation of 0.208 / 2.54 = 0.082.

Let the prediction be X and let the stable component of the teacher's true value-added be Y. The conventional goodness-of-fit statistic for prediction models is the R-squared, which can be computed as $R^2 = (\mathrm{StdDev}(X) / \mathrm{StdDev}(Y))^2$ and is interpretable as the share of the variation in stable value-added that can be predicted. Table 5 lists the standard deviations of the stable components of teacher value-added; for the state math test, this is 0.143. So the R-squared of the prediction model is $(0.082/0.143)^2 = 0.327$. This is quite modest; as a point of comparison, a prediction of a college student's freshman GPA given her high school GPA has an R-squared around 0.4.[18]

What does an R-squared of 0.327 mean? Consider the set of all teachers with predicted value-added at the 25th percentile—that is, with a predicted value-added score (relative to the mean) of -0.055. As actual value-added has a standard deviation of 0.143, this prediction would place the teacher at the 35th percentile. In order for a teacher predicted to be at the 25th percentile to actually fall in the bottom quartile, she has to do worse than predicted. The actual 25th percentile value-added (relative to the mean) is -0.096, so the unpredicted component of her value-added has to be -0.096—-0.055 = -0.041 or less. Given a standard deviation of this component of 0.117 (calculated as the square root of the difference between the variances of Y and X), this will be true for approximately 36% of such teachers. A similar calculation indicates that 32% of teachers predicted to be at the 25th percentile will in fact fall in the top half of the stable value-added distribution.

Value-added on the state math test is somewhat more predictable than the other three outcomes considered in Table 9; for the others, R-squared ranges from 0.167 to 0.310 and the share of teachers predicted to be at the 25th percentile who actually fall in the bottom quarter ranges from 33% to just under 36%, while the share who actually fall in the top half ranges from 33% to

38%. Value-added for state ELA scores is the least predictable of the measures, with an R-squared of 0.167. A teacher with predicted value-added for state ELA scores at the 25th percentile is actually more likely to fall in the top half of the distribution than in the bottom quarter!

Notes and References

[1] *Working with Teachers to Develop Fair and Reliable Measures of Effective Teaching*. MET Project White Paper. Seattle, Washington: Bill & Melinda Gates Foundation, 1. Retrieved December 16, 2010, from
http://www.metproject.org/downloads/met-framing-paper.pdf.

[2] The five observation protocols are Danielson's Framework for Teaching; the Classroom Assessment Scoring System developed by Bob Pianta and Bridget Hamre; Mathematical Quality of Instruction, developed by Heather Hill and Deborah Loewenberg Ball; Protocol for Language Arts Teaching Observations, developed by Pam Grossman; and the Quality Science Teaching Instrument, developed by Raymond Pecheone. See p. 6-7 of the report.

[3] *Learning about Teaching: Initial Findings from the Measures of Effective Teaching Project*. MET Project Research Paper. Seattle, Washington: Bill & Melinda Gates Foundation. Retrieved December 16, 2010, from
http://www.metproject.org/downloads/Preliminary_Findings-Research_Paper.pdf.

[4] The student perception surveys were developed by the Tripod Project for School Improvement, founded by Ronald F. Ferguson. Thirty-six questions are grouped into seven constructs, labeled Care, Control, Clarify, Challenge, Captivate, Confer, and Consolidate. See p. 11-12 of the report.

[5] *Learning about Teaching: Initial Findings from the Measures of Effective Teaching Project*. MET Project Policy Brief. Seattle, Washington: Bill & Melinda Gates Foundation. Retrieved December 16, 2010, from
http://www.metproject.org/downloads/Preliminary_Finding-Policy_Brief.pdf.

[6] See:

Kane, T.J. & Staiger, D.O. (2002, Fall). The promise and pitfalls of imprecise school accountability Measures. *The Journal of Economic Perspectives 16*(4), 91-114.

[7] See, e.g.:

Kane, T.J. & Staiger, D.O. (2008). *Estimating Teacher Impacts on Student Achievement: An Experimental Evaluation*. Working paper #14607, National Bureau of Economic Research. Retrieved December 16, 2010, from
http://www.nber.org/papers/w14607.

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

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

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

[9] For three methods that have been used, any of which could be applied in the MET data, see:

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

Hanushek, E.A. & Rivkin, S.G. (2010). *Constrained Job Matching: Does Teacher Job Search Harm Disadvantaged Urban Schools?*. Working paper #15816, National Bureau of Economic Research. Retrieved December 21, 2010, from
http://www.nber.org/papers/w15816.

Clotfelter, C.T., Ladd, H.F. & Vigdor, J.L. (2006, Fall). Teacher-student matching and the assessment of teacher effectiveness. *Journal of Human Resources 41*(4), 778-820.

[10] These are the approximate ratios of the variance of the transitory component to the total variance of each measure.

[11] The VAM used resembles a random effects estimator, and assumes that teachers' effects are uncorrelated with any of the control variables included in the model. These variables include the student's previous test score, race, gender, free or reduced price lunch status, English language learner (ELL) status, and participation in gifted and talented programs. Importantly, the VAM also controls for the classroom-level average of each of these variables. This ensures that the average estimated effectiveness of teachers of, for example, ELL classes is no different than that of teachers of gifted and talented classes, and similarly that estimated effectiveness is uncorrelated with student- or classroom-level average prior achievement.

[12] For example, the models distributed by the University of Wisconsin Value-Added Research Center or the SAS Institute's Education Value Added Assessment System. See, respectively:

Meyer, R.H. (2010). "Value-added systems, accountability, and performance management in education: promises and pitfalls." Presentation at the 2010 Annual Conference of the Iowa Educational Research and Evaluation Association, Cedar Falls, IA, December 2. PowerPoint slides retrieved December 21, 2010 from
http://varc.wceruw.org/PowerPoints/Iowa_Research_2010_Meyer_slides.ppt.

Ballou, D., Sanders, W. & Wright, P. (2004, Spring). Controlling for student background in value-added assessment of teachers. *Journal of Educational and Behavioral Statistics 29*(1), 37-65.

[13] *Working with Teachers to Develop Fair and Reliable Measures of Effective Teaching*. Seattle, Washington: Bill & Melinda Gates Foundation, 5. Retrieved December 16, 2010, from http://www.metproject.org/downloads/met-framing-paper.pdf.

The current report describes the student assessments as a designed to provide specific feedback for teachers: "[S]tudents are asked to give feedback on specific aspects of a teacher's practice, so that teachers can improve their use of class time, the quality of the comments they give on homework, their pedagogical practices, or their relationships with their students" (p. 7).

[14] See, e.g.,

Kane, T.J. & Staiger, D.O. (2008). *Estimating Teacher Impacts on Student Achievement: An Experimental Evaluation*. Working paper #14607, National Bureau of Economic Research. Retrieved December 16, 2010, from
http://www.nber.org/papers/w14607.

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

Jacob, B.A., Lefgren, L. & Sims, D.P. (2010, fall). The persistence of teacher-induced learning. *Journal of Human Resources 45*(4), 915-943.

[15] Campbell, D.T. (1979). Assessing the impact of planned social change. *Evaluation and Program Planning 2*(1), 67-90.

[16] This point was made effectively by:

Rubin, D.B., Stuart, E.A. & Zanutto, E.L. (2004, Spring). *Journal of Educational and Behavioral Statistics 29*(1), 103-116.

[17] See, e.g.:

Springer, M.G., Ballou, D., Hamilton, L., Le, V., Lockwood, J.R., McCaffrey, D., Pepper, M., & Stecher, B. (2010). *Teacher Pay for Performance: Experimental Evidence from the Project on Incentives in Teaching*. Nashville, TN: National Center on Performance Incentives at Vanderbilt University. Retrieved December 16, 2010 from
http://www.performanceincentives.org/data/files/pages/POINT%20REPORT_9.21.10.pdf.

McCaffrey, D.F. & Hamilton, L.S. (2007). *Value-Added Assessment in Practice: Lessons from the Pennsylvania Value-Added Assessment System Pilot Project*. Santa Monica, CA: RAND Corporation.

[18] Rothstein, J.M. (2004, July-August). College performance predictions and the SAT. *Journal of Econometrics 121*(1-2), 297-317.

DOCUMENT REVIEWED:	**Learning About Teaching: Initial Findings from the Measures of Effective Teaching Project**
AUTHORS:	Thomas J. Kane and Steven Cantrell
PUBLISHER/THINK TANK:	Bill & Melinda Gates Foundation
DOCUMENT RELEASE DATE:	December 10, 2010
REVIEW DATE:	January 13, 2011
REVIEWER:	Jesse Rothstein, University of California at Berkeley
E-MAIL ADDRESS:	rothstein@berkeley.edu
PHONE NUMBER:	(510) 643-8561

SUGGESTED CITATION:

Rothstein, J. (2011). *Review of "Learning About Teaching: Initial Findings from the Measures of Effective Teaching Project."* Boulder, CO: National Education Policy Center. Retrieved [date] from http://nepc.colorado.edu/thinktank/review-learning-about-teaching.