NEPC
NATIONAL EDUCATION POLICY CENTER

Review of Two Culminating Reports from the MET Project

Reviewed By

Jesse Rothstein, University of California-Berkeley

William J. Mathis, University of Colorado Boulder

January 2013

Summary of Review

The Gates Foundation's Measures of Effective Teaching (MET) project was a multi-year study of thousands of teachers in six school districts that concluded in January 2013. This review addresses two of the final MET research papers. One paper uses random assignment to test for bias in teachers' value-added scores. The experimental protocol was compromised, however, when many students did not remain with the teachers to whom researchers had assigned them; other students and teachers did not participate at all. This prevents conclusive answers to the questions of interest. The second paper examines how best to combine value-added scores, classroom observations, and student surveys in teacher evaluations. The data do not support the MET project's premise that all three primarily reflect a single general teaching factor, nor do the data support the project's conclusion that the three should be given roughly equal weight. Rather, each measure captures a distinct component of teaching. Evaluating teachers requires judgments about which components are the most important, judgments that are not much informed by the MET's masses of data. While the MET project has brought unprecedented vigor to teacher evaluation research, its results do not settle disagreements about what makes an effective teacher and offer little guidance about how to design real-world teacher evaluation systems.

Kevin Welner
Project Director

William Mathis
Managing Director

Erik Gunn
Managing Editor

National Education Policy Center

School of Education, University of Colorado
Boulder, CO 80309-0249
Telephone: (802) 383-0058

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

Publishing Director: Alex Molnar

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 *Have We Identified Effective Teachers?* and *A Composite Estimator of Effective Teaching*: Culminating findings from the Measures of Effective Teaching Project

Jesse Rothstein, University of California-Berkeley
William J. Mathis, University of Colorado Boulder

I. Introduction

Teacher evaluation has emerged as a prominent educational policy issue. Debate over teacher compensation, hiring and firing, which once centered on traditional salary matrices and teacher observation systems, is increasingly focused on concrete outcome measures—particularly student test score gains.

The Bill and Melinda Gates Foundation has been prominent in these debates. Among other roles, it sponsored a high-profile set of studies known as the Measures of Effective Teaching (MET) project. After initial reports in 2010 and 2012,[1] the foundation released the final results from this multi-year, multi-million-dollar study in January 2013. Results were reported in a summary brief and a set of principles, intended for general audiences, and in three research papers.[2]

This review focuses on two of the three research papers. These two were selected because they form the foundation for the study's policy conclusions.

One describes an experiment that used random assignment to assess the validity of value-added estimates (computed from student test scores) as measures of teacher impacts on student achievement. Value-added measures have been highly controversial, in part because they may be biased by patterns of student assignments to teachers. The experiment aimed to assess the importance of this bias.[3]

The second research paper examines the relationships among three different performance measures: test scores, classroom observations, and student opinion surveys. It investigates how these measures can be combined into a single composite measure of teaching effectiveness.

The third research paper—which is not reviewed here—examines the well-known problem of reliability in teacher observations. This paper provides solid practical observations about the design of an observation system, but it does not investigate how these observations should fit into a broad system of teacher evaluations.

II. Findings and Conclusions of the Reports

Research Paper: *Have We Identified Effective Teachers?* (Random Assignment)

A central focus of the MET study was the use of value-added (VA) scores to measure teacher effectiveness. These scores are based on the end-of-year test scores of a teacher's students. A teacher whose students' scores exceed expectations, which depend on the students' past achievement and other characteristics, gets a high VA score. A teacher whose students achieve below expectations receives a low score.

One persistent concern about VA-based evaluations is that they may unfairly reward or penalize teachers based on the students they teach rather than on their true impacts. Although VA models attempt to adjust for differences in student assignments by comparing student scores with predictions based on achievement in the prior year (and sometimes on other factors as well, such as student race and poverty), they cannot do this perfectly. There may be teachers, for example, who are consistently assigned difficult-to-teach students who will predictably perform below their statistical predictions. Think of a Spanish-speaking teacher who gets all of the limited-English-proficient students in a grade—students who tend to underperform on English-language assessments. This teacher's VA score will be systematically lower than her true effectiveness. Other teachers may get disproportionate numbers of students who can be expected to outperform their predictions; these teachers will receive inflated VA scores.

The degree of bias, if any, in VA scores is of central importance for proposals to use these scores for teacher evaluations. If they are biased, it would be unfair to base a teacher's evaluation on them, and doing so would create perverse incentives: Even very good teachers would rightly hesitate to take on assignments that lead to poor VA scores. Bias in scores assigned to teachers with unusual assignments is thus of particular concern.

The centerpiece of the MET study was an enormous random-assignment experiment aimed at assessing the magnitude of the biases. The research design built on an earlier study by two of the MET investigators, Thomas J. Kane and Douglas O. Staiger.[4] Randomization was used to eliminate systematic patterns in student assignments that might bias VA scores.

MET project staff worked with principals at participating schools to identify teachers who would be eligible for random assignment. Eligibility hinged on there being two or more MET teachers scheduled to teach the same grade and subject, and on the principal viewing each of the teachers as suitable for any of the others' students. Schools assigned rosters of

students to the teachers however they saw fit. Researchers then randomly shuffled the rosters among teachers. In principle, this should have ensured that teachers who ordinarily get unusual student assignments are no more and no less likely to have been assigned such students than are other teachers teaching the same courses at the same schools. Estimates of teachers' impact on the randomized students were then compared with pre-experimental VA scores to assess the magnitude of biases in the latter.

The study's headline result was that pre-experimental VA scores are good (i.e., unbiased) predictors of the teachers' impacts in the experiment, at least in mathematics. This is what

> *Even the best predictors of outcomes in the second year do not predict very accurately, simply because the measures vary substantially for the same teacher across years.*

one would expect if the pre-experimental scores are largely free of bias. (Estimates were consistent with substantial bias in English language arts VA scores, though sampling errors were large and thus the results were inconclusive.) Moreover, teachers with higher pre-experimental mathematics VA scores also tended to produce larger effects on alternative, more conceptually demanding tests, though these effects were notably smaller.

The report offers two important cautions: First, the researchers point out that students were not randomly assigned to schools, so the study cannot assess the possibility of bias in comparisons of teachers across different schools. Second, the researchers note that VA measures "are prone to substantial error" (p.38)—year-to-year correlations are generally around 0.5 for elementary teachers. Moreover, if the test scores were used for high-stakes purposes "the amount of error could also worsen" (p. 39) as teachers and others act—by, for example, teaching to the test or neglecting non-tested subjects—to improve their scores.

Research Paper: *Combining Measures to Identify Effective Teaching*

This analysis compared alternative performance measures that might be used in teacher evaluations. The study considered four categories of measures: traditional value-added measures; value-added scores computed based on student achievement on alternative, more conceptually demanding tests; classroom observations by trained observers following standardized protocols; and students' self-reported perceptions of their teachers.

A preliminary MET study showed that teachers' value-added scores computed from the regular state tests were positively but weakly correlated with value-added scores computed from the more conceptually demanding tests.[5] A second study found that classroom observations and student survey scores were positively but not strongly correlated with the value-added scores.[6] But a comprehensive analysis of the different measures was held back until the final release.[7] Unlike the other final reports, this study is written in technical

language that is not easily intelligible to a layperson, and it offers little in the way of a simple summary of its results.

The study concludes after analyzing various statistical models "that there is a common component of effective teaching shared by all indicators, but there are also substantial differences in the stable components across measurement modes and across some indicators within a mode" (p. 2). In English, the study found that the different measures capture different dimensions of teaching, with partial but far-from-complete overlap among them. The results indicate that an equally weighted average of value-added scores, classroom observations, and student surveys does nearly as well as any other measure of capturing the common component and is more reliable than many alternatives. .Thus, the summary report for policymakers and practitioners concludes that the equally weighted composite "demonstrated the best mix of low volatility from year to year and ability to predict student gains on multiple assessments."

III. The Reports' Rationales for Their Findings and Conclusions

The Gates Foundation has long held that teachers should be recognized and rewarded (or sanctioned) for their effectiveness. This, of course, requires that effectiveness be measured, and the MET project aimed to better understand how to do this. It began with two highly controversial premises: "First, that 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."[8]

These assertions are highly contestable. It is quite possible that test-score gains are misleading about teachers' effectiveness on other dimensions that may be equally or more important. If so, it would be unwise to put much weight on student test scores in teacher evaluations, and it would be equally unwise to discount alternative measures that may better capture those other dimensions simply because they do not correlate highly with test-score gains.

While these premises treat VA as the standard against which all else must be evaluated, a second idea motivating the MET study is that "multiple measures" of teaching effectiveness are needed. The project considered three broad classes of measures: test score gains,[9] student surveys and classroom observations.

The MET premises lead to a particular conception of the roles played by these multiple measures. The study presumes a single "general" teaching quality factor and aims to use the three classes of measures to uncover this factor. Its "composite" measures of teaching effectiveness are simply weighted averages of the different measures, with weights chosen to best predict one of the measures in the following year. (That is, one composite weights the measure to best predict the next year's VA; another to best predict the next year's classroom observation. The report also considers composites that place pre-specified

weights on the measures. For example, one weights the three measures equally.) Insofar as the measures all reflect the general teaching quality factor, all of the composites should be interpretable as measuring the general factor, with only small differences in reliability among them.

As discussed below, however, the results are not consistent with this. Evidently, the three MET premises—that there is a single dimension of underlying teacher effectiveness; that student achievement gains are a valid measure of this; and that multiple measures should be used to best identify effective teachers—are not all valid.

IV. The Reports' Use of Research Literature

The authors are prominent researchers in the area and demonstrate a good command of the statistical literature. The reports do not consider the substantial qualitative literature that considers teaching as a multidimensional enterprise that serves a variety of purposes beyond test-score improvement.[10]

V. Review of the Reports' Methods

Random Assignment Analysis

The MET study was a monumental undertaking. The random assignment study encompasses tens of thousands of students and over a thousand teachers, spread across hundreds of schools. There has never before been an experiment of this magnitude in education. The findings are important and should be carefully considered. They are generally supportive of the interpretation of VA scores as measures of teachers' causal impacts on the student test scores used to compute the VA model. However, three methodological concerns caution against over-interpretation of the results, and in particular against generalizing the findings to the broader population of teachers:

1. The sample used for the randomization study was not representative of the schools included, nor were these schools representative of non-MET schools and districts. Particularly notable is that the students included in the randomization—that is, those assigned by the school to one of the included teachers—scored substantially higher (by about 0.15 standard deviations) on the baseline math and ELA tests than did other students at the same schools who were not included. Relatedly, the teachers excluded from the randomization had, in the year prior to the experiment, higher shares of special education students and English language learner students than did those included, and the MET district schools that did not participate in the randomization had much higher shares of these students than those that did. Finally, it appears that teachers whose prior year classrooms were especially high- or low-

achieving were generally not included. As noted earlier, these exclusions are particularly important because they go to the heart of prior concerns about bias in VA approaches.

2. There was a great deal of noncompliance with the experiment. Among the MET districts, the fraction of students who actually were taught by the teacher to whom they were randomly assigned ranged from about one-quarter to about two-thirds. Most of the other one-third to three-quarters of the students remained in the original school but were shifted to other teachers. In some cases this was because teachers were shifted to other grades or course sections after rosters were submitted to the researchers; in others, the school simply did not implement the intended assignments. Once again, the concerns about potential bias in VA scores center on the students who are unlikely to comply with random assignments because it matters which teacher they have.

The report used standard statistical techniques for analyzing experiments with noncompliance. These methods are designed to use the students who did comply with their experimental assignments to isolate teachers' effects. The resulting estimates do not indicate any bias in pre-experimental estimates of these effects, at least in math. However, the high rates of noncompliance make the estimates quite imprecise. This prevents any conclusions at all about ELA effects, for which the estimates indicate large biases—a red flag—but for which these estimates are so imprecise that the hypothesis that the apparent bias is due to chance cannot be ruled out.

Together, these results combine to reduce the usefulness of the study. The estimates are informative only about the teachers and students who actually complied with their assignments; the experiment cannot measure bias in the VA scores of teachers who wound up teaching students other than those to whom they were initially assigned. It appears that there were a great many of these.

Importantly, the teachers for whom bias in VA is of greatest concern—those who specialize in teaching distinctive types of students who might be expected to over- or under-perform their statistical predictions on the end-of-year tests—appear to be underrepresented in the random assignment sample, and may well be those whose students were least compliant with the experimental assignment.

Combining Alternative Measures

The second report takes up a very different question: Given a number of evaluation measures in one year, how can one best predict a teacher's performance in a subsequent year? Importantly, there is no consensus about how to measure performance in that second year. Some agree with the MET project's initial premise that student test-score gains are the overarching standard and that other measures are valuable only to the extent that they predict these, while others believe that classroom observations are a better

standard to compare against.[11] The final report explores predictions of each of the different performance measures. Results indicate:

1. The best predictor of value-added for the state test in the second year is the teacher's value-added on the state test in the first year. If one's goal is solely to predict value-added, one will put very little weight on either classroom observations or student surveys.

2. This result is symmetric: The best predictor of classroom observations in the second year is the observation score in the first year. Student surveys convey a bit of additional information, while value-added adds much less. And the best predictor of student surveys in the second year is the student survey score from the first year; neither value-added nor classroom observations provide meaningful amounts of additional predictive power (especially in middle school).

3. Different views of what best measures effectiveness yield very different assessments of which teachers are effective. Evaluations that take value-added as the baseline are correlated only about 0.4 with evaluations that treat classroom observation scores as the best measure of effectiveness.

4. Even the best predictors of outcomes in the second year do not predict very accurately, simply because the measures vary substantially for the same teacher across years. In elementary school, none of the composite measures of effectiveness in mathematics instruction correlates higher than 0.57 from one year to the next. For ELA the composites are even less stable—0.5 or less, with many around 0.4. Stability is higher in middle school, in part because the measures can be averaged across multiple classrooms taught by the same teacher in the base year. At each level, composites that put roughly equal weight on the various measures are generally a bit more stable than those that put most of the weight on individual measures (as do the best predictor composites).

5. None of the three measures does a very good job of predicting value-added for alternative, more conceptually demanding tests, which appear to capture an additional dimension of teacher effectiveness that is not measured by the regular state tests. The best predictor of the teacher's alternative test VA puts most of the weight on the state test VA score, but classroom observations contain important additional information. Moreover, none of the composites correlates higher than about 0.25 with the alternative test value-added in the following year.[12] (By comparison, state test value added and classroom observations are more predictable, with correlations of around 0.4 with the best predictors from the previous year.) There is evidently a dimension of effectiveness that affects the conceptually demanding tests that is not well captured by any of the measures examined by the MET project.

VI. Review of the Validity of the Findings and Conclusions

As noted above, the MET random assignment study built on an earlier random assignment study by Kane and Staiger. That earlier study had two main limitations: It used a highly non-representative sample of teachers that excluded exactly the teachers most likely to be affected by bias in VA measures—those who usually get unusual teaching assignments—and its sample was too small to yield precise estimates.

The MET study promised to overcome these limitations by using a more representative, larger sample. Unfortunately, the high rates of noncompliance reduced the effective sample size in the MET study, leaving a fair amount of imprecision in the estimates. They also reduce confidence in the representativeness of the results. The evidence presented

> *The results of the MET analysis, at least as contained in the two reports reviewed here, say little about how best to conduct teacher evaluations in the real world.*

indicates that teachers with unusual assignments—those who specialize in LEP, special education, disruptive, or high- or low-achieving students—were unlikely to have been included in the randomization sample. And even if they were included, it is not clear whether they actually taught the students to whom they were randomly assigned.

Indeed, it seems reasonable to expect that the teachers with unusual assignments, and the students usually assigned to them, were exactly those who were least likely to comply with the experimental assignment. For example, if Ms. Smith has skills with low-achieving students, while Mr. Jones has skills with high-achievers, one would expect that some low-achieving students would move from Jones to Smith, while high-achievers would move from Smith to Jones, no matter what their experimental assignments say. For all these reasons, the MET results do not tell us much about the magnitude of biases in VA evaluations of a broader population of teachers.[13]

Turning to the second report, concerning the combining of multiple measures, the mass of correlations presented provides a great deal of food for thought. But it does little to resolve the policy question of how teachers should best be evaluated. Indeed, there is little reason for anyone to rethink their approach to the question on the basis of the results.

A reader who comes to this study thinking that value-added is a better measure of a teacher's effectiveness than a high-quality classroom observation will likely conclude that teachers should be evaluated primarily by their value-added, and that there is little use for classroom observations to justify their enormous expense. On the other hand, a reader who comes to the study thinking that a good classroom observer can identify a good teacher and that value-added scores have a questionable relationship to effective teaching will find nothing in this study to dispel that view—rather, the results indicate that VA scores add little value to an observation-based evaluation system.

One strong conclusion that can be drawn is that the choice of criteria matters. There are clearly multiple dimensions of teachers' effects, and many teachers do well on some and poorly on others. There is no statistical procedure that can decide which dimensions should count more in teacher evaluations; that decision must be left to subjective judgments.

Moreover, some of the dimensions of effectiveness—those measured by student performance on alternative, open-response (i.e., not multiple-choice) tests designed to capture higher-order thinking—are not at all well measured by the standardized measures that were the focus of the MET research. An evaluation system based around the MET measures will fail to identify teachers who are effective or ineffective on those other dimensions. Just as importantly, such as system will likely discourage teachers from putting their efforts toward the kinds of learning that the alternative tests measure—in this case, higher-order, conceptual thinking.

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

The MET study has generated an enormous amount of valuable data, and researchers will be reanalyzing and debating the results for many years to come. We applaud the Gates Foundation's commitment to making the MET data available in the near future for outside researchers to study. But the results of the MET analysis, at least as contained in the two reports reviewed here, say little about how best to conduct teacher evaluations in the real world.

As we have discussed, the random assignment study suggests that VA-based measures are unbiased for typical teachers teaching typical students, at least in mathematics. But there are two important caveats to this result. First, it holds only for the particular VA model examined by the MET researchers, which is richer in important ways than some of those being used in practice. It is not clear whether VA measures based on alternative models like those used in many districts are similarly unbiased. Second, it holds only for the types of students and teachers who participated in the random-assignment experiment and complied with their assignments. Because this group was so non-representative, one cannot generalize to the less typical teachers and students for whom bias in VA measures is most likely. The bias concerns are thus left unresolved.

The study of composites constructed from multiple measures of teacher performance is even less encouraging. It generally undermines the movement toward more rigorous teacher evaluation, which has often proceeded on the assumption that any distinctions are better than none. That is, the thinking has been, "Everybody knows who are the good and bad teachers, so any valid measure will yield similar rankings." Yet the MET study makes clear that this is not right—that different ways of evaluating teachers will yield very different rankings, that none are clearly better than the others, and that there are potentially important dimensions of effectiveness that are not captured by any of the available evaluation measures. Distinctions made based on any particular measure will

generally—with a great deal of measurement error—identify teachers who excel on that measure, but these teachers will often be unexceptional along other dimensions.

It remains to educational policymakers to decide for themselves which of the various dimensions are most important for teaching effectiveness. Importantly, this subjective decision must precede the design of the evaluation system, as different decisions have very different implications about what measures should be included.

Only after this decision is made will policymakers confront the question of how to minimize the effects of volatility and inaccuracy in evaluations. The best way to do this will depend on the way that the evaluations are being used. Some uses will call for maximizing accuracy without regard to volatility, while others will require a more stable, if less accurate, measure. The MET project's conclusion that an equally weighted composite does the best job of trading off inaccuracy in predictions against year-to-year volatility is unsupportable absent a specific analysis of the use to which the evaluations will be put.

Finally, two of the most important questions for the design of better teacher quality policies are not at all addressed by the MET study:

1. ***What are the effects of attaching high stakes, either in the form of pay-for-performance or through retention and non-retention decisions, to the various measures of effective teaching?***

It is a well-established principle of social science—known as "Campbell's Law," after noted education researcher Donald Campbell—that a measure that performs well in a low-stakes setting will inevitably be distorted when the stakes are raised. To their credit, the researchers acknowledge this concern. Thus, many analyses of the design of performance evaluation systems recommend that the stakes be kept relatively low, to reduce the incentive to distort the performance measures.

There are a number of ways for a teacher to influence her VA score other than through actually becoming more effective: She can arrange to be assigned a different set of students; she can reduce the amount of class time that she devotes to non-tested subjects (e.g., history) in order to allocate more time to the subjects covered by the tests; she can teach test-taking skills in place of substantive knowledge; and she can try to rally students to give their best effort on a test which is usually, from the student's perspective, entirely meaningless. Student surveys are clearly at least as susceptible to this kind of influence, and classroom observations may be as well.

Because the MET study took place exclusively in a low-stakes setting (where teachers faced moderate- or high-stakes evaluations, those evaluations did not depend on the MET measures themselves), it cannot inform us about the sensitivity of teacher evaluations to the distortion efforts that will inevitably arise from higher stakes.

2. ***How can we design policies to use better teacher-effectiveness measures, be they value-added scores, classroom observations, or something else, to improve teaching?***

The answer to this question may seem obvious. But it is not. There are, at an abstract level, three options: One can use the performance measures to provide feedback to teachers about how to improve instruction, to reward teachers who are rated highly, or to punish teachers who are rated poorly.

We have little evidence that any of these approaches can be used productively for teachers, and we have even less evidence that value-added scores, in particular, can contribute to such use. Providing feedback requires texture about the areas in which a teacher is performing well or badly. It is not at all clear that value-added scores—which amount to a single number—can be used for this kind of formative purpose.

Value-added seems more promising as a tool for summative evaluation. But experimental analyses of pay-for-performance programs in the United States have yielded disappointing results; teachers eligible for bonuses do not produce higher value-added scores—let alone better teaching along other dimensions—than do control-group teachers who are ineligible for bonuses. There have been no comparable experimental evaluations of the use of value-added for retention decisions, but there is highly suggestive evidence from the Air Force Academy, which bases its adjunct professor retention decisions in part on measured student achievement, that such uses lead to important distortions of the performance measure.[14]

Notes and References

1 The earlier reports were reviewed by the National Education Policy Center. See:
Rothstein, Jesse (2011). *Review of "Learning about Teaching: Initial Findings from the Measures of Effective Teaching Project."* Boulder, CO: National Education Policy Center. Retrieved January 30, 2013, from http://nepc.colorado.edu/thinktank/review-learning-about-teaching.

Guarino, S. & Stacy, B. (2012). *Review of "Gathering Feedback for Teaching: Combining High-Quality Observation with Student Surveys and Achievement Gains."* Boulder, CO: National Education Policy Center. Retrieved January 30, 2013, from http://nepc.colorado.edu/thinktank/review-gathering-feedback.

2 The MET study reports from 2010, 2012 and 2013 releases may be accessed at http://www.metproject.org/reports.php.

The five papers comprising the 2013 Final Report are:

- *Ensuring Fair and Reliable Measures of Effective Teaching: Culminating Findings from the MET Project's Three-Year Study: Brief*
  http://www.metproject.org/downloads/MET_Ensuring_Fair_and_Reliable_Measures_Practitioner_Brief.pdf;
- *Feedback for Better Teaching: Nine Principles for Using Measures of Effective Teaching*
  http://www.metproject.org/downloads/MET_Feedback%20for%20Better%20Teaching_Principle%20Paper.pdf;
- *A Composite Estimator of Effective Teaching: Research Paper*
  http://www.metproject.org/downloads/MET_Composite_Estimator_of_Effective_Teaching_Research_Paper.pdf;
- *The Reliability of Classroom Observations by School Personnel: Research Paper*
  http://www.metproject.org/downloads/MET_Reliability_of_Classroom_Observations_Research_Paper.pdf;
- *Have We Identified Effective Teachers? Validating Measures of Effective Teaching Using Random Assignment: Research Paper*
  http://www.metproject.org/downloads/MET_Validating_Using_Random_Assignment_Research_Paper.pdf.

3 Results are reported in:
Kane, T. J., McCaffrey, D. F., Miller, T., & Staiger, D. O. *Have We Identified Effective Teachers? Validating Measures of Effective Teaching Using Random Assignment* (research paper). Seattle, WA: Bill & Melinda Gates Foundation. Retrieved January 30, 2013, from http://www.metproject.org/reports.php.

4 Kane, T.J. & Staiger, D.O (2008). *Estimating teacher impacts on student achievement: An experimental evaluation* (Working Paper No. 14607). Cambridge, MA: National Bureau of Economic Research.

5 See :
Bill & Melinda Gates Foundation (2010). *Learning about Teaching: Initial Findings from the Measures of Effective Teaching Project* (Research Paper). Seattle, WA: Author. Retrieved December 16, 2010, from http://www.metproject.org/downloads/Preliminary_Findings-Research_Paper.pdf.

and
Rothstein, Jesse (2011). *Review of "Learning About Teaching: Initial Findings from the Measures of Effective Teaching Project."* Boulder, CO: National Education Policy Center, 3. Retrieved January 30, 2013, from http://nepc.colorado.edu/thinktank/review-learning-about-teaching.

6 Kane, T.J. & Staiger, D.O., *et al.* (2012, January). *Gathering Feedback for Teaching: Combining High-Quality Observation with Student Surveys and Achievement Gains* (MET Project research paper). Seattle, WA: Bill & Melinda Gates Foundation. Retrieved January 18, 2012, from http://www.metproject.org/reports.php.

7 Mihaly, K., McCaffrey, D.F., Staiger, D.O., & Lockwood, J.R. (2013). *A Composite Estimator of Effective Teaching* (MET Project research paper). Seattle, WA: Bill & Melinda Gates Foundation. Retrieved January 30, 2013, from http://www.metproject.org/downloads/MET_Composite_Estimator_of_Effective_Teaching_Research_Paper.pdf.

8 Bill & Melinda Gates Foundation (2010). *Working with Teachers to Develop Fair and Reliable Measures of Effective Teaching* (MET Project white paper). Seattle, WA: Author, 1. Retrieved December 16, 2010, from http://www.metproject.org/downloads/met-framing-paper.pdf.

9 The MET studies consider two different sets of test-based measures, the state exams and alternative, more conceptually demanding tests. They presume that only the former would be used for teacher evaluations, as few districts are likely to administer the alternative tests on large enough scales.

10 The two reports also do not engage with research about the often-perverse incentives that arise from placing high-stakes (for schools or teachers) on students' test scores. While these incentive issues are of great policy import, they are understandably not addressed in the MET technical reports.

11 A previous MET report was criticized precisely for validating other measures only against value-added score:

Guarino, S. & Stacy, B. (2012). *Review of "Gathering Feedback for Teaching: Combining High-Quality Observation with Student Surveys and Achievement Gains."* Boulder, CO: National Education Policy Center. Retrieved January 30, 2013, from http://nepc.colorado.edu/thinktank/review-gathering-feedback.

12 The report presents the correlation between the composites and the "stable component" of the outcome to be predicted (in this case, the alternative-test VA). This is larger than the correlation between the composite and the next year's actual outcome. We compute the latter for elementary teachers based on estimates elsewhere in the report of the proportion of variance of the alternative-test VA that is attributable to the stable component. The report does not provide enough information to perform a similar computation for middle-school teachers.

13 Further analysis of the MET data could potentially shed light on some of these questions. For example, the six MET districts varied substantially in the experimental compliance rates; it would be of great interest to see whether the experimental results were similar in each. Also of interest would be an analysis of differences between the most compliant schools, where the compliant sample is most representative, and the schools where compliance rates were lower. But the existing MET report does not address these issues.

14 Carrell, S.E. & West, J.E. (2010). Does professor quality matter? Evidence from random assignment of students to professors. *Journal of Political Economy, 118* (3), 409-432.

DOCUMENTS REVIEWED:	**Ensuring Fair and Reliable Measures of Effective Teaching: Culminating Findings from the MET Project's Three-Year Study**
AUTHORS:	*Have We Identified Effective Teachers?*: Thomas J. Kane, Daniel F. McCaffrey, Trey Miller, and Douglas O. Staiger. *A Composite Estimator of Effective Teaching*: Kata Mihaly, Daniel F. McCaffrey, Douglas O. Staiger, and J.R. Lockwood.
PUBLISHER/THINK TANKS:	Bill and Melinda Gates Foundation
DOCUMENT RELEASE DATE :	January 8, 2013
REVIEW DATE:	January 31, 2013
REVIEWERS:	Jesse Rothstein, University of California Berkeley; William J. Mathis, University of Colorado Boulder
E-MAIL ADDRESSES:	rothstein@berkeley.edu; wmathis@sover.net
PHONE NUMBERS:	Rothstein: (510) 643-8561; Mathis: (802) 383-0058

SUGGESTED CITATION:

Rothstein, J. & Mathis, W.J. (2013). *Review of "Have We Identified Effective Teachers?" and "A Composite Estimator of Effective Teaching": Culminating findings from the Measures of Effective Teaching Project*. Boulder, CO: National Education Policy Center. Retrieved [date] from http://nepc.colorado.edu/thinktank/review-MET-final-2013.