NEPC
NATIONAL EDUCATION POLICY CENTER

Review of *Gathering Feedback for Teaching*

Reviewed By

Cassandra Guarino, Indiana University
Brian Stacy, Michigan State University

March 2012

Summary of Review

This second report from the Measures of Effective Teaching (MET) project offers ground-breaking descriptive information regarding the use of classroom observation instruments to measure teacher performance. It finds that observation scores have somewhat low reliabilities and are weakly though positively related to value-added measures. Combining multiple observations can enhance reliabilities, and combining observation scores with student evaluations and test-score information can increase their ability to predict future teacher value-added. By highlighting the variability of classroom observation measures, the report makes an important contribution to research and provides a basis for the further development of observation rubrics as evaluation tools. Although the report raises concerns regarding the validity of classroom observation measures, we question the emphasis on validating observations with test-score gains. Observation scores may pick up different aspects of teacher quality than test-based measures, and it is possible that neither type of measure used in isolation captures a teacher's contribution to all the useful skills students learn. From this standpoint, the authors' conclusion that multiple measures of teacher effectiveness are needed appears justifiable. Unfortunately, however, the design calls for random assignment of students to teachers in the final year of data collection, but the classroom observations were apparently conducted prior to randomization, missing a valuable opportunity to assess correlations across measures under relatively bias-free conditions.

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 *Gathering Feedback for Teaching: Combining High-Quality Observation with Student Surveys and Achievement Gains*

Cassandra Guarino, Indiana University
Brian Stacy, Michigan State University

I. Introduction

The Bill and Melinda Gates Foundation has funded a large-scale study to evaluate several methods of judging teacher effectiveness. This is timely as school districts are looking to implement teacher-evaluation systems to identify high- and low-performing teachers, and, in many cases, are using these evaluations to make important decisions. This second report to come out of the study, *Gathering Feedback for Teaching: Combining High-Quality Observation with Student Surveys and Achievement Gains*,[1] focuses on an evaluation method in which videotaped lessons, evaluated by trained raters, are used to assess teachers.

This is the first large-scale study of classroom observations as a means of evaluating teachers. Using classroom observations, specific feedback can be given to teachers to help improve their instruction. Instruction of non-cognitive skills or particular types of cognitive skills that are not captured by student test scores can be observed. Also, classroom observations may be more difficult than test-score-based evaluations to game.

Such a system raises several concerns, as well. Observation systems will likely be expensive, so the benefits of this system must be compared with its costs. It is unclear how well these measures correlate with student outcomes. The level of reliability of classroom observations in typical applications is unknown. Finally, it is not clear whether an evaluator can separate the effect of teachers from the effect of the composition of students in the classroom.

The MET study makes an important contribution to answering some of these questions. The authors study the reliability of classroom observations. In addition, they examine how well classroom observation ratings can predict student test score gains and discuss the possible benefits of using multiple measures in conjunction with each other.

II. Findings and Conclusions of the Report

The authors report five primary findings:

1. The study examines five classroom observation systems and finds that they all "were positively associated with student achievement gains" (p. 6).
2. For all five observation instruments, "Reliably characterizing a teacher's practice requires averaging scores over multiple observations" (p. 8).
3. "Combining observation scores with evidence of student achievement gains and student feedback improved predictive power and reliability" (p. 9).
4. "In contrast to teaching experience and graduate degrees, the combined measure identifies teachers with larger gains on the state tests" (p. 10).
5. "Teachers with strong performance on the combined measure also performed well on other student outcomes" (p. 11).

The authors also put forward three primary conclusions:

1. "Achieving high levels of reliability of classroom observations will require several quality assurances: observer training and certification; system-level 'audits' using a second set of impartial observers; and use of multiple observations whenever stakes are high" (p. 13).
2. "Evaluation systems should include multiple measures, not just observations or value-added alone" (p. 14).
3. "The true promise of classroom observations is the potential to identify strengths and address specific weaknesses in teachers' practice" (p. 14).

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

The authors provide a wide range of statistics and figures to support their findings. The five classroom observation instruments produce teacher effectiveness scores that are positively related to measured student achievement gains on standardized tests.[2] A variety of correlations are considered. The raw correlations between the observation instruments and test score gains in mathematics from prior years range from .09 to .27, and statistics that attempt to retrieve the correlation with "underlying value-added" are slightly larger, ranging from .12 to .34 (Table 13, p. 46). The correlations for English language arts are also positive but slightly lower than those for mathematics. They range from .06 to .08 for the raw correlations and .09 to .12 for the correlation with underlying value-added (Table 17, p. 53).

An important property of an evaluation method is reliability, which is discussed at length. The reliability of a classroom observation score is defined as the fraction of the variation in observation scores due to the true variation in teacher quality along the dimension studied by the instrument. This gives us a sense of how much noise there is in the measure relative to the signal of teacher quality. A reliability of one would indicate that any variation in the

classroom observation score is solely due to differences in teacher quality. A reliability near zero would indicate that the measure is so contaminated with other factors that it offers little information on teacher quality. As the authors point out, several factors besides the quality of the teacher can affect observation scores. For example, there could be differences in how raters perceive instructional quality. Also, each teacher has lessons that are stronger or weaker than others, so a particular lesson rated may be unrepresentative of the average quality of the teacher across several lessons. Other idiosyncratic factors in the classroom at a given time may also influence a rater's perception of the teacher.

In the MET study, the reported reliability of an overall teacher effectiveness score derived from the observation of a single lesson ranges from .14 to .37, depending on the instrument used (Table 11, p. 35). To add some context to this number, after making some simplifying normality assumptions and assuming the noise in the measure follows classical

It is unclear what samples are contributing to the various analyses presented in the report.

measurement error assumptions, if the reliability is .37, then the probability that a teacher one standard deviation above the average in teaching quality is rated below average using the measure is approximately 22%. Adding in multiple observations from different lessons and by different observers increases the reliability. Having four lessons scored and taking the average increases the reliability to a range of .39 to .67. Having a reliability of .67 means the probability of misclassifying a teacher as below average when he or she is actually one standard deviation above the mean is around 8%. This is certainly an improvement but the measure is still far from definitive – particularly for high-stakes applications. The authors' conclusion that averaging scores over multiple observations is important is certainly warranted, but, as their analyses reveal, even this may not be sufficient to avoid a nontrivial amount of misclassification.

When classroom observation measures, student evaluations, and teacher value-added measures are averaged together, the combined measure correlates more highly with a teacher's "underlying value-added" than does the observation measure alone. For example, the correlation between the Framework for Teaching instrument and underlying value-added for mathematics is .19 using the instrument alone but jumps to .72 when the instrument score is averaged with value-added measures and student evaluations (Table 16, p. 51). The corresponding correlations for the Framework for Teaching instrument and underlying value-added in English and language arts are .11 and .40 (Table 20, p. 55).

The combined measure of classroom observations using the FFT instrument, student achievement gains, and student survey information better predicts student achievement gains than does teaching experience or degree level. The difference in student achievement gains on standardized tests associated with teachers with a master's degree and those without is statistically insignificant at the 5% level, and the point estimates suggest that

teachers with master's degrees had student test gains that were 0.03 standard deviations larger than those without in math and actually 0.02 standard deviations lower gains for ELA. The difference in gains associated with teachers with 12 or more years of experience and those with fewer than three years are also statistically indistinguishable, but the point estimates suggest a gain that is 0.01 standard deviations larger for math and 0.02 larger for ELA for those with 12 or more years of experience (Figure 5, p. 10). However, there are reasons to think this result may be understated.[3] The difference in student achievement gains between teachers rated in the top 25% and those in the bottom 25% on the combined measure is larger in magnitude, 0.21 standard deviations for math and 0.07 for ELA, than the other differences and statistically significant at the .001 level. The authors therefore conclude that student achievement growth has a stronger relationship to the combination of measures being evaluated in this study than to teachers' levels of education and experience.

The observation ratings are also positively related to scores on other (more open-ended) standardized tests in math and English language arts and to self-reported student psychological outcomes. The instruments predict differences between the top and bottom quartiles on these outcomes in a statistically significant way, typically at the 1% level. The estimated difference for the Balanced Assessment in Mathematics exam was 0.13 standard deviations and 0.13 for the Stanford 9 reading exam. The difference between those in the top and bottom quartiles for student effort, as reported on student surveys, is 0.24 standard deviation units in math classes and 0.22 in ELA. For positive emotional attachment, the difference between the top and bottom quartile is 0.46 standard deviation units in math and 0.33 in ELA. (Tables 15 and 19, pp 49 and 55).

IV. The Report's Use of Research Literature

The report is sparing in its reference to research literature and does not review all the pertinent literature on the topic. Some points that we discuss below are raised in literature that is not discussed in the report.

V. Review of the Report's Methods

The study spans two years, 2009-10 and 2010-11, and collects several types of data on teacher effectiveness using a sample of nearly 3,000 volunteer teachers from grades 4 through 9 in six districts selected by the research team from different parts of the country. District selection criteria are not discussed nor is there a description of incentives, if any, offered to teachers in exchange for their participation. No information is given regarding the distribution of teachers across grades within the sites.

The data include student test scores, classroom observations, and student evaluations of teacher performance. The test scores comprise state assessments in reading and mathematics and one other, more open-ended, type of assessment in each of the two subject areas administered by the study team.

The test score data used in the report were collected for the year 2009-10, supplemented by test score data on 44% of the sampled teachers for 2008-09. Thus, value-added measures of teacher performance are computed for all teachers in 2009-10 and for a subset of teachers in 2008-09 as well. More than half of the teachers taught more than one section of their course in 2009-10, so more information was available on those particular teachers. For this subsample, value-added measures were also computed separately for different sections taught by the same teacher.

A key feature of the study design is that teachers were randomized to classrooms in 2010-11, thus affording researchers the opportunity to compute value-added measures of teacher performances with fewer validity issues for that year. However, the test scores used in this particular report were from 2009-10 and were thus apparently collected prior to randomization.

The classroom observations consisted of ratings of video-taped lessons in 2009-10 using five different rubrics.[4] Efforts were made to collect tapes of more than one lesson per teacher.

The student evaluations of teacher performance were based on a survey conducted in 2009-10. A description can be found in the initial MET report.[5] However, response rates are not reported, and we are not told whether the survey was voluntary or mandatory.

It was necessary to piece together the above description to the best of our understanding based on information interspersed throughout the report. It is difficult to locate a concise and complete description of the study design in the report (or in the prior report). This is problematic because the lack of clear and complete study design information effectively prevents other researchers from fully judging the import of the study's findings.

The research design appears to have several desirable features.

Multiple measures constitute a core strength of the study. Test scores are collected on two different types of tests—a high-stakes test in multiple-choice format and a low-stakes test in a more open-ended format—for each subject area. Four classroom-observation instruments are used for math and three for reading. The student evaluations add a dimension of teacher performance that is often neglected in studies of effectiveness.

Having repeated measures for the same teacher is of great importance. As a result, the study sheds considerable light on the reliability of various observation tools, and that of value-added measures as well.

The randomization of teachers to classrooms in the study's second year constitutes another key strength of the study by adding credibility to the value-added measures of teacher performance in that year. However, this report does not utilize post randomization data.

The study design also appears to have weaknesses, some of them easier to surmount than others. Generalizability is compromised by the fact that teachers were volunteers. The volunteer teachers were somewhat younger than average for their districts. It may be the

case that younger teachers are more willing to undergo randomization to students than more experienced teachers who may be used to getting students who are somewhat easier to teach. Younger teachers may be better or worse teachers than experienced teachers, the quality of their teaching may be more likely to fluctuate from one lesson to another, they may have more incentive to change their teaching if they do not have tenure, and they may be more subject to a Hawthorne effect—i.e., modifying their teaching as a result of being studied. Despite these considerations, unless entire schools or districts could be persuaded to require the participation of all their teachers, it would be difficult to find a way to enlist participation that overcomes the problems associated with voluntary participation.

An important concern is that it appears that neither the classroom observations nor the student evaluations will be conducted again after randomization. If this is the case, then it is very unfortunate, as these measures may be affected by non-random assignment of students to teachers in much the same way that test-based measures may be affected. It appears that the design has missed an important opportunity to detect true parallels and contemporaneous correlations across different measures.

There are other possible weaknesses in the design about which we can only speculate because not enough information about the design is provided. For example, the report does not reveal how districts were selected or whether any of the districts used value-added measures of teacher performance for evaluation purposes with stakes attached at the time of the data collection. What incentives were awarded to the teachers who participated is also not reported. Moreover the report provides no information on how the randomization was conducted and how teacher volunteerism may have played into this. Can a principal successfully randomize teachers to classrooms if not all teachers volunteer for the project and if non-random assignment has been used in the past to improve retention of particular teachers? We don't know.

In addition, the issue of variability in measurement across districts and grades is insufficiently discussed. It is unclear what samples are contributing to the various analyses presented in the report. Although the study includes nearly 3,000 teachers, several restrictions noted in the appendix reduce the sample size to 1,333. When these are further divided by district and grade, it may be that certain analyses are being conducted on unevenly distributed or small samples. The fact that we have no information on the distribution of teachers across districts and grades and no "N"s reported on the various tables gives rise to questions. In the effort to reach more national sites and more grades, the study designers have sacrificed some potential uniformity in the sample and thus increased the noise in their aggregate measures. The authors should have reported Ns on the tables as well as the final composition of the sample broken down by district and grade, since the relationships explored may differ along these dimensions.

There are issues with some of the methods used in the report. An issue related to the value-added approach is the use of average residuals rather than coefficients on teacher dummies in computing teacher performance measures. The authors use a specification in which student test scores are regressed on prior test scores, other student characteristics, and classroom characteristics. The residuals in these regressions are then averaged by

classroom to create the teacher value-added performance measure. Student achievement gains are measured by residuals in a regression of current test scores on prior test scores controlling for a set of student and classroom characteristics. Since teachers are not randomly assigned to classrooms in this round of analysis, correlation between lagged test scores (or other student characteristics) and teacher assignment is not "partialed out" in the residual method and can lead to bias.[6] Regression analysis accomplishes this partialling out when the teacher dummies and student characteristics are both explicitly included in the specification. The average residualing method is biased when teacher assignment is nonrandom. It may be the case that the correlation is minimal in this sample—for example, tracking of some sort is a precondition for the correlation and there may be virtually no tracking in these classrooms—however, no information is provided on whether or not attempts were made to detect any type of tracking.[7] Clearly the teacher dummy variable specification was not used in order to allow for the inclusion of classroom "peer" variables without a large loss of sample. However, it is not clear that the one sacrifice was more justifiable than the other. Although the authors mention in a footnote (footnote 36, p. 42) that their specification differs little from the teacher fixed-effects specification, it is unclear whether they have verified this with their data or are simply making this assumption.

An additional methodological issue concerns the computation of the inflated correlations from raw correlations, which is done to remove the effect of noise in the measures. This computation assumes the measure is bias-free and that the error is independent of the "true score." There are reasons to question both assumptions.[8] First, because the randomization is not yet done, the measures may be biased. Second, the different characteristics of the state tests used in different districts may lead to non-random differences in the ways teachers are being evaluated.

Other issues arise from the report's presentation and omissions. Its preferred statistic, the difference between the top and bottom quartile, is somewhat difficult to interpret. The myriad types of correlations used in the various tables are not always clearly explained or labeled. The lack of reported standard errors or p-values is an important omission.[9] The lack of a simple raw correlation table including *all* the teaching effectiveness measures and significance is an omission that is both surprising and frustrating. A full technical appendix showing all equations and the exact formulas used in calculations would have been helpful.

VI. Review of the Validity of the Findings and Conclusions

The study aims to shed light on the question of how to measure effective teaching and provides some excellent data for that purpose. Many of the analyses, however, focus on a somewhat different research question: how well do classroom observations and student evaluations predict teacher value-added measures on tests? Unfortunately, none of the measures including value-added measures (even after randomization), can claim to provide a complete picture or gold standard of effectiveness. The title of the last chapter, "Validating Classroom Observations with Student Achievement Gains," signals a point of

view that considers value-added measures of teacher performance to be valid *a priori*, a view that is still disputed in the research literature.[10]

Classroom observations do not necessarily have to be strongly related to student test score gains for the measures to provide useful information on teacher effectiveness. Many of the competencies evaluated may point out contributions to learning that are not well captured on standardized tests or that may be more strongly related to non-cognitive skills, which some studies have suggested greatly influence later life outcomes.[11] As an example, one of the instruments, the *Framework for Teaching*, assesses teachers on the following dimensions: creating an environment of respect and rapport, establishing a culture of learning, managing classroom procedures, managing student behavior, communicating with students, using questioning and discussion techniques, engaging students in learning,

One hoped-for outcome of the study will be to revise classroom observation instruments.

and using assessments in instruction (p. 22). A teacher with expertise in managing student behavior, for instance, may not produce notably large test score gains as a result of this expertise, but may have a long-term impact on skills students need to be successful later in life. Validating classroom assessments may require more complex approaches than a comparison of the measures with test-score gains. And since the report concludes that "evaluation systems should include multiple measures, not just observations or value-added alone," it would be helpful to take a somewhat agnostic view of the relative validity of various performance indicators.

Combining the measures together did improve both the predictive power for student test score gains and overall reliability compared with the classroom observation ratings on their own. However, the improved predictive power and reliabilities are driven by the high correlation between student achievement gains and underlying value-added and the high reliability of the student surveys, which is evident in tables 16 and 20 (pp. 51 and 55). It would be interesting to see the correlation and reliability with only the student achievement gains and student surveys together. If the sole goal is to maximize reliability and correlation with test scores, then the classroom observations do not appear to add much and dropping them altogether might be best. We view the evidence presented in the report as suggesting that classroom observations are somewhat noisy and weakly related to student achievement gains. It would be possible to come to a different conclusion than the authors: that classroom observations are not worth the time, cost, and effort. However, test scores may not pick up all the useful skills that students learn in schools, and classroom observations may convey different information on teacher quality.

It should be noted that the authors did take some steps to collect other indicators of student achievement, such as the survey information on student effort and emotional attachment. However, it is not clear that these measures adequately capture all of the

important contributions teachers make either. More research will be needed to validate classroom observations as a measure.

Much training on the part of raters was needed to calibrate the ratings of classroom observations. Moreover, the disattenuated correlations below 0.7 in Table 10 across some of the instruments suggest that the instruments may emphasize different skills. Moreover, the observation measures seemed to do a good job of picking up certain skills like classroom management but were less reliable in other important areas for which teachers might need feedback, such as "questioning," "working with students," or "intellectual challenge" (Table 11, p. 35). The observation rubrics differed in what they tried to assess. It is not surprising that the two "generic" instruments, the CLASS and the FFT, which focused more on classroom dynamics, were more strongly correlated with each other than with the instruments that attempted to get at the transmission of content knowledge. Given these findings, the claim that, "The true promise of classroom observations is the potential to identify strengths and address specific weaknesses in teachers' practice," seems premature. It is not clear that the observation rubrics analyzed in this study deliver that information in a clear and reliable manner at this time.

These issues should not obscure the fact that the report lays solid groundwork in analyzing and synthesizing the results of different observation instruments—work that can be used to improve these instruments or the methods used to combine them to capture key aspects of teaching quality. One hoped-for outcome of the study will be to revise classroom observation instruments. Given that this form of evaluation is relatively expensive, it is important to have a good justification for adopting it and for adopting particular instruments.

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

The MET study has undertaken considerable work to quantify effective teaching. The data and information contained in the report further our knowledge of the interrelationships among different measures of teaching quality and provide a valuable contribution to teacher-effectiveness research. Furthermore, the authors are committed to making the data available to other researchers in the future, which will likely produce many helpful studies. In addition, data from the MET study can help lay the groundwork for the development of better measures, based on both tests and observations, in the future.

This particular report studies classroom observation ratings in depth and holds them up to greater scrutiny than they have been held to in the past. This is a particularly helpful contribution to the knowledge base on teacher effectiveness measures because until now, much of the national debate over the validity and precision of teacher effectiveness measures has focused almost exclusively on value-added based on test scores. This report shows that observation measures likely have at least as many validity and reliability issues as do measures based on standardized tests. Thus they should be considered equally controversial.

This report illuminates the complexity of the construct "effective teaching." One explanation for the finding that classroom-observation scores and measured teacher value-added are positively but weakly related could be that teaching may be more multi-faceted than is measured by either standardized tests or observation rubrics. In fact, they may be capturing somewhat different aspects of teacher effectiveness, with classroom observations slightly more adept at picking up the transmission of non-cognitive skills and test scores somewhat more helpful in picking up the transmission of cognitive skills. Despite the imprecision of the various indicators, taken together they might supply valuable, though not definitive, information regarding a teacher's effectiveness.

In addition to addressing several questions related to classroom observation instruments, the report raises several other interesting questions. Given the cost and painstaking efforts required to train classroom observers, can districts implement a classroom-evaluation system that creates a highly reliable measure? Do the classroom observations pick up on non-cognitive skills? Can the feedback they provide improve teaching effectiveness in a meaningful way?

As policymakers consider the policy implications of this study's findings, two points made in the report and reiterated in the research literature should be kept in mind. One is that the status quo—minimal standards for tenure and salary premiums based on education and experience—may be missing aspects of teacher quality that can be picked up by classroom observation, student surveys, and value-added data. The other is that, given the current state of the art, high stakes should not be attached to any of these measures in isolation.

Notes and References

1 Kane, T.J. & Staiger, D.O., et al. (2012, Jan.). *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.

2 Student achievement gains are measured by residuals in a regression of current test scores on prior test scores controlling for a set of student and classroom characteristics.

3 As explained in Rockoff (2004), selection bias in the effect of experience on student achievement gains in a cross-sectional study could be created by less effective teachers being more likely to fired, more effective teachers moving on to other higher-paying occupations, teacher quality varying by cohort, as well as other reasons. Rockoff actually finds evidence that experience is a statistically significant predictor of reading as well as math computation test scores after using panel data. The relationship between experience and student score gains could be understated in this report.

Rockoff, J. (May 2004). The impact of individual teachers on student achievement: evidence from panel data. *The American Economic Review 94* (2), 247-252.

4 The rubrics used in the study are: CLASS, Framework for Teaching, PLATO Prime, MQI Lite, and UTOP. CLASS and Framework for Teaching can be used for all subjects. PLATO Prime assesses English language arts teaching. MQI assesses mathematics. UTOP assesses math, science, and computers. A brief description of the rubrics is available in Table 3 (19).

5 The Tripod survey is described in (December 2010) *Learning About Teaching: Initial Findings from the Measures of Effective Teaching Project*. MET Project Research Paper. Seattle, WA: Bill & Melinda Gates Foundation. Retrieved January 18, 2012, from http://www.metproject.org/reports.php.

6 See Guarino, C., Reckase, M., & Wooldridge, J. (2012). *Can Value-Added Measures of Teacher Performance be Trusted?* Working Paper.

7 For methods of detecting tracking, see Dieterle et al. (2012), Rothstein (2010), Aaronson, Barrow, and Sander (2007), and Clotfelter, Vigdor, and Ladd (2006). Rothstein (2011) also raises this issue in his review of the prior MET report.

8 Rothstein (2011) raises similar issues in his review of the prior MET report.

Rothstein J. (2011). *Review of "Learning About Teaching: Initial Findings from the Measures of Effective Teaching Project."* Boulder, CO: National Education Policy Center. Retrieved January 20, 2012, from http://nepc.colorado.edu/thinktank/review-learning-about-teaching.

9 Where it is difficult to compute standard errors, such as perhaps in the adjusted correlations, bootstrapping procedures could be used.

10 Several papers cast doubt on the validity of the value-added measures of teacher performance under certain conditions. See, for example,

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

Corcoran, S., Jennings, J., & Beveridge, A., (2011). *Teacher effectiveness on high- and low-stakes tests*. Working Paper. Retrieved February 8, 2012, from https://files.nyu.edu/sc129/public/papers/corcoran_jennings_beveridge_2011_wkg_teacher_effects.pdf.

Guarino, C., Reckase, M., & Wooldridge, J. (2012). *Can Value-Added Measures of Teacher Performance be Trusted?* Working Paper.

Further, it isn't clear whether all the skills that make for a good teacher are identified using test score gains. A recent paper does suggest, however, that higher value-added estimates are related to positive later life outcomes for students:

Chetty, R., Friedman, J., & Rockoff, J. (2011). *The Long-Term Impacts of Teachers: Teacher Value-Added and Student Outcomes in Adulthood*. NBER Working Paper No. 17699. Retrieved January 31, 2012, from http://www.nber.org/papers/w17699.

11 See Heckman, J., Rubinstein, Y., (May 2001). The importance of noncognitive skills: Lessons from the GED testing program. *The American Economic Review 91* (2), 145-149.

This idea is also discussed in the article *How Does Your Kindergarten Classroom Affect Your Earnings? Evidence From Project STAR*.

Aaronson, D., Barrow, L., & Sander, W. (2007). Teachers and Student Achievement in the Chicago Public High Schools. *Journal of Labor Economics, 25*(1), 95-135.

Chetty, R., Friedman, J., Hilger, N., Saez, E., Schanzenbach, D., Yagan, D. (2010). *How Does Your Kindergarten Classroom Affect Your Earnings? Evidence From Project STAR*. NBER Working Paper No. 16381. Retrieved January 31, 2012, from http://www.nber.org/papers/w16381.

Chetty, R., Friedman, J., & Rockoff, J. (2011). *The Long-Term Impacts of Teachers: Teacher Value-Added and Student Outcomes in Adulthood*. NBER Working Paper No. 17699. Retrieved January 31, 2012, from http://www.nber.org/papers/w17699.

Clotfelter, C.T., Ladd, H.F., & Vigdor, J.L. (2006). Teacher-Student Matching and the Assessment of Teacher Effectiveness. The Journal of Human Resources, 41(4), 778-820.

Dieterle, S., Guarino, C., Reckase, M, & Wooldridge, J. (2012). *How Do Principals Assign Students to Teachers? Finding Evidence in Administrative Data and the Implications for Value-Added*. Unpublished Draft.

Corcoran, S., Jennings, J., & Beveridge, A., (2011). *Teacher effectiveness on high- and low-stakes tests*. Working Paper. Retrieved February 8, 2012, from https://files.nyu.edu/sc129/public/papers/corcoran_jennings_beveridge_2011_wkg_teacher_effects.pdf.

Guarino, C., Reckase, M., & Wooldridge, J. (2012). *Can value-added measures of teacher performance be trusted?* Working Paper.

Heckman, J. & Rubinstein, Y. (May 2001). The importance of noncognitive skills: Lessons from the GED testing program. *The American Economic Review 91* (2), 145-149.

Rockoff, J. (May 2004). The impact of individual teachers on student achievement: evidence from panel data. *The American Economic Review 94* (2), 247-252

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

Rothstein J. (2011). *Review of "Learning About Teaching: Initial Findings from the Measures of Effective Teaching Project."* Boulder, CO: National Education Policy Center. Retrieved January 20, 2012, from http://nepc.colorado.edu/thinktank/review-learning-about-teaching.

DOCUMENT REVIEWED:	**Gathering Feedback for Teaching: Combining High-Quality Observation with Student Surveys and Achievement Gains**
AUTHORS:	Thomas J. Kane and Douglas O. Staiger
PUBLISHER:	Bill and Melinda Gates Foundation
DOCUMENT RELEASE DATE:	January 2012
REVIEW DATE:	March 2012
REVIEWERS:	Cassandra Guarino, Indiana University Brian Stacy, Michigan State University
E-MAIL ADDRESSES:	guarino@indiana.edu stacybri@msu.edu
PHONE NUMBERS:	Cassandra Guarino: (812) 856-2927 Brian Stacy: (843) 696-5496

SUGGESTED CITATION:

Guarino, C. & Stacy, B. (2012). *Review of "Gathering Feedback for Teaching."* Boulder, CO: National Education Policy Center. Retrieved [date] from http://nepc.colorado.edu/thinktank/review-gathering-feedback.