

# **Learning from Recent Advances in Measuring Teacher Effectiveness**

**Meeting Summary**

August 9, 2012

## Location

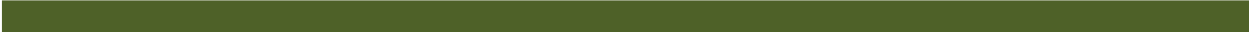
Institute of Education Sciences (IES) Board Room  
80 F Street NW  
Washington, DC 20001

## Research Community Participants

Damian W. Betebenner, Ph.D., Senior Associate, National Center for the Improvement of Educational Assessment  
Henry Braun, Ph.D., Director of the Center for the Study of Testing, Evaluation and Education; Boisi Professor of Education and Public Policy, Boston College, Lynch School of Education  
Sean P. Corcoran, Ph.D., Associate Professor of Educational Economics, New York University's Steinhardt School of Culture, Education, and Human Development; Affiliated Faculty Member, Robert F. Wagner Graduate School of Public Service  
Linda Darling-Hammond, Ed.D., Charles E. Ducommun Professor of Education, Stanford University; Faculty Co-Director, Stanford Center for Opportunity Policy in Education  
John N. Friedman, Ph.D., Assistant Professor of Public Policy, Harvard Kennedy School; Faculty Research Fellow, National Bureau of Economic Research  
Dan Goldhaber, Ph.D., Director, Center for Education Data & Research; Professor in Interdisciplinary Arts and Sciences, University of Washington Bothell  
Andrew Ho, Ph.D., Assistant Professor, Harvard Graduate School of Education  
Thomas Kane, Ph.D., Professor of Education and Economics, Harvard Graduate School of Education; Faculty Director, Center for Education Policy Research  
Helen F. Ladd, Ph.D., Edgar T. Thompson Distinguished Professor of Public Policy and Professor of Economics, Duke University  
Robert C. Pianta, Ph.D., Dean, Curry School of Education, University of Virginia; Director, Center for Advanced Study of Teaching and Learning, Novartis Professor of Education  
Jonah E. Rockoff, Ph.D., Sidney Taurel Associate Professor of Business, Columbia Graduate School of Business; Faculty Research Fellow, National Bureau of Economic Research  
Jesse Rothstein, Ph.D., M.P.P., Associate Professor of Public Policy and Economics, University of California, Berkeley; Research Associate, National Bureau of Economic Research

## IES and U.S. Department of Education Staff

John Q. Easton, Director, IES  
Elizabeth Albro, Acting Commissioner/Associate Commissioner for Teaching and Learning, National Center for Education Research  
Jo Anderson, Jr., Senior Advisor to the Secretary  
Karen Cator, Director, Office of Educational Technology  
Nadya Chinoy Dabby, Associate Assistant Deputy Secretary, Office of Innovation and Improvement  
Deb Delisle, Assistant Secretary, Office of Elementary and Secondary Education  
Monica Herk, Executive Director, National Board for Education Sciences, IES  
Brad Jupp, Senior Program Advisor on Teacher Initiatives



Tyra Mariani, Deputy Chief of Staff for Operations and Strategy, Office of the Secretary  
Ruth Curran Neild, Associate Commissioner, Knowledge Utilization, National Center for  
Education Evaluation and Regional Assistance

Ellie Pelaez, Assistant to the IES Director

Audrey Pendleton, Associate Commissioner, Evaluation, National Center for Education  
Evaluation and Regional Assistance

Jason Snyder, Deputy Assistant Secretary for Policy, Office of Elementary and Secondary  
Education

Elizabeth Warner, Economist, National Center for Education Evaluation and Regional Assistance

Joanne Weiss, Chief of Staff, Office of the Secretary

Ann Whalen, Director, Policy and Program Implementation, Implementation and Support Unit

Michael Yudin, Acting Assistant Secretary, Office of Elementary and Secondary Education

# Meeting Summary

## Welcome, Introductions, and Framing of the Meeting

### **John Q. Easton, Ph.D., and Joanne Weiss**

Dr. Easton called the meeting to order at 10:08 a.m. and thanked the participants for coming. He emphasized that IES believes strongly in the use of research as a powerful tool for improving practice and policy. The past 5 years have brought an explosion in research on teacher effectiveness, value-added measures (VAMs), and student growth models (SGMs), but no unequivocal conclusions on which to draw, said Dr. Easton. Not only do different researchers have different perspectives and interpretations of the literature, but also it is not clear that research is being disseminated effectively across communities. Therefore, IES convened this meeting to collect input from a multidisciplinary group of experts with a range of perspectives. Dr. Easton said the meeting is designed to gather, from researchers, a better understanding of areas of common ground, as well as the bases underlying any areas of disagreement.

The participants introduced themselves. In advance of the meeting, many participants provided brief written summaries outlining their perspectives on recent advances in the use of VAMs and SGMs as measures of teacher effectiveness (see Appendix A).

Ms. Weiss said she believes there is consensus that teachers matter enormously in education relative to other school-level inputs, but there is a need to assess what is working well, what is not, and what policy and practice changes should be undertaken to ensure that the teaching workforce is effective in helping kids. Over the past 4 years, much has changed as some teacher effectiveness policies have kicked in. The culture of education is becoming more data driven, said Ms. Weiss. However, the field lacks the capacity to use data well. Often data are used for sorting and ranking, not for improvement. Ms. Weiss hoped that the meeting participants would elucidate what is working, what is not, and how IES can help with policies and implementation support.

## Recent Research Advances

### **Henry Braun, Ph.D., M.Sc., Lead Discussant**

#### **Tyra Mariani, Facilitator**

Dr. Braun pointed out that in their advance summaries, most of the participants mentioned publications by Jesse Rothstein<sup>1</sup> and by Raj Chetty, Jonah E. Rockoff, and John N. Friedman.<sup>2</sup> Many participants focused on the dynamic, non-random allocation of students to teachers in Dr. Rothstein's paper, but few discussed that paper's findings on the sources of bias in VAMs and how to collect data on bias. The Chetty et al. paper significantly informed discussion by providing evidence that individual teachers matter and that VAMs and other methods have a role in identifying effective teachers and in distinguishing classroom effects from teacher effects.<sup>3</sup> Dr. Braun noted that much research will be framed around the findings of these publications.

In terms of VAMs, Dr. Braun said that their stability, fairness, and utility are “in the eyes of the beholder,” ultimately, that leaves validity as the only objective basis for judging VAMs. Most research focuses on evidential or scientific validity, which is appropriate, but some meeting participants have identified the need to consider consequential validity—that is, the consequences of introducing VAMs or other indicators into teacher accountability systems. Scientific validity takes into account whether results that have been found by researchers are reasonably generalizable to the real-world settings where VAMs will be applied. Consequential validity considers the ways in which VAMs have been implemented in professional accountability systems. How do differences in the way VAMs are implemented play out in terms of policy consequences?

Two schools of thought are present in the research, said Dr. Braun:

- For one school, there is enough evidence that VAMs provide real information about teacher effectiveness to support their role in teacher evaluations, particularly in concert with other indicators.
- For the other school of thought, there are real problems with VAMs; it is not clear how they will function in real-world settings, and so VAMs should only be used with caution.

Both viewpoints are reasonable, said Dr. Braun, but it is important to recognize that states are already using VAMs for teacher evaluation with various degrees of forethought and analysis. He recommended taking advantage of these natural experiments in the states to make up for the insufficient number of studies to date of VAMs being used for actual teacher evaluation. He urged that researchers be “planful” about designing their studies to extract the most information possible from the natural experiments that are available.

Looking to the future, said Dr. Braun, it is important that we consider both the scientific and consequential validity of measures. The Gates Foundation’s Measures of Effective Teaching (MET) will be highly influential and will inform the design of teacher evaluation systems. Researchers should consider how to maximize the utility of the enormous database the project will yield. Finally, Dr. Braun cited a Wall Street Journal opinion piece on the Affordable Care Act (ACA) that questioned whether imposing pay-for-performance measures on physicians would do more harm than good by diminishing their overall sense of professional responsibility. He suggested that a similar question might apply to pay-for-performance as applied to teachers.

## Discussion

Dr. Darling-Hammond noted that one of the significant benefits of the MET studies supported by the Gates Foundation is that they are using multiple measures of learning to validate observational and other tools used to assess teachers. There is a set of research that is not only about the use of value-added modeling for direct measurement of teachers but also about using value-added strategies to assess the potential of other measures of teacher effectiveness. That research has been both interesting and fruitful, she said.

Dr. Rothstein commented on Dr. Darling-Hammond's statement, saying that there is no agreement among researchers on which direction the validation of teacher evaluation measures should go. For example, suppose that two different teacher evaluation measures fail to correlate with each other. They might be equally good, but simply measuring different things; even to the extent that they are measuring the same thing, we don't know which one is doing the better job of it. Researchers end up talking past one another because they have different views of the hypothesis being tested.

Dr. Goldhaber commented, and Dr. Rockoff agreed, that the value of VAMs for teacher assessment can only be determined relative to the effectiveness of the available alternatives for teacher assessment.

Dr. Rockoff went on to point out that a teacher's value-added, as measured by different tests, is positively correlated but perhaps not as highly correlated as some might think—indicating that “what you're testing matters.” If the school system is using a test that does not measure learning that the school system cares about, then it is going to get a VAM that tells it about teacher performance relative to learning that it doesn't care about. Systems really have to take a stand that if they're going to use a VAM, it should be based on a student test that assesses the material that they want students to be learning.

Dr. Rockoff went on to discuss the evidence related to teacher specialization—that is, when there is a high value-added teacher relative to a specific test, then is that teacher good at raising scores on that test for all students, or just for certain subsets of students, such as high-achieving students or low-achieving students? Dr. Rockoff stated that the balance of evidence suggests there is a general construct of teacher effectiveness that spans students and to some extent grades and schools (i.e., a high value-added teacher tends to be a high value-added teacher in a variety of contexts), but there is also evidence of some specialization (i.e., that particular teachers are more skilled at raising the test scores of particular subsets of students). He noted that these questions merit further research.

Dr. Betebenner raised the distinction between *evaluating* teacher effectiveness and *measuring* it.

In addition, some managers have recently pointed out that a hitter's on-base percentage far superior to his batting average is a predictor of the most important goal, namely scoring runs.

Dr. Ladd criticized the frequently used analogy to baseball, pointing out that batting averages may be important measures of effectiveness for some players but not for others such as pitchers; moreover, it is only one measure in the context of several that can tell you something about a player and possibly predict future performance. She also pointed out that in building a team, players with a variety of skills are needed; one wouldn't want a team made up exclusively of good batters but lacking in defensive skills. With respect to the question of instability, she added that batting averages are calculated by observing many at-bats over the course of a long season, and thereby generate greater signals relative to teacher VAMs, which are based on many fewer data points. Dr. Corcoran extended the criticism of the baseball analogy, saying a low batting average offers a clear indication of where improvement is needed, since it measures a specific activity (hitting), while teacher VAMs do not.

Dr. Corcoran also said that the state of the art in value-added modeling is near its frontier, and he was skeptical that models would improve very much due to technical advances in modeling. In other words, inadequacies of VAMs for teacher evaluation are not likely to be resolved by better statistical models.

Dr. Rockoff echoed that opinion and suggested that better data may be what is required to improve models' accuracy. One example of what is needed is better data on which teachers taught a particular student and for how long, rather than current data, which typically lists only the student's teacher on the day of the test or the teacher who administered the test, he said. Dr. Ho agreed that richer data would yield richer models.

Dr. Ho said the less we know about a given domain of expertise such as teaching, the fewer dimensions we assume it has. As an example, Dr. Ho said that he does not know much about baseball, so he simply assumes that there are good baseball players and bad baseball players, but he does not understand or appreciate the difference between batting averages and on-base percentages, and the different information they convey. Similarly, in Olympic track and field, there is the decathlon and there is the 100 meter dash. One requires all-around athletic talents and the other requires a very specific talent. A person who doesn't know much about track and field might think there are simply "good athletes" who could excel at either event. Turning to teaching, there is little appreciation, even among teachers, of just how many dimensions there are in teacher effectiveness, Dr. Ho commented, and so we end up having very simple, unidimensional rankings of teacher effectiveness.

Dr. Darling-Hammond stated that the properties of the student tests have not been considered adequately in the research and that instead we treat scores from different tests as if they mean the same thing. She explained that the demographics of the students taking the test matter, as does whether the test has a high "ceiling" or "floor." Dr. Darling-Hammond described a teacher of gifted students in Houston, TX, who did not show much "value-added" because the test had such a low ceiling that all her students were close to 100 percent even before she taught them.

---

Conversely, the California state test has a high ceiling and a high floor. The state requires students to take the state test in English after only 10 months in the country. The rankings of teachers who work with a lot of English language learners (ELLs) may suffer as a result because the ELLs score below the test's floor; therefore, their progress cannot be measured. Dr. Darling-Hammond said there is no research agenda looking across states at the different properties of state tests, the demographics of the students taking them, and the timing of the tests.

Going back to Dr. Rockoff's comments, Dr. Rothstein pointed out that in addition to more detailed data on student's exposure to individual teachers, we also need more data concerning the effect on student learning of the interaction among the various teachers a student experiences during his or her school years. The effect of the previous year's teacher on a student matters, but researchers don't know how to handle that methodologically. Nevertheless, that is not a defense for ignoring the effect, he said. Turning to another point, Dr. Rothstein stated that researchers tend to use convenience samples for their research, but that the properties of teacher effectiveness measures can be expected to vary with the stakes placed on them. He called for more IES support of research on this issue. For example, do the correlations between two different measures of teacher effectiveness change when high stakes are attached to one of the measures but not to the other?

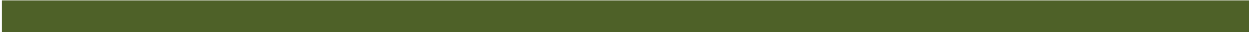
Dr. Ho said one of his colleagues is seeking to embed anchor items within tests that would help detect whether score inflation is occurring after a test becomes a high-stakes test.

Regarding the interaction effects that multiple teachers have on a student, Dr. Pianta pointed out that although we conceive of education as a cumulative experience, researchers' value-added models to date have focused on students' exposure to one teacher in a particular year. He stated that in his own research, using simple pre- and post-growth measures during the pre-K/early elementary grades, his team has observed teacher interaction effects on student learning.

Dr. Kane noted that academics excel at emphasizing the nuances and complexities inherent in research findings, but policymakers ultimately need to make decisions and they are seeking information that will help them make those decisions. Dr. Kane said that one of the lessons of the MET study has been how hard it is to reliably measure teacher effectiveness using alternatives to value-added approaches. For example, it is much harder and more expensive to generate reliable measures of teacher effectiveness from classroom observations than from VAMs. He added that, surprisingly, student survey results are more reliable for assessing teacher effectiveness than measures based on anything less than four classroom observations. Rather than look at the weakness of any single measure—such as value-added—in isolation, we should be focused on the strengths and weaknesses of a combination of measures.

Dr. Braun echoed Dr. Rockoff's and Dr. Betebenner's comments regarding the need for more data, on students' exposure to individual teachers and on the context in which individual teachers teach, in order to better understand VAM scores and make individual teacher





evaluations fairer. Dr. Braun also stated that on the student side we need to assess a broader range of desirable student outcomes, including a broader range of academic subjects, attitudes, and dispositions. By ignoring this broader range of student outcomes, our VAMs do not reflect them and we're missing part of what good teacher practice is meant to produce.

Dr. Ladd added that another limitation is that existing data for VAMs are limited to math and reading scores for students in grades 3–8. Hence, sophisticated value-added models cannot be applied to subjects other than math and reading, to grades K-2, or to high school.

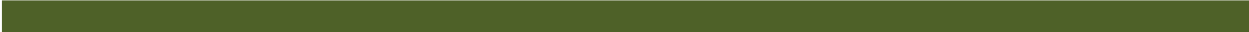
Dr. Pianta stated his sense that researchers agree that VAMs do detect some “signal” regarding teacher effectiveness despite the “noise” in them. However, in the future, he would like to see value-added modeling that controls for the effect of exposure to a specific curriculum, since some curricula are more effective than others, regardless of teachers’ fidelity of implementation.

Dr. Kane agreed with Dr. Darling-Hammond that it is a problem when “top-coding” occurs (i.e., when the ceiling of a test is too low for many students) and suggested developing a set of standards for determining when a state has a top-coding problem.

Dr. Kane then addressed earlier points by Dr. Darling-Hammond and Dr. Rothstein concerning how to validate classroom observation measures of teacher effectiveness. He stated that if we are going to continue using classroom observation measures to assess teacher effectiveness, then we must validate them against student achievement outcomes. It makes no sense to think of “excellent teaching practice” in isolation, and separate from student outcomes. All such observation rubrics purport to measure practices which have been shown by research to be related to student achievement. Those claims should be regularly tested with updated data. In case the measures become corrupted or are implemented poorly or are not measuring the right practices in the first place, it is vital to continue to validate them against student achievement. Otherwise, the classroom observation is just one expert’s untested opinion of what excellent teaching is. There would be no means to compare one rubric against another.

While Dr. Kane maintained that gains on the current state tests should be one of the outcomes they are validated against, he added that they should be validated against the results of other assessments and student outcomes.

Participants debated whether current state tests are an adequate measure of student learning. Dr. Darling-Hammond pointed out that as long as teachers’ livelihoods depend on the specific outcomes of state tests, teachers will teach to the test. She expressed the opinion that state tests in the United States do a poor job of assessing students’ career and college readiness, in part because they depend heavily on multiple-choice questions. Dr. Kane agreed, saying that if there are better ways to measure certain skills or to measure different skills then they should be included on state tests. However, he argued that the lack of assessment of some skills on



the state tests was not an argument for eliminating student results on state tests completely when assessing teacher effectiveness. Ms. Weiss said many states are finding ways to use multiple measures of student learning. Dr. Darling-Hammond emphasized the need to keep in mind how narrowly focused current tests are relative to the goals set for students and relative to other countries' approaches.

Dr. Ladd asked whether other participants agreed with her that one cannot use VAMs to compare teachers across schools (as opposed to comparing teachers within the same school). Several participants agreed, noting that it is difficult if not impossible to compare effectiveness across schools, but other participants believe that the research supports the use of VAMs to compare teachers across schools.

In general, participants saw agreement on a number of facts but variation in their interpretations of the facts and the implications of research findings on specific policies. Dr. Kane suggested that IES use the advance summaries and the notes from this meeting to identify the subset of findings with which almost all of the researchers can agree.

## **Lunch**

Participants adjourned for lunch at 11:47 a.m., and the meeting resumed at 12:18 p.m.

## Practical Implications (Including Applications and Limitations)

**Andrew Ho, Ph.D., M.S., Lead Discussant**

**Brad Jupp, Facilitator**

Dr. Ho referred participants to his summary (see Appendix A) and focused on three areas that he believes hold significant promise for the future.

### Advancing from Symptoms to Diagnosis and Treatment

Dr. Ho compared determining whether a teacher is effective to evaluating a patient's physical health within the field of medicine. Doctors use multiple methods to assess symptoms, including observation, patient self-report, and diagnostic tests. Doctors consider the relative usefulness of each method, but they do not assign specific weights to each measure because the emphasis is on diagnosing and treating the problem. In using VAMs and other tests to determine teacher effectiveness, Dr. Ho stated that there is a tendency to first reduce different symptoms to a single summary statistic, "value-added," and then pay little attention to diagnosis or treatment. In medicine, making a diagnosis is complicated and each test only tells a portion of the story. Education takes a simplistic approach to teacher evaluation, when it should be taking into consideration how the various results should inform diagnosis and treatment. Furthermore, it is important to ask who the "doctors" are in the education system (that are making the diagnoses), who the treatment providers are, and what they need to know.

### Differences that Matter: The Statistical Properties of Accountability Indices

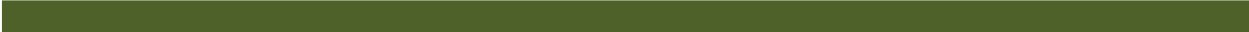
At the state level, users of VAMs and other measures of teacher effectiveness tend to put their measures into categories and then average the results for each teacher across the categories to create a single measure. Much more attention should be paid to the statistical qualities of these aggregate indices and how users at the state level and elsewhere are employing these indices to make decisions.

### Auditing for Unintended Consequences

Unintended consequences are rarely unanticipated—in fact, the consequences often are so obvious that they are better described as "intentional." In using VAMs to assess teacher effectiveness, a predictable consequence is inflation in student test scores. Dr. Ho strongly urged the use of embedded audit tests within student tests used for value-added, as well as additional research to improve test auditing methodology.

## Discussion

In light of recent advances in VAM research (discussed during the morning session), Mr. Jupp asked each participant how they would advise state and local leaders to "do the best and avoid the worst" with regard to using VAMs to assess teacher effectiveness.



Dr. Braun said that sometimes good teachers can score poorly on VAMs, so states need to ensure that local audits of results are included to provide checks and balances. Local audits, in turn, should be subject to review at higher levels, so that local leaders do not use them to protect teachers who do not deserve protections. Referring to examples described earlier in the meeting of cases in which good teachers received low VAM scores due to unusual aspects of their situations, Dr. Braun stated, “Hard cases make bad law.” He explained that extreme examples are not a reason to dismiss VAMs but rather argue for including local audit results.

Dr. Goldhaber focused on attribution of students to the correct teachers and exposure of students to teachers, noting that the existing data often do not support how users want to employ VAMs and SGMs. To support the credibility of high-stakes decisions made on the basis of VAMs and SGMs it is important to adequately capture which students were taught by which teachers, even if it doesn’t actually change individual teachers’ scores. Dr. Goldhaber’s second point was that different models produce different estimates depending on the types of students that are being taught. Those differences should be made transparent to decision-makers up front so that they can better understand and sign-off on the implications of choosing a particular model (rather than an alternative) before it is implemented. Dr. Goldhaber argued that such transparency would help stakeholders more readily accept the adopted model.

Dr. Pianta stated that he would advise leaders to “do something” rather than nothing. He stated that there is “signal” in the VAMs and hoped leaders would not be paralyzed by the research literature describing problems with VAMs. That said, he advised leaders to use VAMs as part of a decision-making system that includes other measures and provides safeguards, supports, structures, and mechanisms for second chances that enable leaders to make sensible and defensible decisions. Dr. Pianta further suggested that within their human resource systems for teachers, leaders should use a triage or gated model, in which, at various points or gates, teachers are linked to systems for improvement in an overall system that ties together evaluation and improvement.

Dr. Rockoff said auditing must be part of any high-stakes system that uses student test scores and that the auditing structure should be in place in advance. Second, he said that people have a natural tendency to categorize teachers based on effectiveness, and every evaluation system involves establishing cut-off points if consequences result from teachers’ scores on the system. Nevertheless, information gathered from teacher assessment should not be reduced simply to the cut-off scores and categories. Thirdly, Dr. Rockoff stated that it is important to communicate clearly to teachers what the particular measures tell us—for example, that it assesses the ability of teachers to raise student achievement on a specific test, or that it yields a comparison of performance relative to other teachers who teach similar students. Clear messaging around the evaluation measures will be important for getting teachers behind them. Finally, Dr. Rockoff stated that the timing of when results are released is important. One issue is that principals should receive teachers’ value-added scores only after they have

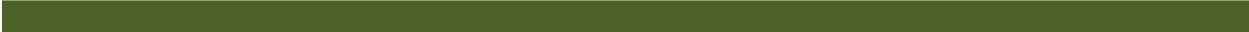
completed their other evaluations of each teacher. Otherwise, principals will game the system to prevent having their teachers with low VAMs be rated as “ineffective.” Another issue is that to be useful, VAMs need to be provided to teachers quickly, because if teachers receive their scores 4 months after the school year ended, they’re not going to be useful. However, getting scores faster costs money, Dr. Rockoff said.

Dr. Corcoran stated that he agreed with the comments of the preceding speakers. He advised leaders to use appropriate caution in devising teacher assessment systems, but he said that based on his experience in New York, his sense was that some states were doing a good job by recognizing uncertainty and not using single-point estimates to assess teachers. Dr. Corcoran expressed some concern that value-added teacher assessment systems and reforms such as Race to the Top are one-size-fits-all approaches that discourage experimentation. These standardized approaches at the state level provide transparency but handcuff schools and districts from trying new approaches. (He noted that charter schools, for example, do not use VAMs as much as public schools.) Dr. Corcoran echoed earlier comments calling for common sense and auditing with regard to high-stakes testing.

Dr. Kane advised leaders to continue using composite measures that include VAMs to evaluate teachers, but he would allow local discretion in handling teachers that fall short on the composite measure. Dr. Kane would change the default action (such that the default would be that a teacher who falls short on the composite score would not receive tenure, since this is the “tough decision” for principals to make), but he would allow principals to override the default decision with some extra effort on their part, such as writing to all the parents in the school why the default decision was being overridden. Second, Dr. Kane advised that teachers should have an opportunity to verify which students are on their teaching rosters to address the student-teacher attribution issue. Finally, he advised that efforts to improve the quality of student tests should focus first on English Language Arts (ELA). Currently, such tests involve multiple-choice questions related to reading comprehension, which most teachers stop teaching after third grade. Updated ELA tests should require some student writing.

Dr. Ladd explained that since she believes that VAMs only allow for comparison of teacher effectiveness *within* schools (as opposed to comparisons of teachers across different schools), VAMs are not a very useful tool for state or local leaders. At the school level, she recommended that principals use multiple measures of teacher effectiveness and that they should be allowed discretion in using these tools. Dr. Ladd emphasized that VAMs are not directly useful for improving student achievement or for ensuring equitable access to good teachers across schools within a district or state.

Dr. Betebenner said states should consider how to increase education efficacy—not necessarily achievement—in order to accelerate student velocity with regard to learning. Dr. Betebenner said that his experience as a data analyst had taught him that data can tell many stories, and results can be presented in many ways. Policymakers should craft the narrative



resulting from their data around the education policy they want to pursue. Policymakers should think about how policy travels across levels, from decisions at the state level, to districts, schools, and classrooms. He said VAMs can be applied across multiple policy levels to evaluate educator effectiveness, but getting leadership across policy levels to communicate consistently is a significant challenge of the effort to get educators to use data more effectively.

Dr. Darling-Hammond cautioned state and local policymakers that VAMs can provide distorted information that leads to dysfunctional consequences. States should ask themselves whether their state tests have a low ceiling and a high floor because VAMs based on such tests may not provide accurate results for some teachers, such as those teaching gifted students, ELLs, and special education students with high needs. Dr. Darling-Hammond urged state and local leaders never to base decisions on a single VAM. Rather she urged them to consider a “basket of evidence” that provides varied information. Evaluation of student achievement should go beyond multiple-choice tests and include, for example, student essays written at the beginning and end of the school year. Policymakers should focus not just on the ranking of teachers — but on how to facilitate teacher improvement. Because personnel decisions will be involved, policymakers must make every effort to ensure that any evidence used to evaluate teachers leads to accurate decisions. State and local leaders should be concerned about the effect of any teacher assessment system on recruitment and retention, since there is the possibility of losing good teachers if they feel that they are being evaluated by an erratic and inaccurate system.

Dr. Friedman echoed earlier speakers in urging state and local policymakers to use VAMs cautiously. However, he also pointed out that the use of analytic methods is the hallmark of modern management in almost every industry. He encouraged leaders to use data intelligently—not just for evaluating individual teachers but for understanding how the educational system is working. For example, VAMs could be helpful in better assessing how well teacher development programs are working. Such applications would help individuals become more familiar with VAMs in a context in which less is at stake, there is less incentive to cheat, and there is less concern about equity. While an inaccurate teacher assessment system could drive teachers away, an effective one could draw teachers into the profession because it rewards excellence at the individual or program level.

Dr. Rothstein disagreed with Dr. Pianta; he said it is not clear that doing something is better than doing nothing in the area of teacher assessment. A poorly designed system could easily be worse than the status quo. Dr. Rothstein expressed his view that the emphasis of teacher assessment policies should be to make teachers better, not to fire them. An effective evaluation system should include investment in observation and training as well as a commitment to provide feedback and improvement opportunities following evaluation. This is expensive, however, and may not be feasible given current fiscal constraints. In that case, it would be better to do nothing than to rush ahead with a bargain-basement system. Dr. Rothstein agreed with Dr. Darling-Hammond in recommending that human judgment be

retained as part of any assessment process, with the supervisor being allowed to weigh the overall “basket of evidence” as they see fit, rather than assigning required weights to particular components. However, Dr. Rothstein cautioned that because those making the assessments are not highly numerate, they will tend to treat value-added scores as more “scientific” than they actually are. Dr. Rothstein also suggested that leaders think hard about dimensions of student learning that are not well measured now, such as arts, citizenship, or non-cognitive skills, and that they figure out ways to monitor these dimensions for evidence that they are being neglected under a new evaluation system.

Because it is not clear what a better teacher assessment system should look like, Dr. Rothstein recommended that both state and federal leaders institute many pilot programs in order to learn from them before rolling out large new assessment systems. Policymakers should have realistic expectations and recognize that successful teacher assessment systems will yield only small improvements in student outcomes. Finally Dr. Rothstein said that he would advise state policymakers to advocate in Washington, DC, for better federal policies—as in his view, recent federal policies surrounding teacher evaluation have done more harm than good.

Dr. Ho recommends against the practice of ranking teachers and schools by averaging across “proficiency” statistics from multiple measures. Averaging over “proficiency” categories throws away information and decreases precision. In general, averaging over multiple measures helps ranking but not diagnosis or treatment. Dr. Ho used the analogy of the medical profession, where blood pressure and pulse are not averaged as much as considered each in turn to determine diagnosis and subsequent treatment.

Mr. Jupp posed some additional questions for consideration. First, he asked Dr. Ladd to elaborate on the advice she would give to school leaders about teacher assessment systems.

Dr. Ladd questions the usefulness of VAMs even at the school level. Dr. Ladd agrees that VAMs do contain information about teachers and that principals should have access to them, but they should understand how to interpret VAMs and recognize that VAMs are just one piece of data on which to make decisions about how to use resources effectively. She notes as well that because VAMs only compare teachers within a school, their use could provide an incentive for principals to encourage the weaker teachers within their school to leave and move to other schools.

Dr. Goldhaber commented that not all the researchers participating in the meeting agreed that fixed-effect VAMs work only for within school comparisons.

Mr. Jupp’s second question, addressed the entire group: ***What can we learn from VAM research that includes additional dimensions of teacher effectiveness in value-added measurement that we are currently missing?***

---

Dr. Rockoff said that New York allows local districts to add a locally developed measure of teacher effectiveness and give it as much weight in the teacher effectiveness score as the growth or VAM from the state. However, he acknowledged that many small districts do not have psychometricians and do not have the resources to develop a reliable and valid local measure, especially if they have to develop a different test for each different local course.

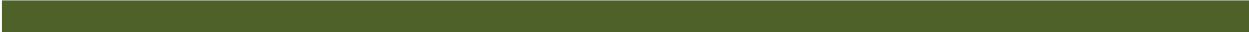
Dr. Corcoran noted that state standards identify numerous dimensions of learning that are not captured through current state tests, such as the ability to engage in a debate in the class setting about a current issue. Some of these learning objectives may not be testable, but the current state standards are a good place to start looking for additional dimensions to include in value-added measurement. Turning to a different point, Dr. Corcoran stated that any personnel decision rule that is based on multiple measures and takes into account the uncertainty and variation across these measures will necessarily have to be a very conservative decision rule that only has strong consequences (e.g., dismissal) for those at the extremes of the distribution. The difficulty that Dr. Corcoran saw with that approach is that such a system would not give useful information to the vast majority of teachers in the middle of the distribution; it would only tell them that they are in the middle. In terms of research, he stated that he thought it should focus on understanding the tails of the distribution, since it is there that the high-consequence decisions will occur.

Speaking to Dr. Corcoran's point about unassessed student learning outcomes, Dr. Pianta would like to see more assessment of student gains in non-cognitive skills, such as social skills. On a different point, Dr. Pianta acknowledged that it is important to include multiple measures in an assessment system, but he advocated for keeping the system simple and limiting the number of indicators to three or four because that is the maximum number of indicators that can be useful. He added that at least one indicator should capture students' learning of social skills.

Dr. Darling-Hammond stressed the importance of measuring the right outcomes of student learning and doing so richly. She noted that other countries with top-rated academic performance use testing that is more open-ended, varied, and includes assessment of students' work on collaborative projects. However, these sorts of tests must be developed in collaboration with the students' teachers, but with high-stakes testing, teachers cannot be part of the planning, development, or grading of the tests.

Dr. Rothstein added that many important dimensions of student outcomes cannot be measured with VAMs. As an example, Dr. Rothstein stated his opinion that much of a teacher's job is to serve as a social worker to students, helping them learn non-cognitive skills such as how to resolve disputes. It is hard to imagine a VAM that would assess this component of teachers' jobs, he stated. If we put a lot of weight on teacher assessments that emphasize quantitative value-added type measures, then we are effectively saying that the rest does not matter, and we will fail to assess much of what teachers do.





Dr. Kane stated that during the next couple of years principals will begin conducting more classroom observations of teachers, and he believes that will create a larger storm than value-added measurement, because it will affect the daily life of principals and teachers in classrooms much more profoundly. As a result, he believes that it is vital to come up with better approaches to conducting classroom observations. Dr. Kane believes that the use of digital video for classroom observations holds great promise for several reasons:

1. It makes it easier to have someone from outside the school that doesn't have a personal relationship with the teacher rate the teaching.
2. It provides direct feedback to the teacher about their teaching that they can't get from any other method because they can watch the video themselves.
3. It makes it cheaper and allows principals to time-shift their work because they can watch the video at any time.
4. Despite the novelty, having a camera in class may be less distracting for the teacher and students than having the principal standing in the back of the classroom frowning and taking notes.

Dr. Kane stated that the biggest challenge in teacher assessment will be in conducting the classroom observations well, and the biggest breakthrough there will be in digital video technology.

## Directions for Future Research

### **John Q. Easton, Ph.D., Facilitator**

Dr. Easton invited participants to comment on areas for future research beyond VAMs.

Dr. Braun commented that the Common Core Standards have been adopted by 46 states, and yet they had not entered into the group's conversation about VAMs. Teachers and principals face new, more challenging standards; new student tests; and being held accountable in a new way (based on value-added). This could be a disaster, at least in the short term. Part of the research agenda should be how researchers should monitor and help states, districts, and schools understand what is happening and help them adjust to it, because some districts and schools do not have adequate capacity to adjust to these sweeping changes. Second, Dr. Braun cautioned that we should moderate our expectations for the new student assessments that will be coming online from the PARCC and Smarter Balanced consortia, despite how good they hope to be. The plans of both consortia are fairly limited with regard to support of teacher assessments. Given the number of challenges facing these consortia, Dr. Braun felt that it was unrealistic to expect them to do more than they were planning in regard to supporting teacher assessment.

Dr. Ho stated that he thinks the roll-out of the new teacher assessment system should include guidelines for creating indices of teacher effectiveness, as opposed to just reporting individual metrics, because it's inevitable that these metrics will be combined with other measures. States might also want to have some metrics that are used for auditing and some metrics that are used for decision-making. Turning to another point, Dr. Ho agreed with Dr. Pianta that the number of dimensions included in a teacher assessment system should not be excessive. On the other hand, drawing on an analogy with medical practice, he urged policymakers not to simply average a teacher's scores on the various dimensions to come up with a composite score. Rather, like a doctor making a diagnosis, Dr. Ho urged better understanding on the part of the decision-makers of the different information the various measures convey.

Dr. Pianta recommended research on the roll-out of the new teacher assessment systems so that we can understand how well these systems are being implemented, what the implementation processes are, what the auditing structures are, and so on. He also recommended studying how the behavior of these systems varies with design variations, since there will be a great deal of such variation. Dr. Pianta expressed concern that the field does not have a robust basic science agenda around teacher effectiveness. He asked whether we are building and funding such an agenda today so that we can be having a different set of conversations about teacher effectiveness in 5-10 years. He believes that this is a hole in the federal research investment and contributes to the weakness of our measures of teacher effectiveness.


Dr. Kane stated his opinion that the introduction of the Common Core Standards is a huge opportunity for research and that IES should fund the development of classroom observation tools and student surveys that are more aligned with the kinds of instruction relevant to the

Common Core Standards. Turning to another point, Dr. Kane agreed with those who suggested that there should be better measures of desired non-cognitive student outcomes, such as student persistence. He stated that they tried to measure some of these non-cognitive outcomes in the Gates Foundation's MET Project, but that they were not very successful. So if the goal is to reward teachers on the basis of their students' progress toward these desired non-cognitive outcomes, then researchers need to be developing better ways of measuring these constructs. Dr. Kane mentioned research that the Gates Foundation funded on measuring student persistence and engagement by measuring galvanic skin response. He expressed his opinion that biometric approaches to measuring non-cognitive aspects of student learning might overcome some of the problems associated with current survey-based approaches to measuring non-cognitive learning and attitudes.

Dr. Goldhaber expressed the opinion that improving teaching may depend more on the politics and culture of schools than it does on the amount of information contained in VAMs. He stated that how one views value-added or any other type of measure depends on whether one thinks the current system of teacher assessment is working well and what the alternatives are. He alluded to the Widget report (<http://widgeteffect.org/>), which indicates that currently, across a wide range of school districts, almost all teachers are rated at the very top of whatever teacher assessment scale is being used. Dr. Goldhaber stated his opinion that there should be more research on what information related to teacher performance is provided to principals and whether the provision of that information, by itself, changes the conversation between principals and teachers about performance. If the question of teacher improvement is more about the politics and culture of schools rather than what information we do or don't have about teachers, then that implies different solutions.

Dr. Ladd raised the question of what the purpose of teacher assessment systems is. Is it to evaluate teachers (for bonuses, promotions, dismissal, etc.) or is it to try to improve teachers? Dr. Goldhaber pointed out that this is important for determining the theory of action of teacher assessment systems, and Dr. Braun pointed out that one must know the goal of the system prior to determining its theory of action.

Dr. Braun continued that the theory of action for teacher assessment systems needs to be based on what makes a good school or an improving school. This in turn leads to the question of how the teacher assessment system supports the capacity of schools to improve, which, he stated, is something the states have not explicitly addressed. Dr. Braun stated that the school improvement literature contains the idea that building collegiality and shared responsibility among teachers leads to school improvement. He said there should be research on whether collegiality and shared responsibility do contribute to school improvement. If so, how do we align the systems of teacher assessment and accountability so that they support school and teacher improvement?

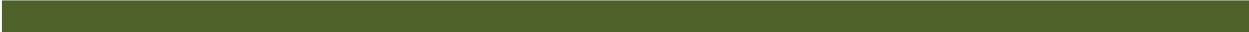


Dr. Friedman stated that he did not see why the use of VAMs for individual teacher accountability was at all in conflict with using it for broader system improvement. He stated that measuring what one can and then accounting for what one can't measure is something that every organization does. Once an organization has the information, it can implement a wide variety of strategies: it can try to hire good teachers; it can try to improve the teachers it has; it can release some teachers. It all starts with good measurement, he stated. Turning to another point, Dr. Friedman stated we should be encouraging schools to run many pilots of teacher assessment systems and to run them in a way that allows us to draw sound conclusions about the pilot's impact. He stated that much of what we will learn over the next 5-10 years will concern how these different state and local efforts work and don't work. We will learn that some things we thought were important, aren't important and vice versa. However, these pilots need to be implemented in a way that is analyzable rather than in a haphazard way.

Dr. Rothstein suggested five possible research topics. First, what information is useful in a formative evaluation system? What can value-added scores contribute? Second, how should VAMs be used? Dr. Rothstein stated his belief that much of the discussion around value-added has been predicated on the assumption that once we have these measures we'll know what to do with them. He stated that figuring out what to do with them should particularly be a priority for IES. Third, how does the addition of high stakes change VAMs? How much are they distorted? Fourth, how do we compare teachers who are teaching in different schools or different student contexts? Researchers tend to focus on within-school or within-context comparisons, because the other comparisons are methodologically difficult. However, Dr. Rothstein stated that he thinks it is implausible that policy debates will be limited to within-school comparisons, so researchers should work on developing responsible ways of comparing teachers across advantaged and disadvantaged (or, alternatively, high-performing and low-performing) schools. Finally, Dr. Rothstein agreed with Dr. Friedman in saying that we need to learn more from the natural experimentation that's happening in the states around teacher assessment.

Dr. Rockoff stated that an important prerequisite for learning from the state experimentation will be that someone collects information on what the states are doing and makes it available in a standardized format that researchers can study. He stated that this could be an important role for the federal government to play.

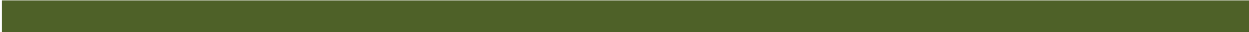
Dr. Pianta talked about the need to create some sort of rapid funding mechanism for researchers. State-level developments in teacher assessment are happening very rapidly, but it can take 2 years for a researcher to receive funding for a research proposal, at which point the state variation has already occurred. Dr. Ho stated that the National Science Foundation (NSF) has a rapid funding mechanism. Dr. Ladd commented that the National Institutes of Health have a similar arrangement. Dr. Ho stated that the turnaround time is 1 month.



Dr. Darling-Hammond stated that we need more research on how teacher assessment systems that are built on multiple measures actually work in practice. Research should look at the defensibility and accuracy of the teacher evaluation scores from multiple lenses. Research should also look at the consequences for the labor force and for how teachers react. Another research topic is the effects of these teacher assessment systems on student learning, depending on the mix of measures used. Another question is how these teacher assessment systems affect collegial culture and professional capacity. It is also important to study how the teacher evaluation scores are implemented to trigger personnel decisions in school districts. Another issue is to look at the implementation of teacher assessment tools, such as classroom observations, student surveys, etc. Otherwise, we could end up with a lot of really good assessment tools but not know how to implement them in schools and districts. Another research topic is to look at how the measures of student learning affect what teachers teach.

Dr. Ladd commented that the introduction of charter schools and vouchers has led to much more movement of students between traditional public schools and charter or private schools. Another research topic is the implications of that movement for teacher evaluation systems, especially when students are moving mid-year. Dr. Easton commented that online schools were also part of this issue.

Dr. Kane commented on his observation that the two testing consortia (PARCC and Smarter Balanced) have been conducting their work for the most part independently of the research community on teacher effectiveness. Dr. Kane felt that it was important that the student tests be designed in ways that don't interfere with their later use for teacher assessment and feedback to individual teachers. Dr. Darling-Hammond stated that she is working with one of the consortia, and that the teacher evaluation question is very much on the table. Dr. Goldhaber echoed Dr. Kane, saying that a criticism one often hears of value-added is that the student test involved was not designed for value-added measurement. It would be a shame if teacher assessment is not part of the conversation while these new student tests are being designed. Dr. Braun stated that he is wary of having the teacher accountability "tail" wag the student assessment "dog." Dr. Braun expressed his opinion that the point of the new student assessments was to build the best tests possible to reliably assess student learning outcomes vis-à-vis the Common Core Standards. If the tests do that well then they will form a better basis for teacher assessment because they will be broader and won't depend so strongly on multiple-choice testing. Nevertheless, Dr. Braun commented that the PARCC consortium, which he consults with, is struggling with how to support the inevitable use of their test results in teacher evaluation. There are necessarily tradeoffs involved in designing the student tests to maximize their suitability for teacher assessment and designing them to support the many other positive outcomes desired for the test. Dr. Rothstein commented that a single assessment cannot be used for all purposes. He alluded to Derek Neal's argument that a test to measure a student's achievement level needs to be designed differently than one to assess a student's learning growth, and that a test for (teacher or system) accountability needs to be designed differently still. Dr. Rothstein stated that it all comes back to the theory of action of what the test is intended to do.



Dr. Betebenner stated that another topic for the research agenda is determining how to design the interfaces that teachers, principals, and parents use to access the teacher assessment data so that each of those groups has the user experience that we want them to have with this information. Dr. Betebenner voiced the opinion that it was important to “start at the end” and consider in detail the conversations that we want teachers, principals, and parents to have when they see the teacher evaluation scores. Dr. Betebenner expressed the opinion that many stakeholders lack an in-depth understanding of what the teacher assessment numbers mean. He commented that for the educational system to improve, stakeholders will need to become better at using evidence of all kinds ranging from value-added to classroom observations. So the research question is as follows: How do we build better interfaces to help people make better decisions based on evidence, knowing that we’re going to be feeding them more and richer sources of evidence?

Dr. Ladd raised the question of why we are spending so much time talking about teacher evaluation when her own research indicates that the quality of the principal is an important determinant of teacher effectiveness.

Dr. Corcoran commented that many of the questions that were raised during the day—such as implementation issues, and the effect of teacher evaluation systems on teachers’ behavior and collegiality—are not amenable to a randomized controlled trial methodology. So where do these sorts of research projects fit within the IES research framework?

Dr. Braun pointed out that value-added approaches create a distribution of teacher scores. Yet there are no guidelines for how decision-makers should meaningfully set the cut-off scores or standards that classify teachers into different categories of effectiveness (e.g., “ineffective,” “highly effective”). He stated that there are ways to collect data to set those thresholds in reasonable ways and to provide a rationale for them.

Dr. Ho stated that setting the cut-off scores or standards should be secondary to understanding what the measure itself means. Therefore, he agreed with Dr. Goldhaber in saying that there should be more research on how people use such scales and interpret them. The better we understand the scale, the better we will be able to say where “proficient” or “passable” falls on that scale.

Dr. Easton brought the discussion on research priorities to a close by summarizing the discussion into six categories:

1. There was much discussion of how the Common Core Standards and their associated assessments will have a huge impact on teaching and learning, and researchers are going to have to stay on top of those developments.
2. There were many questions about how formative assessments would be used to improve teaching and learning.
3. There are many, many experiments in educational reform going on in the states and localities. Will researchers be able to seize that opportunity? Do we have the data

systems to collect the data we need? As practitioners are beginning to think about these experiments, can we get in there early enough to help them be more systematic so that they can be more rigorous in their evaluations?

4. Many discussants focused on the organization of schools, the components of successful schools, successful school improvement, and the role of principals as important topics of research.
5. Another issue is the rapidly changing landscape of school organization and configuration: charter schools, online learning, hybrid learning. If we define “school” too rigidly we will exclude more and more students from what we study.
6. When multiple types of teacher assessment data are aggregated, not only do they need to be communicated effectively to stakeholders but also the process of standard setting—the cut scores—that occurs behind the scenes needs to be fully thought through and approached very carefully.

Dr. Ladd added a seventh category:

7. It would be helpful for researchers to spell out the various theories of action regarding how teacher assessment systems might improve student learning and then detail what we know regarding the steps in those various theories of action.

## Closing Remarks and Adjournment

### John Q. Easton, Ph.D., and Joanne Weiss

Ms. Weiss and Dr. Easton thanked participants for sharing their time, expertise, and knowledge. Dr. Easton concluded that a summary of the meeting, along with the summaries provided by the participants in advance of the meeting, would be circulated for review by the participants, and then published by IES. Dr. Easton adjourned the meeting at 2:48 p.m.

---

<sup>1</sup> Rothstein, J. (2009). Student Sorting and Bias in Value-Added Estimation: Selection on Observables and Unobservables. *Education Finance and Policy*, 4(4): 537–571; and Rothstein, J. (2010). Teacher Quality in Educational Production: Tracking, Decay, and Student Achievement. *Quarterly Journal of Economics*, 125(1): 175–214.

<sup>2</sup> Chetty, R., Friedman, J., & Rockoff, J. (December, 2011). The Long-term Impacts of Teachers: Teacher Value-added and Student Outcomes in Adulthood. National Bureau of Economic Research Working Paper #17699.

<sup>3</sup> After the meeting, Dr. Friedman requested that the following two paragraphs be added: the first about the relationship between Rothstein (2010) and Chetty et al. (2011), and the second about the relationship between Darling-Hammond et al. (2012) and Chetty et al. (2011).

The findings of Rothstein (2010) and Chetty et al. (2011) do not conflict in the literature; rather, Chetty et al. (2011) reconciles the findings of previous work on value-added, including Kane and Staiger (2008) and Rothstein (2010). Rothstein reports two important results, both of which Chetty et al. replicate in their data. First, there is significant grouping of students into classrooms based on twice-lagged scores (lagged gains), even conditional on once-lagged scores (Rothstein 2010, Table 4). Second, this grouping on lagged gains generates minimal bias in VA estimates: controlling for twice-lagged scores does not have a significant effect on VA estimates (Rothstein 2010, Table 6; Kane and Staiger 2008, Table 6). The results from Table 2 and Figure 2 in Chetty et al. are consistent with Rothstein’s conclusions. Therefore, the literature is in agreement that VA measures do not suffer from bias in this way.

The findings in Chetty et al. (2011) directly address the three main critiques of VA measures raised in a recent *Phi Delta Kappan* article by Stanford education professor Linda Darling-Hammond and her colleagues (Darling-Hammond, L., Anrein-Bardsley, A., Haertel, E., & Rothstein, J., Evaluating Teacher Evaluation. *Phi Delta Kappan*, March, 2012). The article shows directly, using quasi-experimental tests, that standard VA measures are not biased by the students assigned to each teacher. Hence, value-added metrics successfully disentangle teachers’ impacts from the many other influences on student progress. It also shows that although VA measures fluctuate across years, they are sufficiently stable that selecting teachers even based on a few years of data would have substantial impacts on student outcomes such as earnings.