Darling-Hammond, L. (2011) Evaluating teacher Evaluation: We Know About Value-Added Models and Other Methods. Phi Delta Kappan

Evaluating Teacher Evaluation: What We Know about Value-Added Models and Other Methods[1]

Linda Darling-Hammond, Stanford University
Audrey Beardsley, Arizona State University
Edward Haertel, Stanford University
Jesse Rothstein, University of California at Berkeley

There is a widespread consensus among practitioners, researchers, and policy makers that current teacher evaluation systems in most school districts do little to help teachers improve or to support personnel decision making. For this reason, new approaches to teacher evaluation are being developed and tested.

There is also a growing consensus that evidence of teachers' contributions to student learning should be a component of teacher evaluation systems, along with evidence about the quality of teachers' practice. "Value Added Models" (VAMs), designed to evaluate student test score gains from one year to the next are often promoted as tools to accomplish this goal. Policy makers can benefit from research about what these models can and cannot do, as well as from research about the effects of other approaches to teacher evaluation. This brief addresses both of these important concerns.

Research on Value-Added Models of Teacher "Effectiveness"

Researchers have developed value-added methods (VAM) as a means to look at gains in student achievement by using statistical methods that allow them to measure changes in student scores over time, while taking into account student characteristics and other factors often found to influence achievement. In large-scale studies, these methods have proved valuable for looking at a range of factors affecting achievement and measuring the effects of programs or interventions.

The use of VAMs for individual teacher evaluation assumes that measured achievement gains for a specific teacher's students reflect that teacher's "effectiveness." This attribution, however, assumes that student learning is measured well by a given test, is influenced by the teacher alone, and is independent from the growth of classmates and other aspects of the classroom context. None of these assumptions is well supported by current evidence.

Most importantly, research reveals that a student's achievement and measured gains are influenced by much more than any individual teacher. Others factors include:

- School factors such as class sizes, curriculum materials, instructional time, availability of specialists and tutors, and resources for learning (books, computers, science labs, and more)
- Home and community supports or challenges

- Individual student needs and abilities, health, and attendance
- Peer culture and achievement
- Prior teachers and schooling, as well as other current teachers
- Differential summer learning loss, which especially affects low-income children
- The specific tests used, which emphasize some kinds of learning and not others, and which rarely measure achievement that is well above or below grade level.

Most of these factors are not actually measured in value-added models, which rely on statistical controls for past achievement to parse out the small portion of student gains that is actually due to other factors. Within this component, the teacher's effort and skill. As a consequence, researchers have documented a number of problems with VAM models as accurate measures of teachers' effectiveness.

1. Value-added models of teacher effectiveness are highly unstable.

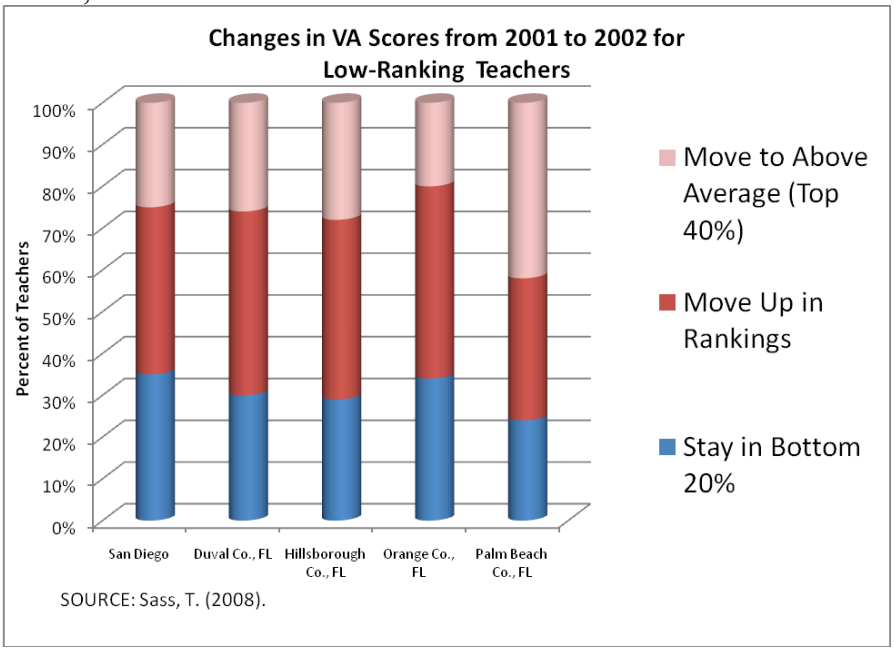
Researchers have found that teachers' effectiveness ratings differ substantially from class to class and from year to year, as well as from one statistical model to the next, as Table 1 shows (Newton et al., 2010).

Table 1: Percent of Teachers Whose Effectiveness Rankings Change

	By 1 or more Deciles	By 2 or more Deciles	By 3 or more Deciles
Across models [a]	56-80%	12-33%	0-14%
Across courses [b]	85-100%	54-92%	39-54%
Across years [b]	74-93%	45-63%	19-41%

Note: [a] Depending on pair of models compared. [b] Depending on the model used.
Source: Newton, Darling-Hammond, Haertel, and Thomas (2010).

A study examining data from five separate school districts found, for example, that of teachers who scored in the bottom 20% of rankings in one year, only 20-30% had similar ratings the next year, while 25 – 45% of these teachers moved to the top part of the distribution, scoring well above average. (See Figure 1.) The same was true for those who scored at the top of the distribution in one year: A small minority stayed in the same rating band the following year, while most scores moved to other parts of the distribution.



Teachers' measured effectiveness varies significantly when different statistical methods are used. (Briggs & Domingue, 2011) For example, when researchers used a different model to recalculate the value-added scores for teachers that were published in the Los Angeles *Times* in 2011, they found that from 40 to 55 percent of them would get noticeably different scores using an alternative VA model that accounted for student assignments in a different way. (Lockwood, et. al., 2007)

Teachers' value-added scores also differ significantly when different tests are used, even when these are within the same content area.[2] (Gates Foundation, 2010) For example:

- In a study using two tests measuring basic skills and higher order skills, 20%-30% of teachers who ranked in the top quartile in terms of their impacts on state tests ranked in the bottom half of impacts on more conceptually demanding tests (and vice versa). (Lockwood et. al., 2007)

- Teachers' estimated effectiveness is very different for "Procedures" and "Problem Solving" subscales of the same math test. (Corcoran et. al., 2011)

- Teacher effects on high-stakes tests are not highly related to their effects on low stakes tests, and dissipate more quickly. (Briggs & Domingue, 2011)

This raises concerns both about measurement error and, when teacher evaluation results are tied to student test scores, about the effects of emphasizing "teaching to the test" at the expense of other kinds of learning, especially given the narrowness of most tests currently used in the United States.

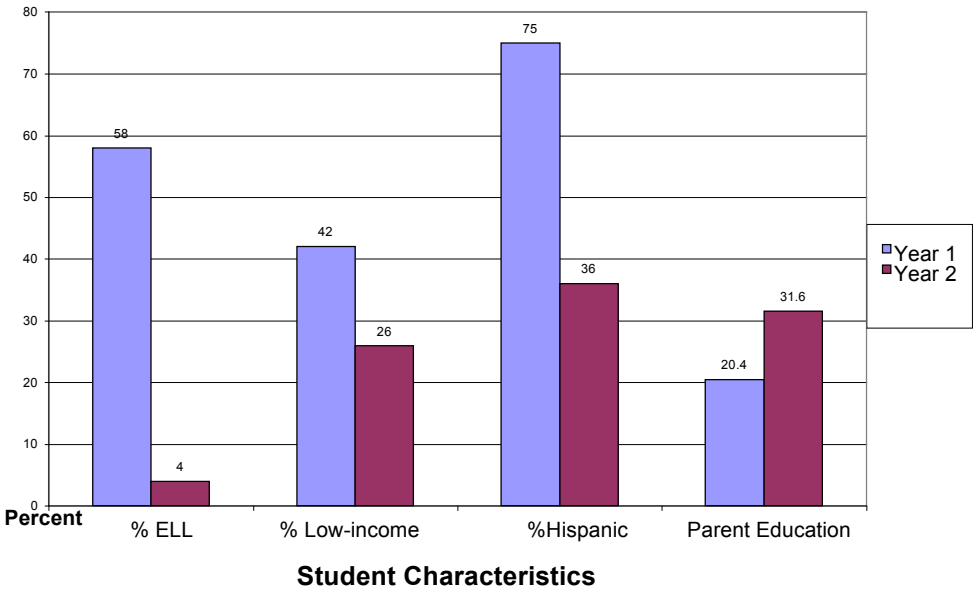
2. Teachers' value-added ratings are significantly affected by differences in the students who are assigned to them

VAMs are designed to identify teachers' effects when students are assigned to teachers randomly. However, students are not randomly assigned to teachers – and statistical models cannot fully adjust for the fact that some teachers will have a disproportionate number of students who have greater challenges (students with poor attendance, who are homeless, who have severe problems at home, etc.) and those whose scores on traditional tests may not accurately reflect their learning (e.g. those who have special education needs or who are new English language learners). These factors can create both misestimates of teachers' effectiveness and disincentives for teachers to teach the students who have the greatest needs.

Even when the model includes controls for prior achievement and student demographic variables, teachers are advantaged or disadvantaged based on the students they teach. Several studies have shown this by conducting tests that look at a teacher's "effects" on their students *prior* test scores. Logically, for example, 5th grade teachers can't influence their teachers' 3rd grade test scores. So a VAM that identifies teachers' true effects should show *no* effect of 5th grade teachers on their students' 3rd grade test scores two years earlier. But studies that have looked at this have shown large "effects" – which indicates that the VAMs wrongly attribute to teachers other influences on student performance that are present when the teachers have no contact with the students. (Tood & Wolpin, 2003)

One study that found considerable instability in teachers' value-added scores from class to class and year to year examined changes in student characteristics associated with the changes in teacher ratings. After controlling for prior test scores of students *and* student characteristics, the study still found significant correlations between teachers' ratings and their students' race/ethnicity, income, language background, and parent education. Figure 2 illustrates this finding for an experienced English teacher in the study whose rating went from the very lowest category in one year to the very highest category the next year (a jump from the 1st to the 10th decile). In the second year, this teacher had many fewer English learners, Hispanic students, and low-income students, and more students with well-educated parents, than in the first year.

Figure 2 - Student Characteristics in Years 1 and 2 for a Teacher Whose Ranking Changed from the 1st to the 10th Decile



This variability raises concerns that use of such ratings for evaluating teachers could create disincentives for teachers to serve high-need students. This could inadvertently reinforce current inequalities, as teachers with options would be well-advised to avoid classrooms or schools serving such students, or to seek to prevent such students from being placed in their classes.

3. Value-added ratings cannot disentangle the many influences on student progress

It is impossible to fully separate out the influences of students' other teachers, as well as school conditions, on their reported learning. No single teacher accounts for all of a student's learning. Prior teachers have lasting effects, for good or ill, on students' later learning, and current teachers also interact to produce students' knowledge and skills. (Carrell & West, 2010) For example, a student's math progress in 4th grade may be importantly influenced by how well she learned the 3rd grade material the previous year; the essay writing a high school student learns through his history teacher may be credited to his English teacher, even if she assigns no writing; and the math he learns in his physics class may be credited to his math teacher. Specific skills and topics taught in one year may not be tested until later, if at all. Some students receive tutoring, as well as help from well-educated parents. A teacher who works in a well-resourced school with specialist supports may appear to be more effective than one whose students don't receive these supports.

It is not clear that "teacher effectiveness" is a stable enough construct that it could be uniquely identified even under ideal conditions (for example, with random assignment of teachers to schools and students to teachers, and with some means of controlling differences in out-of-school impacts). Some teachers may be effective at some forms of instruction or in some portions of the curriculum and less effective in others. If so, their rated effectiveness will depend on whether the student tests used for estimation of the VAM emphasize the skills and topics for which the teacher is relatively more or relatively less effective. As noted above, some teachers would be rated as effective if the test emphasized math procedures but as ineffective if it emphasized math problem solving, or as ineffective when a traditional multiple-choice standardized test is used but as effective when a more conceptually demanding test is used.

Other research indicates that teachers whose students do best on end-of-year tests are not always effective at promoting longer-run achievement for their students. Thus, VAM-style measures may be importantly influenced by the amount of emphasis that the teacher devotes to short-run test preparation. One study even found that teachers who most raised end-of-course grades were, on average, *less* effective than others at preparing the students for the next year's course. (Braun, 2005)

As Henry Braun, then at ETS, noted,

> It is always possible to produce estimates of what the model designates as teacher effects. These estimates, however, capture the contributions of a number of factors, those due to teachers being only one of them. So treating estimated teacher effects as accurate indicators of teacher effectiveness is problematic. (Braun, 2005)

Initial research on the use of value-added methods to dismiss some teachers and award bonuses to others shows that value-added ratings often do not agree with the ratings teachers receive from skilled observers, and are influenced by all of the factors described above.

For example, among a number of teachers dismissed in Houston as a result of their Education Value-Added Assessment System (EVAAS) scores, one teacher, a ten-year veteran, had been voted "Teacher of the Month" and "Teacher of the Year" and was rated each year as "exceeding expectations" by her supervisor. (Amrein-Beardsley & Collins, forthcoming) She showed positive VA scores on 8/16 of tests over four years (50% of the total observations), with wide fluctuations from year to year and both across and within subjects. (See table below.) It is worth noting that this teacher's lower value-added in grade 4, when English learners are mainstreamed, in Houston, was a pattern for many of the other teachers for whom EVAAS data were analyzed as well.

EVAAS Scores (Teacher A)	2006-2007 Grade 5	2007-2008 Grade 4	2008-2009 Grade 3	2009-2010 Grade 3	2010-2011 Grade 5
Math	-2.03	+0.68*	+0.16*	+3.46	n/a
Reading	-1.15	-0.96*	+2.03	+1.81	n/a
Language Arts	+1.12	-0.49*	-1.77	-0.20*	n/a

Science	+2.37	-3.45	n/a	n/a	n/a
Social Studies	+0.91*	-2.39	n/a	n/a	n/a
ASPIRE Bonus	$3,400	$700	$3,700	$0	n/a

*Notes: (1) Scores with asterisks (*) signify that the scores are not detectably different from the reference gain scores of other teachers across HISD within one standard error; however, the scores are still reported to both the teachers and their supervisors as they are here.

The wide variability shown in this teacher's ratings from year to year, like that documented in many other studies, was not unusual for teachers in Houston across this analysis, regardless of whether the teacher was terminated. Teachers reported that they could not identify a relationship between their instructional practices and their ratings on value-added, which appear unpredictable. As one teacher noted:

> I do what I do every year. I teach the way I teach every year. [My] first year got me pats on the back. [My] second year got me kicked in the backside. And for year three my scores were off the charts. I got a huge bonus, and now I am in the top quartile of all the English teachers. What did I do differently? I have no clue.[3] (Amerin-Beardsley & Collins, forthcoming)

Another teacher classified her past three years as "bonus, bonus, disaster." And another noted:

> We had an 8th grade teacher, a very good teacher, the "real science guy," [who was a] very good teacher…[but] every year he showed low EVAAS growth. My principal flipped him with the 6th grade science teacher who was getting the highest EVAAS scores on campus. Huge EVAAS scores. [And] now the 6th grade teacher [is showing] no growth, but the 8th grade teacher who was sent down is getting the biggest bonuses on campus.

This example of two teachers whose value-added ratings flip-flopped when they exchanged assignments is an example of a phenomenon found in other studies which document a larger association between the class taught and value-added ratings than the individual teacher effect, itself. The notion that there is a stable "teacher effect" that is a function of the teacher's teaching ability or effectiveness is called into question if the specific class or grade-level assignment is a stronger predictor of the value-added rating than the teacher.

Another Houston teacher, also consistently rated as "exceeding expectations" or "proficient" by her supervisor, and also receiving positive VA scores about 50% of the time, had a noticeable drop in her value-added ratings when she was assigned to teach a large number of English Language Learners who were transitioned into her classroom. Overall, the study found that, in this system,

- Teachers teaching in grades in which English Language Learners (ELLs) are transitioned into mainstreamed classrooms are the least likely to show "added value."

- Teachers teaching larger numbers of special education students in mainstreamed classrooms are also found to have lower "value-added" scores, on average.
- Teachers teaching gifted students show little value-added because their students are already near the top of the test score range.
- Ratings change considerably when teachers change grade levels, often from "ineffective" to "effective" and vice versa.

These kinds of comments from teachers were typical:

> Every year I have the highest test scores, [and] I have fellow teachers that come up to me when they get their bonuses…One recently came up to me [and] literally cried, 'I'm so sorry.'… I'm like, 'Don't be sorry. It's not your fault.' Here I am…with the highest test scores and I'm getting $0 in bonuses. It makes no sense year-to-year how this works. You know, I don't know what to do. I don't know how to get higher than 100%.

> I went to a transition classroom, and now there's a red flag next to my name. I guess now I'm an ineffective teacher? I keep getting letters from the district, saying 'You've been recognized as an outstanding teacher'…this, this, and that. But now because I teach English Language Learners who 'transition in,' my scores drop? And I get a flag next to my name for not teaching them well?

A Tennessee study of teachers who volunteered to be evaluated based on VAMs, and to have a substantial share of their compensation tied to their VAM results, corroborated this evidence: After three years, 85 percent thought that the VAM evaluation ignored important aspects of their performance that were not measured by test scores, and two thirds thought that the VAM did not do a good job of distinguishing effective from ineffective teachers. (Springer et. al., 2010)

Campbell's Law and the Dangers of Quantification

Over thirty years ago, Donald Campbell pointed out that overreliance on imperfect quantitative measures of performance can do real harm. (Campbell, 1975) He formulated what is now known as Campbell's Law:

> The more any quantitative social indicator is used for social decision-making, the more subject it will be to corruption pressures and the more apt it will be to distort and corrupt the social processes it is intended to monitor.

One of Campbell's examples concerned the use of student achievement gains to evaluate instructional quality. When teachers were paid based on their students' achievement test score gains, they devoted their effort to teaching the answers to specific test items rather than to general instruction. Campbell pointed out that this is a general problem:

> [A]chievement tests may well be valuable indicators of general school achievement under conditions of normal teaching aimed at general competence. But when test scores become the goal of the teaching process, they both lose their value as indicators of educational status and distort the educational process in undesirable ways.

We already have evidence from the Houston study described earlier that Campbell's predictions are coming true. Teachers in Houston report seeking to boost their scores by avoiding certain subjects and types of students, and by seeking assignments to teach particular subjects / grades, while being increasingly confused and demoralized by the system. As a Houston teacher noted, voicing a common concern:

> I'm scared to teach in the 4th grade. I'm scared I might lose my job if I teach in an [ELL] transition grade level, because I'm scared my scores are going to drop, and I'm going to get fired because there's probably going to be no growth.

Although we cannot be certain how prevalent this sort of reaction will be, it will certainly be more prevalent than is already apparent in low-stakes settings, and the VAM scores will thus be even less reliable than is indicated by the research reviewed here.

The long-run implications for teacher recruitment and retention in schools and classrooms serving students who appear to negatively influence measured gains have yet to be studied empirically. There is good reason to worry that over-reliance on student test scores for teacher evaluation will make it harder to staff these schools and classrooms with high-quality teachers.

Professional Consensus about the Use of Value-Added Methods in Teacher Evaluation

For all of these reasons, most researchers have concluded that value-added modeling (VAM) is not appropriate as a primary measure for evaluating individual teachers. A major report by the RAND Corporation concluded that:

> The research base is currently insufficient to support the use of VAM for high-stakes decisions about individual teachers or schools. (McCaffrey et al, 2005)

Similarly, Henry Braun of the Educational Testing Service concluded in his review of research:

> VAM results should not serve as the sole or principal basis for making consequential decisions about teachers. There are many pitfalls to making causal attributions of teacher effectiveness on the basis of the kinds of data available from typical school districts. We still lack sufficient understanding of how seriously the different technical problems threaten the validity of such interpretations. (Braun, 2005)

Finally, the National Research Council's Board on Testing and Assessment concluded that:

> VAM estimates of teacher effectiveness that are based on data for a single class of students should not used to make operational decisions because such estimates are far too unstable to be considered fair or reliable. (National Research Council, 2009)

Other Approaches to Teacher Evaluation

While value-added models based on student test scores are problematic for making evaluation decisions for individual teachers, they are useful for looking at groups of teachers for research purposes – for example, to examine how specific teaching practices or measures of teaching influence the learning of large numbers of students. The larger scale of these studies reduces error, and their frequent use of a wider range of outcome measures allows more understanding of the range of effects of particular strategies or interventions. A crucial aspect of these studies is that results do not have important consequences for the individual study participants, and thus the participants do not have incentives to distort their practices in order to improve their measured outcomes.

These kinds of analyses provide other insights for teacher evaluation, since there is a large body of evidence over many decades concerning how specific teaching practices influence student learning gains. For example, there is considerable evidence that effective teachers:

- Understand subject matter deeply and flexibly
- Connect what is to be learned to students' prior knowledge and experience
- Create effective scaffolds and supports for learning
- Use instructional strategies that help students draw connections, apply what they are learning, practice new skills, and monitor their own learning
- Assess student learning continuously and adapt teaching to student needs
- Provide clear standards, constant feedback, and opportunities for revising work
- Develop and effectively manage a collaborative classroom in which all students have membership. (Darling-Hammond & Bransford, 2005)

These aspects of effective teaching, supported by research, have been incorporated into professional standards for teaching that offer some useful approaches to teacher evaluation.

Using Professional Standards for Teacher Evaluation

Professional standards defining accomplished teaching were first developed by the National Board for Professional Teaching Standards to guide assessments for veteran teachers. Subsequently, a group of states working together under the auspices of the Council for Chief State School Officers created the Interstate New Teacher Assessment and Support Consortium (INTASC), which translated these into standards for beginning teachers, adopted by over 40 states for initial teacher licensing. A recent revision of the INTASC teaching standards has been aligned with the Common Core Standards in order to reflect the kind of teacher knowledge, skills, and understandings needed to enact the standards.

These standards have become the basis for assessments of teaching that produce ratings which are much more stable than value-added measures. At the same time, they incorporate classroom evidence of student learning and they have recently been shown in larger-scale studies to predict teachers' value-added effectiveness, so they help ground evaluation in student learning in more stable ways. Typically the performance assessments ask teachers to document their plans and teaching for a unit of instruction linked to the state standards, adapt them for special education students and English language learners, videotape and critique lessons, and collect and evaluate evidence of student learning.

A number of studies have found that the National Board Certification assessment process identifies teachers who are more effective in raising student achievement than other teachers.[4] Equally important, studies have found that teachers' participation in the National Board process stimulates improvements in their practice. (Bond et al., 2000; Cavaluzzo, 2004; Goldhaber & Anthony, 2005; Smith et al., 2005; Vandevoort, Amrein-Beardsley, & Berliner, 2004) Similar performance assessments, used with beginning teachers in Connecticut and California, have been found to predict their students' achievement gains on state tests. (Athanases, 2004; Sato, Wei & Darling-Hammond, 2008; Tracz, Sienty & Mata, 1994; Tracz et al., 1995) The Performance Assessment for California Teachers (PACT) has also been found to improve beginning teachers' competence and to stimulate improvements in the teacher education programs that use it as a measure. (Wilson & Hallum, 2006; Newton, 2011)

Professional standards have also been translated into teacher evaluation instruments in use at the local level. In a study of three districts using standards-based evaluation systems, researchers found significant relationships between teachers' ratings and their students' gain scores on standardized tests, and evidence that teachers' practice improved as they were given frequent feedback in relation to the standards. (Chung, 2008; Wei & Pecheone, 2010) In the schools and districts studied, assessments of teachers were based on well-articulated standards of practice evaluated through evidence including observations of teaching along with teacher pre- and post-observation interviews and, sometimes, artifacts such as lesson plans, assignments, and samples of student work.

The Cincinnati Public Schools use an unusually careful standards-based system for teacher evaluation. Evaluations involve multiple classroom observations and detailed written feedback to teachers. A recent study finds that these evaluations lead to substantial improvements in the performance of mid-career teachers, as measured by student achievement, that persist for many years thereafter. (Milanowski, Kimball & White, 2004; Milanowski, 2004; Rockoff & Speroni, 2010) This system, like the others described above, does not use student test scores; however, all of them have the effect of improving student achievement.

The Gates Foundation has launched a major initiative to find additional tools that are based on professional standards and validated against student achievement gains to be used in teacher evaluation at the local level. The *Measures of Effective Teaching (MET) Project* has developed a number of tools, including observations or videotapes of teachers, supplemented with other artifacts of practice (lesson plans, assignments, etc.), that can be scored according to a

set of standards which reflect practices associated with effective teaching. (Taylor & Tyler, 2001)

Building Systems for Teacher Evaluation that Support Improvement and Decision Making

Systems that help teachers improve and that support timely and efficient personnel decisions have more than good instruments. Successful systems use multiple classroom observations across the year by expert evaluators looking at multiple sources of data that reflect a teacher's instructional practice, and they provide timely and meaningful feedback to the teacher.

For example, the Teacher Advancement Program, which is based on the standards of the National Board and INTASC, as well as the standards-based assessment rubrics developed in Connecticut (Gates Foundation, 2010; Rothstein, 2011), ensures that teachers are evaluated four to six times a year by master / mentor teachers or principals who have been trained and certified in a rigorous four-day training. The indicators of good teaching are practices that have been found to be associated with desired student outcomes. Teachers also study the rubric and its implications for teaching and learning, look at and evaluate videotaped teaching episodes using the rubric, and engage in practice evaluations. After each observation, the evaluator and teacher meet to discuss the findings and to make a plan for ongoing growth. Ongoing professional development, mentoring, and classroom support are provided to help teachers meet these standards. Teachers in TAP schools report that this system, along with the intensive professional development offered, is substantially responsible for improvements in their practice and the gains in student achievement that have occurred in many TAP schools. (Solomon, White, Cohen & Woo, 2007)

In districts that use Peer Assistance and Review (PAR) programs, highly expert mentor teachers conduct some aspects of the evaluation and provide assistance to teachers who need it. Key features of these systems include not only the instruments used for evaluation but also the expertise of the consulting teachers or mentors – skilled teachers in the same subject areas and school levels who have released time to serve as mentors to support their fellow teachers – and the system of due process and review that involve a panel of both teachers and administrators in making recommendations about personnel decisions based on the evidence presented to them from the evaluations. Many systems using this approach have been found not only to improve teaching, but also to successfully identify teachers for continuation and tenure as well as intensive assistance and personnel action. (NCTAF, 1996;Van Lier, 2008)

Some systems ask teachers to assemble evidence of student learning as part of the overall judgment of effectiveness. Such evidence is drawn from classroom and school-level assessments and documentation, including pre- and post-test measures of student learning in specific courses or curriculum areas, and evidence of student accomplishments in relation to teaching activities. A study of Arizona's career ladder program, which requires the use of various methods of student assessment to complement evaluations of teachers' practice, found that, over time, participating teachers demonstrated an increased ability to create locally-developed assessment tools to assess student learning gains in their classrooms; to develop and evaluate pre- and post-

tests; to define measurable outcomes in hard-to-quantify areas like art, music, and physical education; and to monitor student learning growth. They also showed a greater awareness of the importance of sound curriculum development, more alignment of curriculum with district objectives, and increased focus on higher quality content, skills, and instructional strategies. (Packard & Dereshiwsky, 1991) Thus, the development and use of student learning evidence, in combination with examination of teaching performance, can stimulate improvements in practice.

Some districts in the United States, along with high-achieving countries like Singapore, include a major emphasis on teacher collaboration in their evaluation systems. This kind of measure is supported by studies which have found that stronger value-added gains for students are enabled by teachers who work together as teams (Jackson & Bruegmann, 2009) and by higher levels of teacher collaboration for school improvement. (Goddard & Goddard, 2007)

Summary and Conclusions

New approaches to teacher evaluation should take advantage of research on teacher effectiveness. While there are considerable challenges in the use of value-added test scores to evaluate individual teachers directly, the use of value-added methods can help to validate measures that are productive for teacher evaluation.

With respect to value-added measures of student achievement tied to individual teachers, current research suggests that high-stakes, individual-level decisions, or comparisons across highly dissimilar schools or student populations, should be avoided. Valid interpretations require aggregate-level data and should ensure that background factors – including overall classroom composition – are as similar as possible across groups being compared. In general, such measures should be used only in a low-stakes fashion when they are part of an integrated analysis of what the teacher is doing and who is being taught.

Other teacher evaluation tools that have been found to be both predictive of student learning gains and productive for teacher learning include *standards-based evaluation processes.* These include systems like National Board Certification and performance assessments for beginning teacher licensing as well as district and school-level instruments based on professional teaching standards. Effective systems have developed an integrated set of measures that show what teachers do and what happens as a result. These measures may include evidence of student work and learning, as well as evidence of teacher practices derived from observations, videotapes, artifacts, and even student surveys.

These tools are most effective when embedded in systems that support evaluation expertise & well-grounded decisions, by ensuring that evaluators are trained, evaluation and feedback are frequent, mentoring and professional development are available, and processes are in place to support due process and timely decision making by an appropriate body.

With these features in place, evaluation can become a more useful part of a productive human capital system, supporting accurate information about teachers, helpful feedback, and well-grounded personnel decisions.

References

Amrein-Beardsley, A. & Collins, C. (forthcoming). The SAS® Education Value-Added Assessment System (EVAAS®): Its Intended and Unintended Effects in a Major Urban School System. Arizona State University.

Athanases, S. (1994). Teachers' reports of the effects of preparing portfolios of literacy instruction. *Elementary School Journal, 94*(4), 421-43.

Bill & Melinda Gates Foundation (2010). Learning About Teaching: Initial Findings from the Measures of Effective Teaching Project. Seattle: Author. Rothstein, Jesse (2011). Review of 'Learning About Teaching: Initial Findings from the Measures of Effective Teaching Project. Boulder, CO: National Education Policy Center.

Bond, T. Smith, W. Baker, & J. Hattie (2000). The certification system of the National Board for Professional Teaching Standards: A construct and consequential validity study (Greensboro, NC: Center for Educational Research and Evaluation).

Braun, H. (2005). Using Student Progress to Evaluate Teachers: A Primer on Value-Added Models (Princeton, NJ: Educational Testing Service.

Briggs, D. & Domingue, B. (2011). Due Diligence and the Evaluation of Teachers: A review of the value-added analysis underlying the effectiveness rankings of Los Angeles Unified School District teachers by the Los Angeles Times. Boulder, CO: National Education Policy Center.

Campbell, D. T. (1975). "Assessing the Impact of Planned Social Change," in G. M. Lyons, editor, *Social Research and Public Policies*. Hanover, New Hampshire: The Public Affairs Center, Dartmouth College. Pp. 3-45.

Carrell, S. & West, J. (2010). "Does Professor Quality Matter? Evidence from Random Assignment of Students to Professors," *Journal of Political Economy* 118(3).

Cavaluzzo, L. (2004). Is National Board Certification an effective signal of teacher quality? (National Science Foundation No. REC-0107014). Alexandria, VA: The CNA Corporation.

Chung, R. R. (2008). Beyond Assessment: Performance Assessments in Teacher Education, Teacher Education Quarterly, 35 (1): 7-28.

Corcoran, S.P., Jennings, J. L. & Beveridge, A. A. (2011). Teacher effectiveness on high- and low-stakes tests. Working paper. NY: New York University.

Darling-Hammond, L. & Bransford, J. (2005). *Preparing Teachers for a Changing World: What Teachers should Learn and Be Able to Do.* San Francisco: Jossey-Bass.

Darling-Hammond, L. (2011) Evaluating teacher Evaluation: We Know About Value-Added Models and Other Methods. Phi Delta Kappan
Stanford University makes this peer-reviewed final draft available under a Creative Commons Attribution-Noncommercial License. The published version is available from the publisher, subscribing libraries, and the author
Goddard, Y. & Goddard, R. D. (2007). A theoretical and empirical investigation of teacher collaboration for school improvement and student achievement in public elementary schools. *Teachers College Record, 109*(4), 877–896.

Goldhaber, D., & Anthony, E. (2005). Can teacher quality be effectively assessed? Seattle, WA: University of Washington and the Urban Institute.

Jackson, C. K. & Bruegmann, E. (2009, August). Teaching students and teaching each other: The importance of peer learning for teachers. Washington, DC: National Bureau of Economic Research.

Lockwood, J. R., McCaffrey, D. F., Hamilton, L.S., Stetcher, B., Le, V. N., & Martinez, J. F. (2007). The sensitivity of value-added teacher effect estimates to different mathematics achievement measures. *Journal of Educational Measurement, 44* (1), 47 – 67.

McCaffrey, D. F., Koretz, D., Lockwood, J.R., Hamilton, L.S. (2005). Evaluating Value-Added Models for Teacher Accountability. Santa Monica: RAND Corporation.

Milanowski, A., Kimball, S.M., & White, B. (2004). The relationship between standards-based teacher evaluation scores and student achievement. University of Wisconsin-Madison: Consortium for Policy Research in Education.

Milanowski, A. (2004). The Relationship Between Teacher Performance Evaluation Scores and Student Achievement: Evidence From Cincinnati, Peabody Journal of Education 79 (4): 33-53.
National Commission on Teaching and America's Future (1996). What Matters Most: Teaching for America's Future. NY: NCTAF.

National Research Council, Board on Testing and Assessment (2009). Letter Report to the U.S. Department of Education.

Newton, X., Darling-Hammond, L., Haertel, E., & Thomas, E. (2010) Value-Added Modeling of Teacher Effectiveness: An exploration of stability across models and contexts. *Educational Policy Analysis Archives, 18* (23). http://epaa.asu.edu/ojs/article/view/810.

Newton, S.P. (2011). Predictive Validity of the Performance Assessment for California Teachers. Stanford, CA: Stanford Center for Opportunity Policy in Education, 2010, available at http://scale.stanford.edu.

Packard, R. & Dereshiwsky, M. (1991). Final quantitative assessment of the Arizona career ladder pilot-test project. Flagstaff: Northern Arizona University.

Rockoff, J. & Speroni, C. (2010). Subjective and Objective Evaluations of Teacher Effectiveness (New York: Columbia University, 2010).

Darling-Hammond, L. (2011) Evaluating teacher Evaluation: We Know About Value-Added Models and Other Methods. Phi Delta Kappan

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

Solomon, L., White, J. T., Cohen, D. & Woo, D. (2007). *The effectiveness of the Teacher Advancement Program.* National Institute for Excellence in Teaching, 2007.

Springer, M. G., Ballou, D., Hamilton, L., Le, V., Lockwood, V.R., McCaffrey, D., Pepper, M., & Stecher, B. M. (2010). *Teacher Pay for Performance: Experimental Evidence from the Project on Incentives in Teaching*. Nashville, TN: National Center on Performance Incentives at Vanderbilt University.

Taylor, E. S. & Tyler, J. H. (2011, March). "The Effect of Evaluation on Performance: Evidence from Longitudinal Student Achievement Data of Mid-Career Teachers." National Bureau of Economic Research working paper number 16877.

Todd, P.E. & Wolpin, K. I. (2003). "On the Specification and Estimation of the Production Function for Cognitive Achievement," *Economic Journal* 113(485).

Tracz, S.M., Sienty, S. & Mata, S. (1994, February). The self-reflection of teachers compiling portfolios for National Certification: Work in progress. Paper presented at the Annual Meeting of the American Association of Colleges for Teacher Education. Chicago, IL.

Tracz, S.M., Sienty, S. Todorov, K., Snyder, J., Takashima, B., Pensabene, R., Olsen, B., Pauls, L., & Sork, J. (1995, April). Improvement in teaching skills: Perspectives from National Board for Professional Teaching Standards field test network candidates. Paper presented at the annual meeting of the American Educational Research Association. San Francisco, CA.

Sato, M., Wei, R.C. & Darling-Hammond, L. (2008). Improving Teachers' Assessment Practices through Professional Development: The Case of National Board Certification, American Educational Research Journal, 45: pp. 669-700.

Smith, T., Gordon, B., Colby, S., & Wang, J. (2005). An examination of the relationship of the depth of student learning and National Board certification status (Office for Research on Teaching, Appalachian State University).

Van Lier, P. (2008). Learning from Ohio's best teachers: A homegrown model to improve our schools. Policy Matters Ohio.

Vandevoort, L. G., Amrein-Beardsley, A., & Berliner, D. C. (2004). National Board certified teachers and their students' achievement. *Education Policy Analysis Archives, 12*(46), 117.

Wei, R. C. & Pecheone, R. (2010). Teaching Performance Assessments as Summative Events and Educative Tools. In Mary Kennedy (ed.), Teacher Assessment and Teacher Quality: A Handbook (New York: Jossey-Bass).

Darling-Hammond, L. (2011) Evaluating teacher Evaluation: We Know About Value-Added Models and Other Methods. Phi Delta Kappan

Wilson, M. & Hallum, P.J. (2006). Using Student Achievement Test Scores as Evidence of External Validity for Indicators of Teacher Quality: Connecticut's *Beginning Educator Support and Training* Program. Berkeley, CA: University of California at Berkeley.

[1] This article is developed from a Capitol Hill Briefing sponsored by the American Educational Research Association and the National Academy of Education, held on September 15, 2011.