Evaluating teacher evaluation

Popular modes of evaluating teachers are fraught with inaccuracies and inconsistencies, but the field has identified better approaches.

By Linda Darling-Hammond, Audrey Amrein-Beardsley, Edward Haertel, and Jesse Rothstein

Practitioners, researchers, and policy makers agree that most current teacher evaluation systems do little to help teachers improve or to support personnel decision making. There's also a growing consensus that evidence of teacher contributions to student learning should be part of teacher evaluation systems, along with evidence about the quality of teacher practices. "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.

Value-added models enable researchers to use statistical methods to measure changes in student scores over time while considering student characteristics and other factors often found to influence achievement. In large-scale studies, these methods have proved valuable for looking at factors affecting achievement and measuring the effects of programs or interventions.

Using VAMs for individual teacher evaluation is based on the belief 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 gains in student achievement 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; and
- 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.

However, value-added models don't actually measure most of these factors. VAMs rely on statistical controls for past achievement to parse out the small portion of student gains that is due to other factors,

LINDA DARLING-HAMMOND (ldh@stanford.edu) is the Charles Ducommun professor of teaching and teacher education, Stanford University, Stanford, Calif. **AUDREY AMREIN-BEARDSLEY** is an associate professor of education, Arizona State University, Phoenix, Ariz. **EDWARD HAERTEL** is the Jacks Family professor of education, Stanford University, Stanford, Calif. **JESSE ROTHSTEIN** is an associate professor of economics and public policy, University of California, Berkeley.

of which the teacher is only one. 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 inconsistent.

Researchers have found that teacher 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.

A study examining data from five school districts found, for example, that of teachers who scored in the bottom 20% of rankings in one year, only 20% to 30% had similar ratings the next year, while 25% to 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' value-added scores differ significantly when different tests are used, even when these are within the same content area.

Teacher effectiveness also varies significantly when different statistical methods are used (Briggs & Domingue, 2011; Newton et al., 2010; Rothstein, 2007). For example, when researchers used a different model to recalculate the value-added scores for teachers published in the *Los Angeles Times* in 2011, they found that from 40% to 55% of them would get noticeably different scores (Briggs & Domingue, 2011).

Teachers' value-added scores also differ significantly when different tests are used, even when these are within the same content area (Bill & Melinda Gates Foundation, 2010; Lockwood et al., 2007). This raises concerns both about measurement error and, when teacher evaluation results are tied to student test scores, the effects of emphasizing "teaching to the test" at the expense of other kinds of learning, especially given the narrowness of most tests in the United States.

2. Teachers' value-added performance is affected by the students assigned to them.

VAMs are designed to identify teachers' effects

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, & Thomas (2010).

Thinkstock/iStockphoto

when students are assigned to teachers randomly. However, students aren't randomly assigned to teachers — and statistical models can't fully adjust for the fact that some teachers will have a disproportionate number of students who have greater challenges (e.g., 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).

Teachers are advantaged or disadvantaged based on the students they teach.

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 teacher "effects" on students' *prior* test scores. Logically, for example, 5th-grade teachers can't influence their students' 3rd-grade test scores. So a VAM that identifies teachers' true effects should show *no* effect of 5th-grade teachers on 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 (Rothstein, 2010).

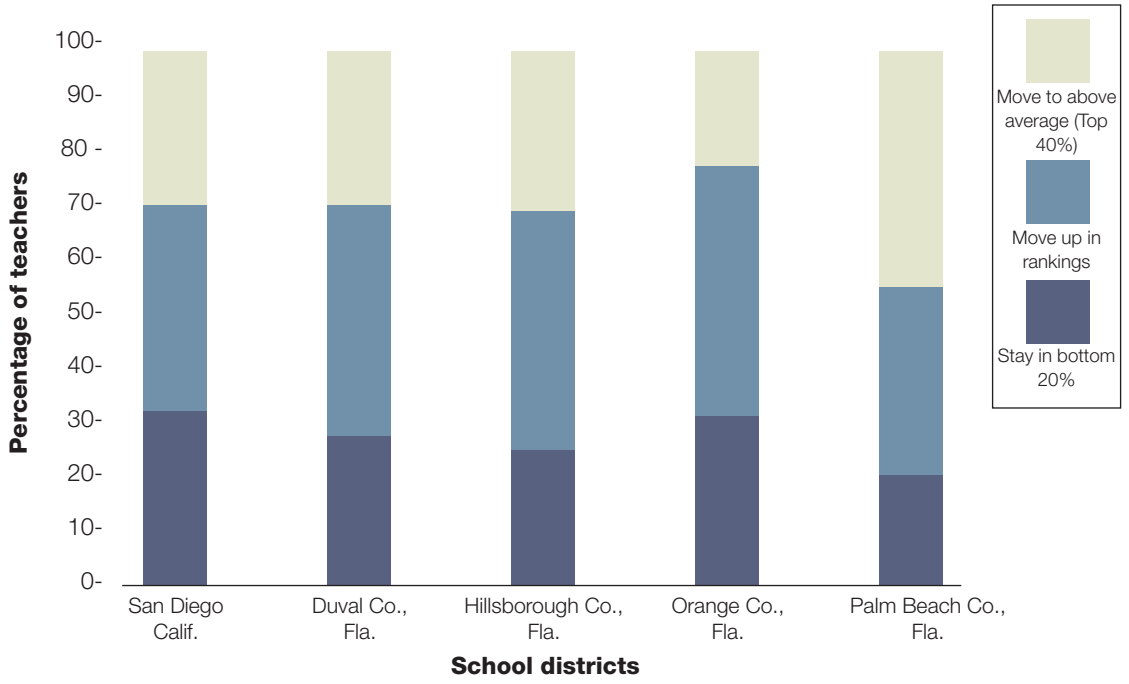
One study that found considerable instability in teachers' value-added scores from class to class and year to year examined changes in student characteristics associated with changes in teacher ratings. After controlling for prior student test scores *and* student characteristics, the study still found significant correlations between teacher ratings and students' race/ethnicity, income, language background, and parent education. Figure 2 illustrates this finding for an experienced English teacher 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.

This variability raises concerns that using such ratings for evaluating teachers could create disincentives for teachers to serve high-need students.

3. Value-added ratings can't disentangle the many influences on student progress.

Given all of the other factors operating, it appears

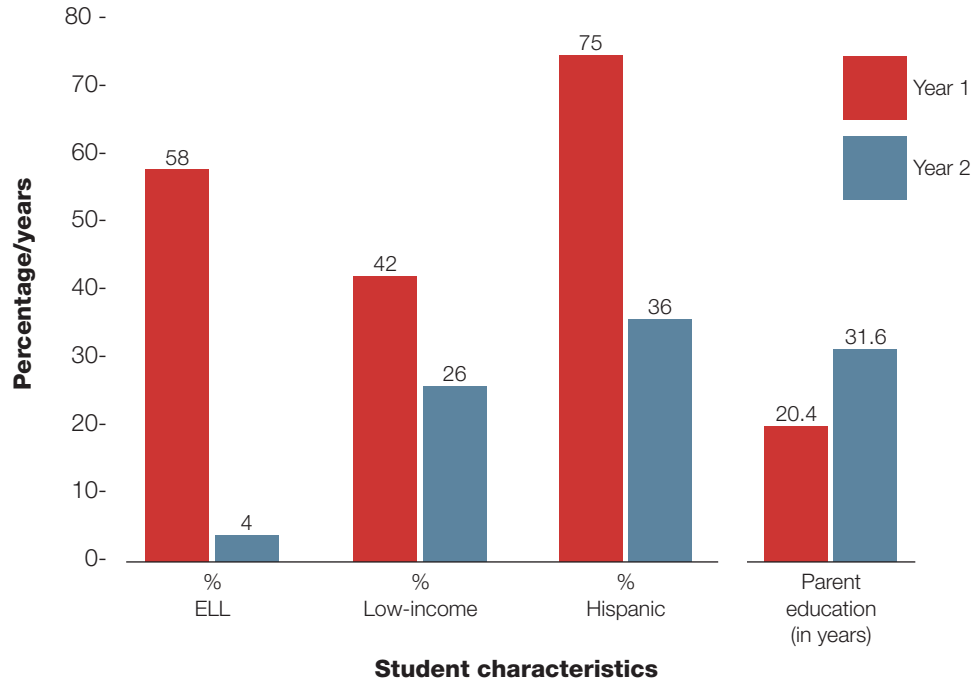
FIG. 1.
Changes in VA scores from 2001 to 2002 for low-ranking teachers



Source: Sass, T. (2008).

FIG. 2.

Student characteristics in years 1 and 2 for a teacher whose ranking changed from the 1st to the 10th decile



that "teacher effectiveness" is not a stable enough construct to 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 effects). Furthermore, 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 would depend on whether the student tests used for the VAM emphasize skills and topics for which the teacher is relatively more or relatively less effective.

Other research indicates that teachers whose students do best on end-of-year tests aren't always effective at promoting longer-run achievement for their students. Thus, VAM-style measures may be influenced by how much the teacher emphasizes short-run test preparation. One study even found that teachers who raised end-of-course grades most were, on average, less effective than others at preparing students for next year's course (Carrell & West, 2010).

Initial research on using value-added methods to dismiss some teachers and award bonuses to others shows that value-added ratings often don't agree with ratings from skilled observers and are influenced by all of the factors described above.

For example, one of the teachers dismissed in Houston as a result of its Education Value-Added Assessment System (EVAAS) scores was a 10-year veteran who 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, in press). She showed positive VA scores on 8 of 16 tests over four years (50% of the total observations), with wide fluctuations from year to year, both across and within subjects. (See Table 2.) It is worth noting that this teacher's lower value-added in 4th grade, when English learners are mainstreamed in Houston, was also a pattern for many other teachers.

The wide variability shown in this teacher's ratings from year to year, like that documented in many other studies, wasn't unusual for Houston teachers in this analysis, regardless of whether the teacher was terminated. Teachers said they couldn't identify a relationship between their instructional practices and their value-added ratings, 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 (Amrein-Beardsley & Collins, in press).

Deepen your understanding of this article with questions and activities in this month's *Kappan* Professional Development Discussion Guide by Lois Brown Easton. Download a PDF of the guide at **kappanmagazine.org**.

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". . . [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 that 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's 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 whose supervisor consistently rated her as "exceeding expectations" or "proficient" and who also was receiving positive VA scores about 50% of the time, had a noticeable drop in her value-added ratings when a large number of English language learners transitioned into her classroom. Overall, the study found that, in this system:

- Teachers of grades in which English language learners (ELLs) are transitioned into mainstreamed classrooms are the least likely to show "added value."
- Teachers of large numbers of special education students in mainstreamed classrooms are also found to have lower "value-added" scores, on average.
- Teachers of 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? (Amrein-Beardsley & Collins, in press).

A study of Tennessee 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% thought the VAM evaluation ignored important aspects of their performance that test scores didn't measure, and two-thirds thought VAM didn't do a good job of distinguishing effective from ineffective teachers (Springer et al., 2010).

Other approaches

For all of these reasons and more, most researchers have concluded that value-added modeling is not appropriate as a primary measure for evaluating individual teachers. (See, for example, Braun, 2005; National Research Council, 2009.)

While value-added models based on test scores

TABLE 2.
2006-2010 EVAAS scores of a teacher dismissed as a result of these scores

EVAAS scores (Teacher A)	2006-2007	2007-2008	2008-2009	2009-2010
	GRADE 5	GRADE 4	GRADE 3	GRADE 3
Math	-2.03	+0.68*	+0.16*	+03.26
Reading	-1.15	-0.96*	+2.03	+1.81
Language arts	+1.12	-0.49*	-1.77	-0.20*
Science	+2.37	-3.45	n/a	n/a
Social studies	+0.91*	-2.39	n/a	n/a
ASPIRE bonus	$3,400	$700	$3,700	$0

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

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. Such analyses provide other insights for teacher evaluation because we have a large body of evidence over many decades concerning how specific teaching practices influence student learning gains. For example, we know 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're 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; and
- 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

The National Board for Professional Teaching Standards (NBPTS) defined accomplished teaching 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 that have been adopted by over 40 states for initial teacher licensing. Revised INTASC teaching standards have been aligned with the Common Core Standards to reflect the knowledge, skills, and understandings that teachers need to enact the standards.

These standards have become the basis for assessments of teaching that produce ratings that are much more stable than value-added measures. At the same time, these standards incorporate classroom evidence of student learning, and large-scale studies have shown that they can predict teachers' value-added effectiveness (National Research Council, 2008; Wilson et al., 2011), so they have helped ground evaluation in student learning in more stable ways. Typically, performance assessments ask teachers to document their plans and teaching for a unit of instruction linked to state standards, adapt them for special education students and English language learners, videotape and critique lessons, and collect and evaluate evidence of student learning.

The notion that there is a stable "teacher effect" that's 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.

Professional standards have also been translated into teacher evaluation instruments at the local level. Cincinnati Public Schools uses an unusually careful standards-based system for teacher evaluation that involves multiple classroom observations and detailed written feedback to teachers. This system, like several others in local districts, has been found both to produce ratings that reflect teachers' effectiveness in supporting student learning gains and to improve teachers' performance and their future effectiveness (Milanowski, Kimball & White, 2004; Milanowski, 2004; Rockoff & Speroni, 2010; Taylor & Tyler, 2011.)

A Bill & Melinda Gates Foundation initiative is identifying additional tools 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 standards that reflect practices associated with effective teaching.

Building better systems

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, and they provide timely and meaningful feedback to the teacher.

For example, schools using the Teacher Advancement Program, which is based on NBPTS and INTASC standards as well as the standards-based assessment rubrics developed in Connecticut (Bill & Melinda Gates Foundation, 2010; Rothstein, 2011),

evaluate teachers four to six times a year using master/mentor teachers or principals certified in a rigorous four-day training. The indicators of good teaching are practices 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 discuss the findings and plan for ongoing growth. Schools provide professional development, mentoring, and classroom support to help teachers meet these standards. TAP teachers say this system, along with the intensive professional development offered, is substantially responsible for improving their practice and for student achievement gains in many TAP schools (Solmon, White, Cohen, & Woo, 2007).

Successful systems use multiple classroom observations, expert evaluators, multiple sources of data, are timely, and provide meaningful feedback to the teacher.

In districts that use Peer Assistance and Review (PAR) programs, highly expert mentor teachers support novice teachers and veteran teachers who are struggling, and they conduct some aspects of the evaluation. Key features of these systems include not only the evaluation instruments but also the expertise of the consulting teachers or mentors, and a system of due process and review in which a panel of teachers and administrators make recommendations about personnel decisions based on evidence from the evaluations. Many systems using this approach have improved teaching while they have also become more effective in identifying teachers for continuation and tenure as well as intensive assistance and, where needed, dismissal (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 teachers to use various methods of student assessment to complement evaluations of teacher practice, found that, over time, participating teachers improved their ability to create tools to assess student learning gains; to develop and evaluate before and after 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).

Some U.S. districts, along with high-achieving countries like Singapore, emphasize teacher collaboration in their evaluation systems. This kind of measure is supported by studies finding that students have stronger achievement gains when teachers work together in teams (Jackson & Bruegmann, 2009) and when there is greater teacher collaboration for school improvement (Goddard & Goddard, 2007).

In conclusion

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

Research indicates that value-added measures of student achievement tied to individual teachers should not be used for high-stakes, individual-level decisions, or comparisons across highly dissimilar schools or student populations. 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're part of an integrated analysis of teachers' practices.

Standards-based evaluation processes have also been found to be predictive of student learning gains and productive for teacher learning. 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 and 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 teaching and learning system, supporting accurate information about teachers, helpful feedback, and well-grounded personnel decisions.

References

Amrein-Beardsley, A. & Collins, C. (In press). *The SAS education value-added assessment system (EVAAS): Its intended and unintended effects in a major urban school system.* Tempe, AZ: Arizona State University.

Bill & Melinda Gates Foundation. (2010). *Learning about teaching: Initial findings from the Measures of Effective Teaching Project.* Seattle, WA: Author.

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.

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

Darling-Hammond, L. & Bransford, J. (2005). *Preparing teachers for a changing world: What teachers should learn and be able to do.* San Francisco, CA: Jossey-Bass.

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.

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

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

Milanowski, A. (2004). The relationship between teacher performance evaluation scores and student achievement: Evidence from Cincinnati. *Peabody Journal of Education, 79* (4), 33-53.

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

National Commission on Teaching and America's Future. (1996). *What matters most: Teaching for America's future.* New York, NY: Author.

National Research Council, Board on Testing and Assessment. (2008). *Assessing accomplished teaching: Advanced-level certification programs.* Washington, DC: National Academies Press.

National Research Council, Board on Testing and Assessment. (2009). Letter report to the U.S. Department of Education. Washington, DC: Author.

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).

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

Rockoff, J. & Speroni, C. (2010). *Subjective and objective evaluations of teacher effectiveness.* New York, NY: Columbia University.

Rothstein, J. (2007). Do value-added models add value? Tracking, fixed effects, and causal inference. *CEPS Working Paper No. 159*. Cambridge, MA: National Bureau of Economic Research.

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

Rothstein, J. (2011). *Review of "Learning about teaching: Initial findings from the Measures of Effective Teaching Project."* Boulder, CO: National Education Policy Center.

Sass, T. (2008). *The stability of value-added measures of teacher quality and implications for teacher compensation policy*. Washington, DC: CALDER.

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

Springer, M., Ballou, D., Hamilton, L., Le, V., Lockwood, V., McCaffrey, D., Pepper, M., & Stecher, B. (2010). *Teacher pay for performance: Experimental evidence from the Project on Incentives in Teaching*. Nashville, TN: National Center on Performance Incentives.

Taylor, E. & Tyler, J. (2011, March). The effect of evaluation on performance: Evidence of longitudinal student achievement data of mid-career teachers. *Working Paper No. 16877*. Cambridge, MA: National Bureau of Economic Research.

Van Lier, P. (2008). Learning from Ohio's best teachers: A homegrown model to improve our schools. *Policy Matters Ohio.* www.policymattersohio.org/learning-from-ohios-best-teachers-a-homegrown-model-to-improve-our-schools

Wilson, M, Hallam, P., Pecheone, R., & Moss, P. (2011). *Investigating the validity of portfolio assessments of beginning teachers: Relationships with student achievement and tests of teacher knowledge.* Berkeley, CA: Berkeley Evaluation, Assessment, and Research Center.