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

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**

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 inuence achievement. In large-scale studies, these methods have proved valuable for [illegible] at factors affecting achievement and

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

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.

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).

Another teacher classi ed 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 ipped 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 ip- opped 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 speci c 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 “pro cient” 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:

- U Ñ /i > V ˆ i Å È Ñ w Ñ € Å > ` i È Ñ ‹ › Ñ å ˆ ‹ V ˆ Ñ learners (ELLs) are transitioned into mainstreamed classrooms are the least likely to show “added value.”
- U Ñ /i > V ˆ i Å È Ñ w Ñ ’ > Å € i Ñ › Þ – L i Å È Ñ w Ñ È students in mainstreamed classrooms are also found to have lower “value-added” scores, on average.
- U Ñ /i > V ˆ i Å È Ñ w Ñ € ‹ w Ó i ` Ñ È Ó Þ ` i › Ó added because their students are already near the top of the test score range.
- U Ñ , > Ó ‹ › € È Ñ V ˆ > › € i Ñ V › È ‹ ` i Å > L ’ ç Ñ å ˆ 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 ag 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 ag 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.

ground evaluation in student learning in more stable ways. Typically, performance assessments ask teachers to document their plans and teaching for a unit

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 ndings 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 Uni ed 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).

National Research Council, Board on Testing and Assessment. (2008). Assessing accomplished teaching: Advanced-level certi cation 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, xed 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 ndings 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.