EDUCATION WEEK
Published Online: April 1, 2014
Published in Print: April 2, 2014, as **Take the Time to Evaluate Teacher Evaluation**

Take the Time to Evaluate Teacher Evaluation

By Tia Sukin, W. Alan Nicewander, Phoebe Winter, Howard Mitzel, Lisa Keller, Matt Schulz

Do you remember New York City's Pascale Mauclair? She was an educator who primarily instructed English-language learners and won accolades for teaching excellence. But then she was labeled the worst teacher in New York City in 2012. Following the release of much-publicized assessment-based ratings by New York City's education department, stunned parents demanded that their children be instructed by a different teacher, and that Ms. Mauclair—whose test-based ratings were low—be fired. This happened despite tremendous **support for the teacher from her principal**. Later, it was revealed that her rating had **failed to account for such factors** as her students' English-language-learner status.

Thankfully, there is some good news for teacher-evaluation systems that could help avoid this type of error. Last June, the U.S. Department of Education agreed to allow some states **to seek an additional year** before they must rely on new evaluation systems that incorporate student test scores. Thus, their deadlines will be extended to the 2016-17 school year, giving those states a total of three years before teacher-evaluation systems must be used for high-stakes purposes, such as identifying teachers for sanctions or rewards.

This time frame is an absolute window of opportunity in which to conduct necessary validity studies. Without studies to support the use of student scores for evaluating educators, good teachers could be dismissed and teachers needing support, or those who should not be teaching at all, may not be identified.

When teachers challenge the validity of evaluation systems, it can appear self-serving. Because of this, it is the responsibility of testing professionals such as us to weigh in on the use of student scores in the evaluation of teachers. Testing professionals must lead the way in providing a framework for evaluating proposed systems that purport to measure teacher quality.

In fact, unless appropriate validity studies are conducted, widespread use of student test scores for evaluating teachers will constitute a serious violation of the **Standards for Educational and Psychological Testing**. These standards were developed collaboratively by the American Educational Research Association, the American Psychological Association, and the National Council on Measurement in Education with the intent of providing test developers, administrators, and users with criteria for evaluating both the quality of a test and its appropriate uses. A large component of the standards consists of guidance for evaluating the validity of proposed uses of test scores.

"Testing professionals must lead the way in providing a framework for evaluating proposed systems that purport to measure teacher quality."

In a **2012 research paper**, Lorrie A. Shepard, the dean of the education school at the University of Colorado at Boulder, emphasized that validation requires testing the viability of the assumptions underlying the use of test scores in teacher evaluation.

The following list identifies some of the assumptions that need to be verified as part of a study to ensure that when an evaluation declares a teacher "effective" or "ineffective," the label carries meaning:

- The instruments (e.g., accountability assessments, teacher-observation protocols, student-satisfaction surveys) that make up the teacher-evaluation system are designed to be sensitive to classroom instruction and changes in classroom instruction across a diverse population of students.
- The administration and implementation of the instruments are consistent with their protocols.
- The scoring rules and rubrics used for instruments are appropriate.
- Scores assigned by raters (e.g., peers, principals, students) are accurate, consistent with scoring protocols, and free of bias.
- Observations used in the evaluation are fair, using multiple observers and representing the variety of conditions that could affect teacher performance (e.g., time of year, time of day, subject area covered), so that results are generalizable to teacher performance as a whole.
- The measurement instruments are sufficiently reliable.
- Teacher-evaluation scores do not significantly correlate with variables associated with the students they teach (e.g., English-language proficiency, prior performance on content, free or reduced-price lunch status). That is, the instruments address factors that can be changed by the teacher.
- The instrument outcomes are related to the desired traits (e.g., those exhibited in classrooms that differentiate between higher- and lower-quality teachers).
- Teachers with higher scores are more effective than teachers with lower scores.
- Raters are able to appropriately assess teacher performance.

Some of these assumptions are easy to test, and data supporting them may already be available. Gathering and analyzing data for other assumptions will require more creative research designs.

Also critical is the evaluation of assumptions related to consequences of policy implementation. For example, policies concerning the use of teacher-evaluation measures typically rest on assumptions that decisionmakers understand and can effectively interpret and use the measures to select teachers for rewards, sanctions, and additional professional development, and that pay-for-performance incentives would increase teacher quality.

Likewise, undesirable consequences need to be explored and vetted for their impact. For example, personal concern for evaluation results and their associated rewards or sanctions may discourage teachers from accepting teaching assignments for specific student populations; or the number of effective teachers may be inadequate to replenish those who are removed through sanctions or who retire in discouragement from the teaching profession.

Most importantly, the public's and the education profession's trust in the labels placed on teachers is vital in enhancing the quality of education in the classroom. (On this matter, **a lawsuit was recently filed in Tennessee** over the state's value-added

MORE OPINION

teacher-evaluation system, which relies on student test scores.) Ultimately, we need to gather evidence to support these labels and address possible consequences.

Visit Opinion.

We plead: Evaluate the validity of claims made about teacher quality before moving forward. We now have an extra year granted to us by the Department of Education. We need to take this time to conduct essential validity studies for the sake of true accountability, student learning, and a just educational measurement system.

Tia Sukin is a senior psychometrician for Pacific Metrics, in Monterey, Calif. W. Alan Nicewander is the chief psychometrician at Pacific Metrics and an associate editor of Psychometrika, the journal of the Psychometric Society. Phoebe Winter was the executive vice president for education policy at Pacific Metrics when this Commentary was written. She has since retired. Howard Mitzel, who died in January, had been the president and principal founder of Pacific Metrics and retained a seat on its board at the time this Commentary was written. Lisa Keller is an associate professor of psychometric methods at the University of Massachusetts Amherst. Matthew Schulz is the vice president of research at Pacific Metrics.

Vol. 33, Issue 27, Pages 28-29