

Fine-Tuning Teacher Evaluation

by Kim Marshall

Classroom observations, student achievement, and feedback from students are important, but they'll only improve education if they're used wisely.

As many states and districts rethink teacher supervision and evaluation, the team at the Measures of Effective Teaching (MET) Project, funded by the Bill and Melinda Gates Foundation, has analyzed thousands of lesson videotapes and studied the shortcomings of current practices.

The tentative conclusion: Teachers should be evaluated on three factors—classroom observations, student achievement gains, and feedback from students. The use of multiple measures is meant to compensate for the imperfections of each individual measure and produce more accurate and helpful evaluations (Kane & Cantrell, 2012).

This approach makes sense, but its effectiveness will depend largely on how classroom observations, achievement data, and student feedback are used. Here are some suggestions.

Classroom Observations

Not surprisingly, MET researchers have found that one classroom observation a year doesn't give an accurate picture of a teacher's work. They suggest several enhancements: using a good rubric for observations, observing teachers four times a year, having more than one observer evaluate each teacher, and improving administrator training. In my view, these don't deal adequately with the serious design flaws in the conventional teacher evaluation model.

For starters, four evaluation visits a year aren't nearly enough to sample what students experience daily, especially given the fact that most official classroom visits are scheduled in advance. If teachers know their evaluator is coming, they tend to take their performance up a notch, which means evaluators are seeing better-than-normal teaching. In addition, when an administrator walks into a classroom, students usually behave better, which again masks quotidian realities of classroom life. Day-by-day teaching practices are what drive student achievement. If administrators don't see those practices, their evaluations are inaccurate, dishonest in terms of quality assurance, and not helpful for improving mediocre and ineffective teaching practices.

Further, detailed feedback after infrequent, full-lesson observations, because of its inauthenticity and bulk, is a "weak lever" for improving teacher performance (DuFour & Marzano, 2009). Filling out elaborate rubrics after every visit, as the MET study suggests, creates an impossible workload for administrators, leaving less time for informal classroom visits and interactions with teacher teams. School leaders in Tennessee are suffering this fate under the state's cumbersome new evaluation system (Anderson, 2012). Time is a precious commodity for overtaxed school leaders, and they can't afford to spend it on bureaucratic tasks of questionable impact.

Here's a better approach: Ten brief, unannounced classroom visits (10–15 minutes each) by the *same* administrator, sampling multiple aspects of each teacher's work—the beginning, middle, and end of lessons; different subject areas or classes; and different times of the day and week. Each observation is followed promptly by a face-to-face coaching conversation and then a brief write-up (Marshall, 2005, 2009). Having the same administrator conduct all 10 visits avoids the problem of mixed signals that might come from different evaluators and enables the administrator to get a fuller picture of a teacher's performance and how it evolves over a year.

I'm continually amazed at how much a visitor can see in a short period of time—in fact, the challenge is deciding on the two or three most important nuggets to discuss with the teacher afterward. When I coach principals on short observations, I suggest that they not burden themselves with laptops, tablets, smartphones, or elaborate rubrics or checklists. Visitors get the best insights by walking around the room, looking over students' shoulders, checking in with a few students, observing what the teacher is doing, and jotting a few quick notes. A brief mental checklist is helpful as administrators scan the classroom and decide what needs affirmation or feedback. One suggestion can be summed up with the acronym SOTEL: How is the teacher doing with Safety, Objectives, Teaching, Engagement, and Learning?

The Hamilton County Schools in Tennessee received a waiver from Tennessee's statewide teacher evaluation requirements and used short observations over the last two years. Teachers and administrators report significant improvements in the quality and impact of classroom visits, and test scores have gone up (Scales & Atkins, 2011). If this approach, accompanied by good training and supervision of administrators, replaced the traditional dog-and-pony show, classroom observations would be more accurate and make far better contributions to instructional improvement.

Student Achievement

The MET team understands the challenges of using test scores to evaluate teachers and has used a sophisticated combination of assessments in its research. My concern is that in the real world, cash-strapped school districts will use one standardized test for high-stakes teacher evaluations, leading to a variety of unfortunate consequences (Darling-Hammond, Amrein-Beardsley, Haertel, & Rothstein, 2012; David, 2010; Johnson, 2012):

- Standardized tests are not designed to evaluate teachers, which will fuel litigation by teachers who suffer negative job consequences based on test-score data.
- Districts will need to collect three years of value-added scores to reduce "noise" in the data (Goldhaber & Hansen, 2008). Three years is an unacceptable time lag for intervening with ineffective teachers.
- Fear of negative consequences may lead teachers to spend an inordinate amount of time on test prep—and may even increase cheating by ethically challenged teachers and administrators.
- Evaluating individual teachers on the basis of test results can have a negative impact on collegiality and teamwork, which are among the most powerful engines of instructional improvement.
- Usable standardized test data are available for only about 20 percent of teachers, raising equity concerns about how the other 80 percent are evaluated.

- Praising or criticizing individual teachers for their students' test scores fails to take into account the work done by pullout teachers, specialists, tutors, and teachers in previous grades, all of whom contribute to student outcomes.

The conclusion is inescapable: It's highly problematic to use standardized test scores to evaluate teachers. The idea sounds appealing, but it will inevitably hit a brick wall. Fortunately, there is a better way to make student learning part of teacher evaluation. Each teacher team could decide on a valid, locally available assessment tool and get the principal's approval. (For example, a 2nd grade team might pick Fountas-Pinnell benchmarks to assess reading.) The team would do a baseline assessment of every student in September; set a goal (for example, getting at least 85 percent of students to Fountas-Pinnell Level M or above by June); teach, assess, and share ideas throughout the year; assess all students at the end of the year; and present the results to the principal. The principal could look at the data, give the team a collective evaluation on its value-added data, and include the team score as one factor in each teacher's individual evaluation.

This approach puts student learning at the heart of the evaluation process. Formative assessment data can be used throughout the year, and the premium is on teamwork, with underperforming team members getting help and motivation from their colleagues. Administrators and other support personnel can provide feedback, suggestions, and support through frequent visits to classrooms and team meetings. And *all* teachers—including primary grades, art, music, and physical education—can be part of using teamwork and data to continually improve their craft.

Student Input

The MET project makes a convincing case for using student surveys as a third factor in teacher evaluation. Despite many educators' immediate objection (*How can children evaluate adults?*), there's no denying that students are with their teachers hundreds more hours than even the most energetic administrator.

Harvard professor Ronald Ferguson and his Tripod Project colleagues have found that students are quite astute at sizing up teachers' instructional competence when they are asked about observable classroom behaviors in kid-friendly language. Here are some sample questions, using a five-point agree/disagree scale:

- Our class stays busy and does not waste time.
- I understand what I am supposed to be learning in this class.
- If I don't understand something, my teacher explains it another way.
- My teacher pushes everyone to work hard.
- My teacher takes the time to summarize what we learn each day.

Ferguson reports that students taught by teachers with high student ratings achieve a full semester better than students whose teachers get low ratings (Sparks, 2012). But here's a concern: Students sometimes don't appreciate tough, demanding teachers until years later. Could high-stakes student surveys lead teachers to lower their standards? In addition, there's evidence that although high student ratings correlate with high achievement in the year surveyed, they correlate less well with success in the next grade or course (Glenn, 2011).

Fortunately, there's a way to use student input that avoids these problems: Ask all teachers to survey their students anonymously each year using questions like Ferguson's.

At the end of the year, teachers look over their data with the principal and answer questions like these: What pleased you most in this year's survey? What surprised you? What are two changes you'll make in your classroom next year? The evaluation would then be based on how well the teacher responds to the feedback. For example, a teacher might express surprise that students said he talked too fast and sometimes left them confused about new concepts. He might resolve to make a conscious effort to slow his pace and use better methods to check for understanding, including dry-erase boards and clickers.

Making Teacher Evaluation Effective

As states and districts rethink their teacher evaluation policies, I urge them to consider these enhancements to classroom observations, the use of achievement data, and student input. I believe these practices will give teachers a stronger voice, use principals' time more effectively, and make teacher evaluation a real player in dramatically improving teaching and learning.

References

- Anderson, J. (2012, February 20). States try to fix quirks in teacher evaluation. *New York Times*, p. 1.
- Darling-Hammond, L., Amrein-Beardsley, A., Haertel, E., & Rothstein, J. (2012). Evaluating teacher evaluation. *Phi Delta Kappan*, 93(6), 8–15.
- David, J. (2010). [Using value-added measures to evaluate teachers](#). *Educational Leadership*, 67(8), 81–82.
- DuFour, R., & Marzano, R. J. (2009). High-level strategies for principal leadership, *Educational Leadership*, 66(5), 62–68.
- Glenn, D. (2011). One measure of a professor: Students' grades in later courses. *Chronicle of Higher Education*, 57(19), A8–A9.
- Goldhaber, D., & Hansen, M. (2008). *Assessing the potential of using value-added estimates of teacher job performance for making tenure decisions*. Washington, DC: National Center for Analysis of Longitudinal Data in Educational Research.
- Johnson, S. (2012). Having it both ways: Building the capacity of individual teachers and their schools. *Harvard Educational Review*, 82(1), 107–122.
- Kane, T., & Cantrell, S. (2012). *Learning about teaching: Initial findings from the measures of effective teaching project*. Seattle, WA: Bill and Melinda Gates Foundation.
- Marshall, K. (2005). It's time to rethink teacher supervision and evaluation. *Phi Delta Kappan*, 86(10), 727–735.
- Marshall, K. (2009). *Rethinking teacher supervision and evaluation*. Hoboken, NJ: Jossey-Bass.
- Scales, J., & Atkins, C. (2011). Hamilton County Department of Education: Rethinking teacher evaluation through project COACH. *District Management Journal*, 7, 12–21.
- Sparks, S. (2012, April 25). MET studies seek more nuanced look at teaching quality. *Education Week*, 31(29), 12.

[Kim Marshall](#), a former teacher, principal, and central-office leader in Boston, is the author of *Rethinking Teacher Supervision and Evaluation* (Jossey-Bass, 2009; 2nd ed., in press).