ISSUE
ANALYSIS
REPORT

TNTP reimagine teaching

‘MET’ MADE SIMPLE

BUILDING RESEARCH-BASED TEACHER EVALUATIONS

'MET' Made Simple: Building Research-Based Teacher Evaluations

New findings from the Gates Foundation's Measures of Effective Teaching (MET) project can help policymakers develop research-based evaluation systems that could unleash the untapped potential in the nation's teaching force.

Introduction

There is no shortage of research on the importance of good teaching. For decades, study after study has shown that there are large differences in effectiveness from one teacher to another and that these differences can have a lifelong impact on students. A recent study that tracked 2.5 million students over 20 years determined that those with highly effective teachers "are more likely to attend college, earn higher salaries, live in better neighborhoods, and save more for retirement. They are also less likely to have children as teenagers."[1]

Yet there has been little research on exactly how schools can get an accurate picture of their teachers' performance in the classroom. States and school districts have been left largely to their own devices when it comes to this singularly important task.

The results have been disastrous. As we documented in our 2009 study *The Widget Effect: Our National Failure to Acknowledge and Act on Differences in Teacher Effectiveness*, most teachers are evaluated infrequently and according to low standards. They rarely receive feedback that helps them improve. Nearly every teacher is labeled "good" or "great," no matter how much progress their students are making. In the end, the entire profession has suffered from this negligent approach.

Groundbreaking new findings from the Bill and Melinda Gates Foundation's Measures of Effective Teaching (MET) project hold the potential to answer crucial questions about how to assess teachers' performance.[2] For the past two years, MET researchers have conducted a research project of unprecedented scope, involving 3,000 teachers in six school districts across the country. Using gold standard research methods, they have tested a number of evaluation approaches, including student achievement data, classroom observations, and surveys of students. Their most recent report, "Gathering Feedback for Teaching," provides a wealth of practical implications for improving teacher evaluations.

This paper is intended for policymakers who are developing better teacher evaluations and are looking for ways to apply new research findings quickly. It summarizes the lessons from MET and provides recommendations on how these lessons can be applied right now.

Our perspective is informed by a careful review of the MET findings as well as our direct experience designing and implementing evaluation systems in states and urban school districts across the country. It is the best of what we know at this moment. We do not pretend that there are easy answers to every challenge, or that our ideas will always be right. However, MET is shedding light on questions that have been poorly understood until now. As educators, we have a responsibility and opportunity to put this valuable new information to use.

[1] Chetty, Raj; Friedman, John; and Rockoff, Jonah. (2011). "The Long-Term Impacts of Teachers: Teacher Value-Added and Student Outcomes in Adulthood." National Bureau of Economic Research Working Paper No. 17699, December 2011.

[2] *Disclosure:* The Bill and Melinda Gates Foundation is one of several major philanthropies that provide funding support to TNTP.

The New MET Report: Four Key Lessons

Lesson #1: Teachers generally appear to be managing their classrooms well, but are struggling with fundamental instructional skills.

The Widget Effect and countless other studies have shown that assigning high ratings to nearly all teachers—as most current teacher evaluation systems do—ignores wide variations in effectiveness and stunts teachers' growth by giving them a false picture of their performance. The MET study suggests that decades of ratings inflation have come at a high price.

As they tested several commonly used classroom observation rubrics in multiple districts, the MET researchers were able to rate the classroom performance of a significant sample of teachers across the country (1,333). The trends were clear. Across multiple rubrics and in multiple districts, carefully trained raters reported similar findings: a majority of teachers had mastered basic classroom management skills but struggled with more advanced instructional skills.

For example, nearly three-quarters of teachers observed using the Danielson Framework for Teaching were rated proficient or higher at "managing classroom behavior," and more than half were proficient or distinguished at "managing classroom procedures." But only about one-third were rated at least proficient in "using questioning and discussion techniques," and less than one-third were proficient or better in "communicating with students"—instructional skills that are essential to helping students master the content of a lesson.[3] The findings were strikingly similar on four other rubrics tested in the project.

Consider these findings for a moment. Among a sample of more than 1,000 teachers, only about one in three was able to lead a classroom discussion or communicate with students at the level defined as "proficient" by this rubric. In the remaining classrooms, teachers performed at either a "basic" or "ineffective" level.

> Better evaluations can give teachers a real opportunity to reach their full potential in the classroom.

These results mirror the findings of another recent study by the Consortium on Chicago School Research[4] and stand in stark contrast to the actual ratings assigned to teachers in most districts—even some districts that have already put more rigorous evaluation systems in place. The gap between official teacher evaluation ratings and actual classroom performance appears to be shockingly wide. Education leaders and policymakers should anticipate that any accurate new evaluation system will result in fairly large downward corrections in ratings.

---

[3] "Gathering Feedback for Teaching: Combining High-Quality Observations with Student Surveys and Achievement Gains." (2012). The Bill and Melinda Gates Foundation. Page 26.

[4] Sartain, Lauren; Stoelinga, Sara Ray; and Brown, Eric R. (2011). "Rethinking Teacher Evaluation in Chicago: Lessons Learned from Classroom Observations, Principal-Teacher Conferences and District Implementation." Consortium on Chicago School Research at the University of Chicago Urban Education Institute. The researchers found that, among a random sample of 280 teachers across Chicago Public Schools, the percentage receiving at least one "unsatisfactory" rating jumped from the district's historic 0.3 percent rate to 8 percent after the adoption of a new teacher evaluation system, including observations that followed the Framework for Teaching rubric. About three-quarters of teachers received positive ratings for classroom management skills; just over half were highly rated on instructional skills.

This is a sobering reality, but it points to the enormous untapped potential in the current teacher workforce. The problem is not that teachers are not working hard enough or are incapable of mastering advanced instructional skills. Inflated evaluation ratings are a symptom of systemic neglect of teachers over many decades: a failure to set clear, rigorous expectations about good instruction, a failure to give teachers the feedback and support they need to meet those expectations, and a failure to be honest with teachers when they fall short. Better evaluations that give teachers honest feedback and meaningful development opportunities represent the first step toward ending this pattern of neglect, and can give more teachers a real opportunity to reach their full potential in the classroom.

Lesson #2: Classroom observations can give teachers valuable feedback, but are of limited value for predicting future performance.

As the findings above demonstrate, classroom observations can say a lot about the overall performance of a large group of teachers. But what can they tell us about the performance of a single teacher?

MET researchers found that it is extremely difficult to evaluate individual teachers accurately using classroom observations alone. The researchers used several different rubrics, ensured that all evaluators were carefully trained and had passed accuracy screens, and conducted multiple observations of each teacher. Under these ideal circumstances, they found that observation ratings did correlate somewhat with student achievement data.[5] But no matter what they tried, observation ratings alone were not very predictive of a teacher's future success at helping students learn.[6]

That's because classroom observations have an inherent limitation: They capture only a few short snapshots of a teacher's performance during the course of a long academic year. This is a problem, because the MET researchers confirmed that classroom performance really does vary from day to day, just as teachers have said for years. Researchers also found that even well-trained observers sometimes disagreed on how to rate a particular lesson. More observations by more observers can mitigate some of these problems, but classroom observations will always provide incomplete information on their own.

> Classroom observations will always provide incomplete information on their own.

This is a crucial insight, because most school systems evaluate teachers primarily or exclusively based on classroom observations. In other words, most schools are using an approach that is likely to produce inaccurate judgments about teacher performance. Researchers did find some meaningful correlation between observation ratings and student learning, but not enough to use observations as the sole factor in performance evaluations.

This does not mean policymakers should abandon observations, of course. They are a critical part of any evaluation and development system, because they help evaluators identify teachers' specific strengths and weaknesses in the classroom—which enables them to give honest feedback that can help teachers improve.

[5] "Gathering Feedback for Teaching," Page 7.
[6] "Gathering Feedback for Teaching," Page 29.

Lesson #3: "Value-added" analysis is more powerful than any other single measure in predicting a teacher's long-term contributions to student success.

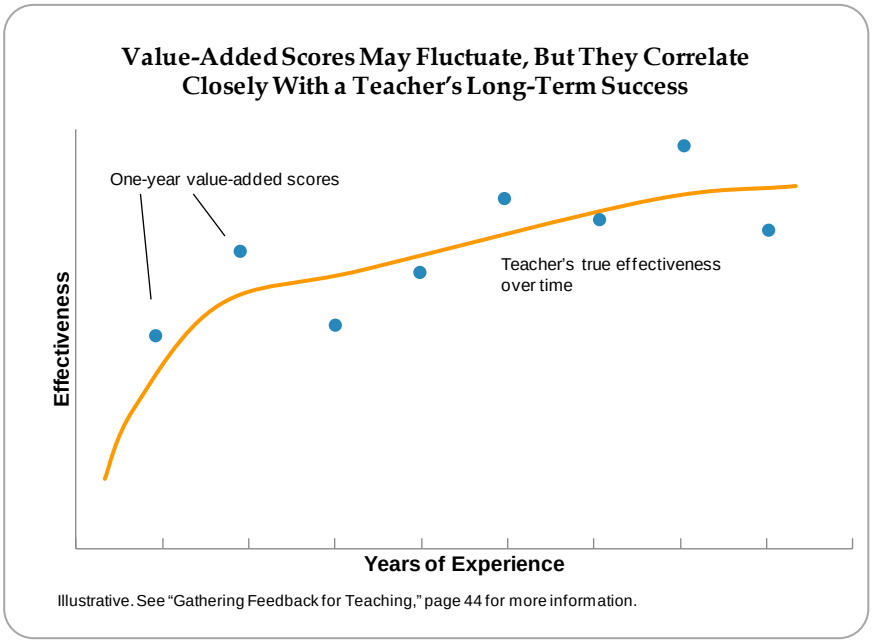
MET researchers found that value-added analysis, which typically uses test results to gauge how much an individual teacher contributes to his or her students' learning growth, is more accurate than any other single measure in predicting success over the course of a teacher's career—more than classroom observations or student surveys.[7]

This finding is bound to be controversial in some corners. Value-added methodology generates great debate because it relies on standardized tests and because it yields only an estimate of teacher performance, not an exact measure.

However, the MET findings make a very strong case that although value-added scores are not perfect (no measure is), they tell us a great deal about how teachers will likely perform in the future. In addition, the findings debunk two common myths.

First, researchers found that high value-added scores are not associated with a "drill-and-kill" approach to instruction. Teachers with high value-added scores helped their students master higher-level thinking skills in addition to helping them score well on traditional standardized tests.[8] And in surveys, students of high value-added teachers reported enjoying school more and trying harder on their classwork.[9] In other words, good teaching is good teaching. Teachers are not generally earning high value-added scores by teaching to the test.

Second, the report also shows that year-to-year fluctuations in an individual teacher's value-added scores should not keep us from using them in evaluations. Although a teacher's score might vary somewhat from year to year, the relationship between any individual year's score and the teacher's long-term success is quite strong. [10]



A good analogy is a baseball player's batting average: it will be higher in some years than in others, but each individual year correlates well to the player's career average. The findings suggest that we should think of a teacher's effectiveness as a trend line, not a fixed data point.

[7] "Gathering Feedback for Teaching," Page 9.
[8] "Gathering Feedback for Teaching," Page 12.
[9] "Gathering Feedback for Teaching," Page 12.
[10] "Gathering Feedback for Teaching," Page 44.

These findings are all the more powerful when viewed in conjunction with newly released data from researchers at Harvard and Columbia Universities showing that teachers with high value-added scores have a major and enduring influence on their students' life outcomes, from their likelihood of going to college to saving for retirement.[11] Both studies suggest that a high value-added score is a strong indicator of a great teacher.

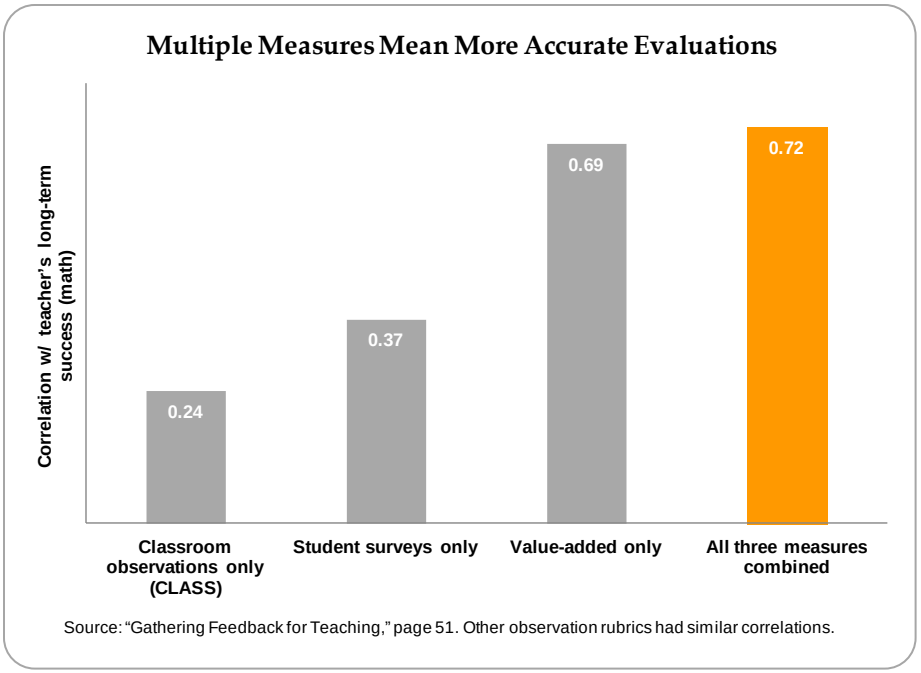
Bottom line: Evaluations should not rely on value-added scores alone, because no single measure can tell the full story of a teacher's performance. But including value-added data makes results significantly more accurate over time, not less.

Lesson #4: Evaluations that combine several strong performance measures will produce the most accurate results.

Although there is strong evidence that value-added results are especially powerful measures of teacher effectiveness, many organizations (including TNTP) have recommended a "multiple measures" approach to teacher evaluations. It is a common-sense idea: No single measure can tell the full story about a teacher's performance, so schools should consider all the information at their disposal. Moreover, evaluation is not just about sorting teachers but providing them with useful feedback and support, which requires direct observations of teachers at work. Nearly every state and school district currently revamping their teacher evaluations is creating a system that uses multiple measures.[12]

MET researchers have confirmed that this is the right approach. They found that evaluations were most accurate when they combined value-added data with rigorous classroom observations and surveys of student perceptions.

If a school wants to predict a teacher's future success in helping students learn, multiple measures will yield the most accurate results—more accurate than any one measure on its own.[13]



Source: "Gathering Feedback for Teaching," page 51. Other observation rubrics had similar correlations.

[11] Chetty et al., 2011.
[12] National Council on Teacher Quality. 2011. *State of the States.*
[13] "Gathering Feedback for Teaching," Page 9.

What to Do Now: Five Recommendations for Policymakers

The new MET project report points to several ways policymakers can ensure the success of their new evaluation systems—many of which have been confirmed by TNTP's firsthand experience in states and school districts across the country.

Recommendation #1: Base teacher evaluations on multiple measures of performance, including data on student academic progress.

Educators who are already building evaluations that include classroom observations and student learning data should continue down that path with heightened confidence and a renewed sense of urgency. Those who are planning or implementing systems based solely on classroom observations—without any measures of student learning—should reconsider. If a goal of evaluating teachers is to ensure student learning, then student learning must be a major part of what's measured. In subjects and grades where common assessments of student learning are not yet available, states and districts should prioritize the development of high-quality assessment tools that will provide data for both instructional and evaluative purposes.

Recommendation #2: Improve classroom observations by making them more frequent and robust.

When it comes to classroom observations, the MET study is clear that more eyes—in the form of more observers and more frequent observations—lead to more telling results.[14] A single observation by a single rater is unreliable as a performance measure. For the many districts where this practice is the norm, it's time to change course.

Instead, states and school districts should require raters to visit teachers' classrooms more often. Where possible, they should help schools reallocate resources and assign different raters for different visits.[15] Creative solutions are possible. For example, observations need not last an entire class period. MET researchers found that even a 15-minute observation can be just as meaningful with some observation rubrics.

Researchers found that four observations by four different raters had the strongest correlation to student learning results, but they also note that this shouldn't be taken as a hard-and-fast requirement. Ratings became more reliable with each additional observation, so anything more than a single classroom visit per year is an incremental improvement. The more observations and the more observers, the better.

[14] "Gathering Feedback for Teachers," Page 37.
[15] New evaluation systems offer examples of how to grow the rater pool. The New Haven Public Schools now includes external validators, and DC Public Schools has designated master teachers as additional evaluators.

Recommendation #3: Use or modify an existing observation rubric instead of trying to reinvent the wheel.

MET researchers tested several different observation rubrics and did not find any that distinguished themselves meaningfully from the rest. There is no such thing as a perfect observation tool, and no single observation can provide a complete picture of a teacher's performance under any rubric.

All of this suggests that states and districts should pick an existing rubric that meets their needs instead of designing one from scratch—especially now that several have been tested thoroughly.

This recommendation won't be easy to follow. Educators will likely want to make modifications to an off-the-shelf rubric to reflect their local context, and this can be a good opportunity to involve teachers and principals in the design process. Certainly, investment in and understanding of these tools is critical. But policymakers should keep in mind that there are diminishing returns to time spent on design —and that time will be better spent on communicating shared definitions of teaching excellence and bringing new systems online. Implementation matters most, and the bulk of time and available resources should be devoted to ensuring that the rubric is used well and consistently.

Choosing an Observation Rubric: Five Key Questions

1. Does the rubric focus on the competencies most connected to student outcomes?
2. Does the rubric set high performance expectations for teachers?
3. Is the rubric clear and precise?
4. Is the rubric student-centered, requiring observers to look for direct evidence of student learning?
5. Is the rubric concise enough for teachers and observers to understand thoroughly and use easily?

For more information, see TNTP's 2011 report, "Rating a Teacher Observation Tool."

Recommendation #4: Give evaluators the training and ongoing support they need to be successful.

Observation rubrics are only as good as the observers who use them. Districts need to provide initial training to ensure that evaluators can conduct accurate observations and provide useful feedback to teachers.

For example, in the MET project, evaluators received 17 to 25 hours of training. They then had to rate a series of videotaped lessons, and were only approved as evaluators if their ratings were close to those of expert observers. This level of preparation may be difficult for school districts to achieve, but it can guide their approach to training observers. For example, schools could easily adopt the low-cost, high-impact strategy of using videotaped lessons to ensure that evaluators' judgment is reasonably consistent.

The MET researchers also make it clear that upfront training is not enough. States and districts will need to monitor the results of evaluations to make sure they are producing accurate ratings. This means collecting and analyzing data on observation ratings and overall evaluation ratings, as well as conducting regular "spot checks" on observers. It may also mean checking to ensure that observation ratings

correlate with student achievement results. These efforts will require states to invest in high-quality data systems that can track and analyze evaluation results.

To be clear: delivering perfunctory training to administrators and sending them out to evaluate teachers is unlikely to yield accurate information. If we want better results, we need to invest in training.

Recommendation #5: Strongly consider using student surveys as a component of teacher evaluation.

Researchers found a strong correlation between positive student learning outcomes and surveys in which a majority of students described the learning environment as focused, engaging and demanding. Students with *effective* – not necessarily "easy" – teachers are more likely to say they feel happy and believe their teacher makes good use of class time. In short, researchers wrote, "the average student knows effective teaching when he or she experiences it."[16]

Student survey results correlated as strongly with student learning as classroom observations did. They were also a more reliable measure than observations—which makes sense considering that students see every lesson a teacher presents during the school year, compared to the handful observed by evaluators. These findings suggest that surveys could be especially useful for evaluating teachers whose students do not take standardized tests, and who therefore will not have a value-added score. Coupling student surveys with strong classroom observations would ensure that no single measure determines a teacher's overall evaluation rating. Surveys may act as a "check" on classroom observation ratings, much like value-added analysis can act as a "check" on observations in tested subjects.

Student surveys do not come close to matching the power of value-added data in predicting a teacher's future success, though. States and districts that opt to use student surveys as performance measures in non-tested grades and subjects should still make it a priority to find or develop accurate, objective measures of student learning that can apply to as many teachers as possible.

Multiple Measures for Better Evaluations

According to MET researchers, teacher evaluations should include:

- **Student achievement data**, such as value-added analysis, to identify and predict teachers' ability to help students learn.
- **Classroom observations**, to diagnose strengths and weaknesses in classroom practice.
- **Student surveys**, which can add reliability to evaluation ratings and assess classroom culture.

[16] "Learning About Teaching: Initial Findings from the Measures of Effective Teaching Project." (2011). The Bill and Melinda Gates Foundation. Page 5.

Conclusion

The MET findings make clear that we do not have to settle for the flimsy teacher evaluations of the past. We can gather an accurate picture of a teacher's work by combining several strong measures that exist today. In particular, where it is possible to incorporate value-added results, we should not delay. But we also need to admit two things.

First, even though infrequent classroom visits are the predominant approach for assessing teachers, they are woefully insufficient. We need to change.

Second, we will never arrive at a perfect measure. Teaching is complex and multifaceted; it draws on a broad array of professional skills. And a teacher's performance is not static from day to day or from year to year. Just as in other professions, our goal should be a fair, consistent approach to evaluation that gives schools and teachers the best possible information.

As research increasingly confirms the usefulness and reliability of value-added analysis and a multiple measures approach to evaluation, the focus shifts now to the school level, where we have an opportunity to translate these findings into better practice. We hope this paper helps to hasten progress toward that critical goal—and toward the goal of ensuring that all students learn from effective teachers every day.

About TNTP

TNTP strives to end the injustice of educational inequality by providing excellent teachers to the students who need them most and by advancing policies and practices that ensure effective teaching in every classroom. A national nonprofit organization founded by teachers, TNTP is driven by the knowledge that effective teachers have a greater impact on student achievement than any other school factor. In response, TNTP develops customized programs and policy interventions that enable education leaders to find, develop and keep great teachers. Since its inception in 1997, TNTP has recruited or trained approximately 49,000 teachers – mainly through its highly selective Teaching Fellows programs – benefiting an estimated 8 million students. TNTP has also released a series of acclaimed studies of the policies and practices that affect the quality of the nation's teacher workforce, including *The Widget Effect* (2009) and *Teacher Evaluation 2.0* (2010). Today TNTP is active in more than 25 cities. For more information, visit www.tntp.org.