THE MET PROJECT

The Wrong $45 Million Question

By equating teacher effectiveness with teacher effects on student test scores, the Measures of Effective Teaching project fails to address what we value most in education.

Rachael Gabriel and Richard Allington

In 2009, the Bill and Melinda Gates Foundation funded the investigation of a $45 million question: How can we identify and develop effective teaching? Now that the findings from their Measures of Effective Teaching (MET) project have been released (see "Initial Findings from the MET Project," p. 47), it's clear they asked a simpler question, namely, What other measures match up well with value-added data?

Although we don't question the utility of using evidence of student learning to inform teacher development, we suggest that a better question would not assume that value-added scores are the only existing knowledge about effectiveness in teaching. Rather, a good question would build on existing research and investigate how to increase the amount and intensity of effective instruction.

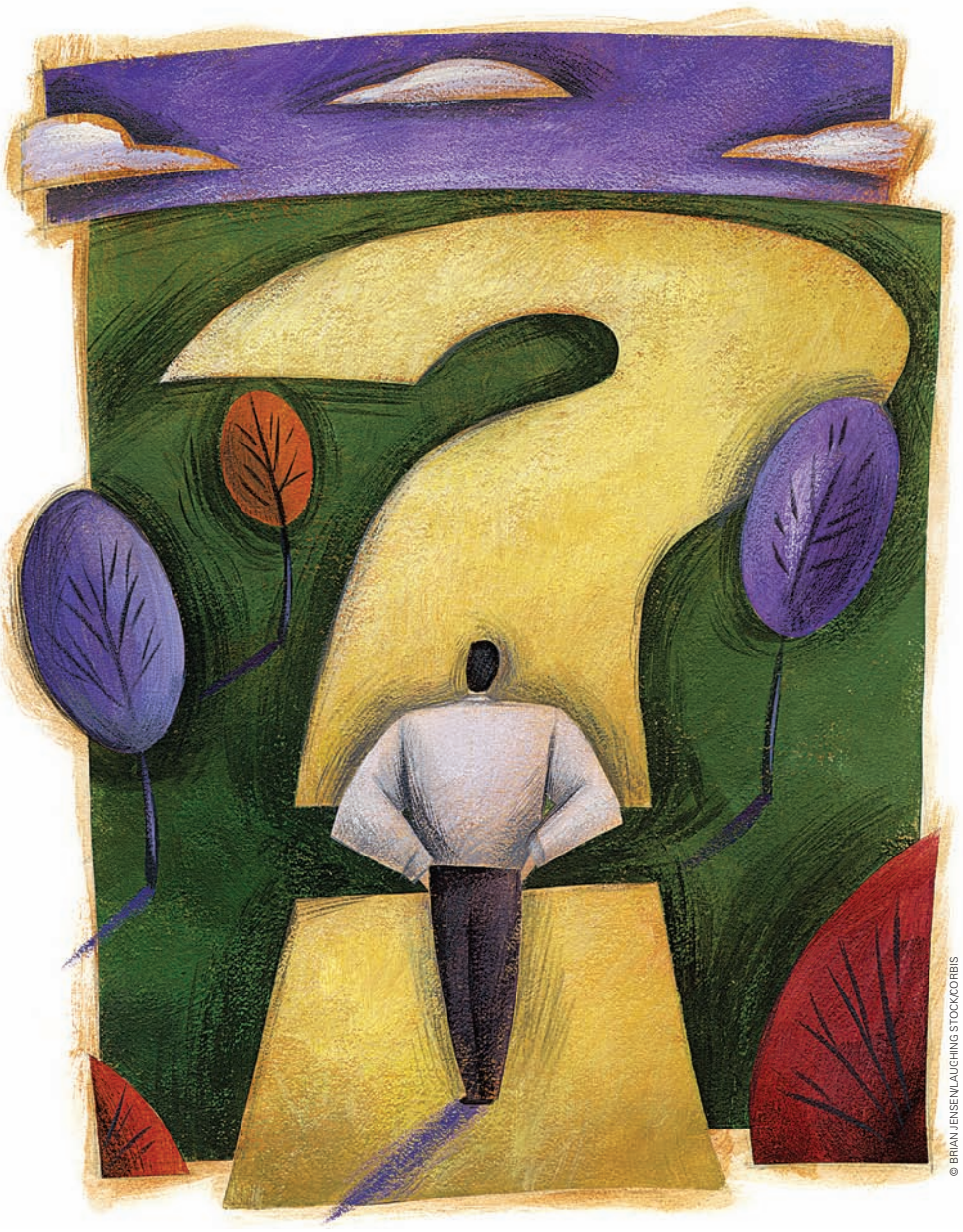
A Narrow View of Effectiveness

In a 2011 *Wall Street Journal* editorial, Bill and Melinda Gates suggest that the field of education has abdicated responsibility for defining effective teaching, thus the need for a different approach:

> It may surprise you—it was certainly surprising to us—but the field of education doesn't know very much at all about effective teaching. . . . This ignorance has serious ramifications. We can't give teachers the right kind of support because there's no way to distinguish the right kind from the wrong kind. We can't evaluate teaching because we are not consistent in what we're looking for. We can't spread best practices because we can't capture them in the first place.

The Gates Foundation has outfunded the U.S. Department of Education on studies of teacher effectiveness at a rate of about 40 to 1, yet the MET project is not the first attempt to understand and measure effectiveness. Like the MET project, studies of effective teaching from the last 50-plus years have often combined both quantitative and qualitative analyses of classroom observations, teacher interviews, and student test scores on a variety of measures. Studies conducted in the mid-1990s had budgets of more than $1 million; the MET project is the first to reach into the tens of millions of dollars.

Perhaps the largest difference between the MET project's findings and existing research on effectiveness is that effectiveness has only recently been conflated with teacher effects on student test scores. Although those involved in the MET project take pains to assure the public that they only look at value-added data, or teacher effect scores, in combination with other measures, they



© BRIAN JENSEN/LAUGHING STOCK/CORBIS

also explain that value-added scores are the gold standard by which they vet all other comparisons. Their initial report (2010) lists as a fundamental premise that "any additional components of the evaluation (e.g., classroom observations, student feedback) should be demonstrably related to student achievement gains" (p. 5).

Almost 50 years ago, Bond and Dykstra (1967) published the findings of a study with similar landmark status in its day as the MET project has today. The researchers combined the efforts of 35 research teams to identify the most effective beginning reading instruction in upward of 300 1st grade classrooms across the United States. Their main finding was that effective reading could not be distilled into one monolithic set of indicators or best practices; instead, there were many—sometimes contradictory—successful approaches.

The no "one right way" finding has held true across federally funded, large-scale national studies of exemplary 1st grade teachers (Pressley, Allington, Wharton-McDonald, Block, & Morrow, 2001); 4th grade teachers (Allington & Johnston, 2002); elementary school teachers (Knapp et al., 1995); middle and high school teachers (Langer, 2002); and 800 1st grade teachers in a study conducted by the National Institute of Child Health and Human Development (Stuhlman & Pianta, 2009). This is because exemplary teaching looks and sounds different across different classrooms and contexts.

Instead of asking how to measure effectiveness, the **Met Project asks** a simpler question: **What other measures match** up well with **value-added data?**

The Supports That Count

Let's take a personal case. One of us—Richard Allington—used to believe that expertise in the mechanics of teaching was the crucial factor in effective teachers, but that was before his children entered school. He used to tell preservice teachers that he didn't care whether they loved his children, that he wanted teachers who could *teach* his children, and that he would provide love at home. However, once his children entered school, he changed his mind and his message.

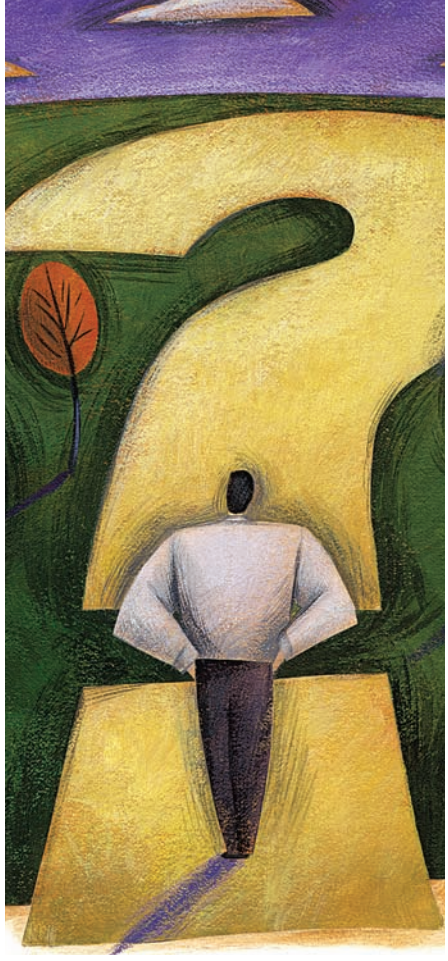
What he wants now are teachers who love his children, and he'll teach his children at home if necessary (and sometimes it has been). This change

in stance was inspired by noting how differently his children achieved when they were in classrooms where they felt loved and respected as children and learners. If the teacher viewed a child as "problematic" (too active, too shy, too aggressive, and so on), learning suffered, as did the child's enthusiasm for school.

This issue of respecting and appreciating students appears in research on teacher effectiveness (Stuhlman & Pianta, 2009), which found that most primary grade teachers did, in fact, demonstrate respect for their students and provide positive socioemotional support. But fewer than a quarter of the teachers studied provided both socioemotional *and* high-quality academic and cognitive support. Of particular concern, children from low-income families were the most likely to be taught by teachers who were rated as providing neither high-quality instruction nor positive emotional support.

Rather than interrogating what counts as evidence of "socioemotional support" or "academic and cognitive support," the MET project simply incorporated Pianta and associates' teacher observation tool—the Classroom Assessment Scoring System (CLASS)—in its own study. (The CLASS tool assesses interactions between teachers and children in three areas: emotional support, classroom organization, and instructional support.) MET researchers then compared the data they collected using this tool with value-added data, using one to validate the other.

The developers of the CLASS tool (Pianta, La Paro, & Hamre, 2008) acknowledge that it's grounded in specific "developmental theory and research" that suggest "that interactions between students and adults are the primary mechanism of student development and learning" (p. 1). It's also grounded in a specific set of values around the kinds of interactions teachers should have with students, such as smiling, laughing, and showing enthusiasm as evidence of a positive classroom climate.



Are students experiencing the education we hope for them? How do we know, and how can we help?

The tool is one of many possible ways to view and evaluate socioemotional and academic support, and it may have more or less validity and instructional utility, depending on a community's philosophies and values. In other words, it may identify some ways of being effective—such as extroverted displays of enthusiasm—but exclude others—such as a quieter manner of displaying support and encouragement. Or it may draw attention to some things that effective teachers have in common but perhaps not to the most important aspects of their practice.

Five Questions to Ponder

The following questions get to the heart of any discussion about effective teacher evaluation tools.

Do evaluation tools inspire responsive teaching or defensive conformity?

Most KIPP schools would get consistently low ratings on the K–3 CLASS assessment because they often require students to SLANT (Sit up, Listen, Ask questions, Nod, and Track the speaker with your eyes). To some, this constitutes physical alertness, respect, and good posture for learning. To others, it constitutes unnecessary restriction of students' freedom of movement and disregard for student perspectives, both of which can limit learning. To us, it constitutes one of many possible value-laden indicators that schools use to estimate effectiveness but that do not, in themselves, constitute effectiveness.

At the end of the day, we don't care whether teachers ask students to SLANT or stand on their heads. Students could learn in both positions, although neither will be appropriate for all students or lessons. The educational value behind such indicators is rooted in the idea that there's a physical aspect to learning and that student engagement is important to learning. However, students will display these behaviors differently across different settings.

Do evaluation tools reflect our goals for public education?

The experience of students, not the ideology of adults, should guide our interpretation of what goes on in classrooms. This is the bigger picture we would observe if we could lift our eyes from lists of indicators and see whether classroom practice actually reflects the education we want for our children.

If we focus too closely on indicators of what we value, rather than on the value itself, we might miss cases where

a teacher uses physical control to create oppressive compliance rather than learning. We would argue that unintended effects are more frequent when teachers perform specified behaviors for the purpose of meeting evaluation requirements rather than as expressions of their professional judgment, inquiry, and reflection. Lessons from the unintended effects of ranking hospitals (more unnecessary surgeries on healthy patients, fewer necessary surgeries on high-risk patients) are evidence that evaluation tools with a focus on indicators (success rates) rather than values (excellent treatment for all) can backfire (Coy, 2002).

Many would argue that minor flaws in lists of indicators come out in the wash as scores are averaged and accumulated, and therefore we should ignore them. Similarly, Kane (2012) has suggested that we should view each of the multiple measures used in the MET project as a superhero with both special powers and a fatal flaw. He argues that instead of rejecting any single measure, we should combine their forces to create a more reliable bundle of tools. Yet when we use these flawed measures to evaluate teachers, they become expressions of what matters in teaching (Darling-Hammond, 1990); we sew the flaws and biases into the fabric of everyday classroom interactions.

Instead of asking which measures correlate most closely with value-added data or whether a rubric based on a certain philosophy is good at predicting value-added scores, we suggest a different set of questions—ones that more wholly reflect the education we want for our children. We acknowledge, however, that it's unlikely these questions would gain any traction in media or policy debates for one simple reason: The United States' current goals for public education are all written in terms of scores or concepts defined by scores.

When we frame education as an economic imperative and the rhetoric involves international comparisons, we measure the goal of education in the form of scores on the National Assessment of Educational Progress (NAEP) or Programme for International Student Assessment (PISA). The Obama administration has used the term "student growth" as a more palatable way of saying test scores compared over time. "College and career readiness" is often a nicer way of saying SAT scores and grade point averages. As long as we define the purpose of public education by scores, we'll define teacher effectiveness as nothing other than a teacher's effect on a test score. There can be no other measures of teaching while there are no other articulated goals for learning.

Do evaluation tools encourage teachers to use text in meaningful ways?

Researchers have identified a relatively stable set of essential practices (see Allington & Johnston, 2002; Duke & Pearson, 2002; Pressley et al., 2003) that develop literate thought, one longstanding goal of instruction. If we cast the development of literate thinkers as the goal of education, we would stop using multiple-choice tests to measure it. We would instead engage students in authentic literacy tasks and observe their attempts to use and develop literacy skills in authentic ways.

For example, borrowing from a national assessment used in New Zealand, we might provide a group of students with a stack of unfamiliar texts and ask them to review and rate them as though they were the acquisitions board for the classroom's library. Teachers would observe or videotape the students' discussion and identify evidence of reading for meaning and enjoyment, building an argument, managing a civil debate, and considering others' perspectives (Johnston, 2005).

If we were measuring and therefore defining student growth in more authentic

Initial Findings from the MET Project

Three Premises

- Whenever feasible, a teacher's evaluation should include his or her students' achievement gains.
- Any additional components of the evaluation (for example, classroom observations, student feedback) should be demonstrably related to student achievement gains.
- The measure should include feedback on specific aspects of a teacher's practice to support teacher growth and development.

Four Findings

- In every grade and subject, a teacher's past track record on value-added measures is among the strongest predictors of their students' achievement gains in other classes and academic years.
- Teachers with high value-added scores on state tests tend to promote deeper conceptual understanding as well.
- Teachers have larger effects on math achievement than on achievement in reading or English language arts, at least as measured on state assessments.
- Student perceptions of a given teacher's strengths and weaknesses are consistent across the different groups of students they teach.

Source: From *Learning About Teaching: Initial Findings from the Measures of Effective Teaching Project* by the Bill and Melinda Gates Foundation, 2010. Retrieved from www.gatesfoundation.org/college-ready-education/Documents/preliminary-findings-research-paper.pdf

ways, we would be more likely to measure teaching in terms of how frequently and successfully teachers orchestrated learning opportunities that lead to the development of literate thought and avoided practices that never have and never will foster thinking—much less literate thinking—such as worksheets, standardized test preparation, or disregard for student engagement or text difficulty.

In interviews, we could evaluate the extent to which teachers are aware of their students' reading preferences, strengths, and weaknesses and the degree to which they plan to address and build on these. In observations, we could evaluate the amount of class time organized around meaningful and authentic uses of texts, relevant topics, and literate conversation. In surveys, we could evaluate how often students feel engaged, interested, and valued when they work with and for their teachers. We could use evidence of meaningful and successful uses of texts to indicate student achievement.

In addition, we could use existing observation rubrics, minus the detailed indicators that are meant to make effectiveness more observable and ratings more reliable, to generate questions about the things we value most about individual classrooms. For example, we might ask evaluators whether they can identify *any* evidence that students are engaged: Are students reading voluntarily outside of class? Do they talk about what they're learning? Do they participate in class discussions? We could triangulate the evaluator's opinion on the authenticity of engagement (rather than the presence of items on a list of indicators) with those of teachers, students, and perhaps other sources (parents, supervisors, or colleagues) to ensure that one principal's version of what engagement looks like doesn't preclude other possibilities.

Thus, the definition of engagement would be shared and meaningful to the school community. This practice would prevent evaluators from mistaking quiet compliance as evidence of learning or perceiving calling out as misbehavior rather than enthusiasm. After all, quietly completing a low-level worksheet could be dubbed "engaged" behavior, but because completing low-level worksheets involves no literate thinking, we would argue that having students quietly fill in blanks isn't evidence of effectiveness.

Exemplary teaching looks and sounds **different across different classrooms** and contexts.

Do evaluation tools spark meaningful conversations with teachers?

Focusing evaluation on underlying values instead of on individual indicators inspires evaluators to ask questions that support both teachers and students, such as, Are students in this classroom engaged? How do you know? If some are not, why not, and how can I help?

If teachers are unaware of what's happening in their classrooms and don't know how to reach more students, they need coaching and conversation. If they persistently fail to acquire this knowledge, a poor evaluation and employment decisions might follow. Whether students are sitting in a specific position or not (the indicator) is irrelevant if they're engaged. Likewise, as a participant at the 2012 annual meeting of the American Educational Research Association pointed out, whether a class will earn a teacher a high or low value-added score is irrelevant on a day-to-day basis. However, the questions that surface for administrators and teachers in the evaluation process—questions that engage professional judgments, inquiry, and reflection—could make all the difference in a student's education.

Of course, the immediate response is that these questions aren't "objectively" or "empirically" measureable. The statistical dazzle and media coverage (Gabriel, 2012; Gabriel & Lester, 2010) around value-added measurement have created an illusion of objectivity (Ewing, 2011) that has obscured the limitations of statistical methods of determining teacher effectiveness. There is a well-documented set of concerns about value-added measurement in terms of error rates, reliability, model differences, and even exclusionary practices.

For example, in their federally funded national study of exemplary 4th grade teachers, Allington and Johnston (2002) declined the use of value-added measurement as a tool to identify the most effective teachers in a district because the only students who "count" in some district value-added calculations are those without special education or English language learner labels who were present in class on a specific day in November.

The claim of objectivity by virtue of being "statistical" has secured solo billing for testing and value-added measurement as the "best we have" in an imperfect world. In fact, this logic is used to excuse the nation's headlong dash to incorporate value-added measurement into state and federal policies despite its flaws.

This is not just at the level of politics and policy. During a 2010 meeting of the teacher evaluation advisory committee in Tennessee that was open to the public, an administrator reported that she preferred a certain observation rubric because it "took her out of the equation" and she didn't need to make any "inferences." She couldn't say which of those behaviors actually had a genuine impact on student achievement, noting that that was something "we don't know yet." She was reserving

judgment, she said, until she got more information from the MET project.

In other words, the administrator was more comfortable counting indicators than she was discussing the values evident in the classroom, and she is waiting on the MET project to tell her what is "true and genuine" about the behaviors she observed.

Do evaluation tools promote valuable education experiences?

We reject the idea that the best we can hope for is a continuous search for objectivity in measures of student learning and teacher effectiveness. Instead, we argue that the best we should hope for is *authenticity* in the tasks we ask students to engage in and the assessments we use to understand their progress. The questions that deserve million dollar price tags should be those that we pose as educators every day: Are students experiencing the education we hope for them? How do we know? If some are not, how can we help?

If your school values creating a democratic citizenry, supporting children's socioemotional needs, or helping students read the world (Friere & Macedo, 1987) not just the word (or nonsense word, in the case of some current progress monitoring), then we may need another $45 million. Neither lists of indicators nor the so-called gold standard of value-added data measure those things. Moreover, a set of multiple measures designed to correlate with test scores doesn't keep such goals in sight.

Beyond Letters and Numbers

As Alfred Tatum (2007) wrote about increased evaluation of students, "Pressure to meet adequate yearly progress (AYP) has led to overlooking young people (OYP)" (p. 83). Let's not let teacher evaluation do the same. EL

Effectiveness has only recently been conflated with teacher effects on **student test scores.**

References

Allington, R., & Johnston, P. (2002). *Reading to learn: Lessons from exemplary 4th grade classrooms.* New York: Guilford Press.

Bond, G., & Dykstra, R. (1967). The cooperative research program in first-grade reading instruction. *Reading Research Quarterly, 2*, 5–142.

Coy, P. (2002, April 1). Economic trends: When hospitals get graded: There's a downside to rankings. *Business Week*, 6.

Darling-Hammond, L. (1990). Teacher evaluation in transition: Emerging roles and evolving methods. In J. Millman & L. Darling-Hammond (Eds.), *The new handbook of teacher evaluation: Assessing elementary and secondary school teachers* (pp. 17–34). Thousand Oaks, CA: Corwin.

Duke, N. K., & Pearson, P. D. (2002). Effective practices for developing reading comprehension. In A. E. Farstrup & S. J. Samuels (Eds.), *What research has to say about reading instruction* (3rd ed., pp. 205–242). Newark, DE: International Reading Association.

Ewing, D. (2011, April 5). Leading mathematician debunks "value-added" [blog post]. Retrieved from *The Answer Sheet* at *The Washington Post* at www.washingtonpost.com/blogs/answer-sheet/post/leading-mathematician-debunks-value-added/2011/05/08/AFb999UG_blog.html

Freire, P., & Macedo, D. (1987). *Literacy: Reading the word and the world.* South Hadley, MA: Bergin and Garvey.

Gabriel, R. (2012, April). *Constructions of value-added measurement and teacher effectiveness in the* Los Angeles Times*: A discourse analysis of the talk surrounding measures of teacher effectiveness.* Paper presented at the conference of the American Educational Research Association, Vancouver, BC, Canada.

Gabriel, R., & Lester, J. (2010, December 15). Public displays of teacher effectiveness. *Education Week.* Retrieved from www.edweek.org/ew/articles/2010/12/15/15gabriel.h30.html

Gates, B., & Gates, M. (2011). Grading the teachers: Schools have a lot to learn from business about how to improve performance, say Bill and Melinda Gates. *Wall Street Journal*. Retrieved from http://online.wsj.com/article/SB10001424052970204485304576641123767006518.html

Johnston, P. (2005). Literacy assessment and the future. *The Reading Teacher, 58*(7), 864–868.

Kane, T. J. (2012, March 28). Measuring effective teaching with a team of superheroes [blog post]. Retrieved from *Voices in Education* at www.hepg.org/blog/74

Knapp, M., et al. (1995). *Teaching for meaning in high-poverty classrooms.* New York: Teachers College Press.

Langer, J. (2002). *Beating the odds:Teaching middle and high school students to read and write well.* Albany, NY: National Research Center for English Learning and Achievement.

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

Pianta, R., La Paro, K., & Hamre, B. (2008). *Classroom assessment scoring system: Manual K–3.* Baltimore, MD: Brookes Publishing.

Pressley, M., Allington, R., Wharton-McDonald, R., Block, C. C., & Morrow, L. M. (2001). *Learning to read: Lessons from exemplary first grade classrooms.* New York: Guilford Press.

Pressley, M., Dolezal, S. E., Raphael, L. M., Mohan, L., Roehrig, A. D., & Bogner, K. (2003). *Motivating primary grade students.* New York: Guilford.

Stuhlman, M. W., & Pianta, R. C. (2009). Profiles of educational quality in first grade. *Elementary School Journal, 109*(4), 323–342.

Tatum, A. (2007). Building the textual lineages of African American male adolescents. In K. Beers, R. Probst, & L. Rief (Eds.), *Adolescent literacy: Turning promise into practice* (pp. 81–85). Portsmouth, NH: Heinemann.

Rachael Gabriel (rachael.gabriel@uconn.edu) is an assistant professor of literacy education at the Neag School of Education, University of Connecticut, Storrs. **Richard Allington** (rallingt@utk.edu) is professor of education at the University of Tennessee, Knoxville.