Teachers, Teaching,
and Reform; edited by Ralph P. Ferretti and James Hiebert
Format: US_6x9 (152.4 × 228.6 mm); Style: Supp; Font: Bembo;
Dir: {Desktop}New folder (2)/TTR/9781138735316_text.3d;
Created: 04/10/2017 @ 15:23:03

4

DOES VAM + MET = IMPROVED TEACHING?

James W. Stigler, James Hiebert, and Karen B. Givvin

UCLA; UNIVERSITY OF DELAWARE; UCLA

In 1993, in a small conference room at the National Center for Education Statistics in Washington, DC, plans were being made for the Third International Mathematics and Science Study (TIMSS). The design of the main data collection on student achievement had been finalized but, anticipating only moderate performance by its students, the United States was interested in additional studies that might reveal some of the reasons for cross-country differences in student performance. These studies would involve smaller subsets of countries. The studies discussed during the meeting in Washington, DC included longitudinal data collection, case studies of school and community environments, and a video study of mathematics teaching practices. Key to these studies would be the participation of Japan. At the time, Japanese students were among the highest achieving in the world (McKnight et al., 1987), and there was great interest in learning about factors that might account for their stellar performance.

Investigators eventually decided on a number of studies and pitched their plans to a Japanese representative from Japan's National Institute for Educational Policy Research, the unit responsible for conducting TIMSS. One after another the studies were presented, and each was rejected, politely, by the Japanese. By the time the video study was presented, there was little hope the Japanese would participate. If they did not, there was little chance the study would be conducted.

The TIMSS video study proposed to the Japanese was an ambitious undertaking in which national samples of eighth-grade teachers would be videorecorded teaching a single, randomly selected, mathematics lesson in their classrooms. The study would be done in three countries: Japan, Germany, and the United States. The videos would be coded by teams of coders from each country, and the codes turned into measures of instructional practices. Instruction would be compared by looking at the occurrence and frequency of different teaching features across the three countries.

Japan would be critical because many researchers thought that the interesting comparisons would come from comparing lower and higher achieving countries. Surprisingly, the Japanese representative almost immediately said yes, and the video study was launched (Stigler, Gonzales, Kawanaka, Knoll, & Serrano, 1999; Stigler & Hiebert, 1999).

Later, one of us had a chance to ask the Japanese representative why, when they had turned down almost all of the other studies, they decided to participate in the video study. The answer was telling: "We want to watch the videos," the Japanese representative explained. "We might get some good ideas for how to teach mathematics."

Behind the Japanese representative's comment lies a very different way of conceptualizing education research. The TIMSS video study was funded by the U.S. Department of Education's National Center for Education Statistics. The value this statistical agency saw in the work was in the measures that would be developed and the statistical comparisons of teaching practices across countries. But, the Japanese were not so interested in the statistical comparisons. In fact, during the project, they more than once wondered aloud why a national probability sample was required for a study like this. The U.S. funders, on the other hand, believed the national probability sample was critical to circumvent the variations created by local contexts and allow reliable comparisons among *average* teaching practices. Whereas the Japanese were interested in variation, the U.S. funders were interested in averaging away that variation. From our perspective, U.S. researchers in the National Center for Education Statistics never showed much interest in watching the videos. U.S. statisticians were more interested in the means and standard deviations – properly weighted, of course.

The story of the TIMSS video study highlights what we see as two very different models of how research on teaching impacts student learning. Trying hard not to resort to caricatures, we will explore these models. We will begin the chapter by outlining the more traditional approach on which many current policies for improving teaching in the United States are based, and then discuss the logic of this approach and the work for which it is well suited. We then lay out an alternative research approach, one that prompted the Japanese representative's response. In this approach, with its roots in quality improvement methodologies, a different theory of improvement is assumed. We then discuss two examples of this latter approach, one from Japan and another more recent one from the United States. At the end, we draw conclusions from these analyses and note the alignment of the alternative approach with the goal of improving teaching.

VAM + MET: One Research Approach for Improving Teaching

We start by unpacking part of the title we've given this chapter: VAM + MET. VAM refers to Value-Added Models – an increasingly popular way of assessing an individual teacher's effect on the learning of her/his students. The idea behind

VAM makes a lot of sense: if you want to measure the effectiveness of a teacher, find a way to measure what students learn during the year they spend with the teacher. Although it might sound simple, this is not easy to do. There are lots of complicating factors. For example, two teachers might use different assessments, or even teach different curricula. But, newer methodologies are giving us ways to account for these potential confounds to yield better and more comparable estimates of each teacher's value-added effect on her/his students' learning (Briggs, 2012; Hanushek & Rivkin, 2010; McCaffrey, Lockwood, Koretz, & Hamilton, 2003).

It turns out that, although they are far from perfect, value-added measures of teacher effectiveness do predict significant variance in student learning outcomes. In other words, it matters which teacher a student gets. However, what is not clear from the results is what it is *about* a particular teacher, or the way they teach, that causes students to learn to varying degrees.

The Bill & Melinda Gates Foundation has invested heavily in answering exactly this question through its support of the Measures of Effective Teaching project, or MET (Kane, McCaffrey, Miller, & Staiger., 2013; Kane & Staiger, 2010, 2012). MET is the largest study of teaching practices ever conducted. In the study, videos were collected in large numbers of classrooms. Researchers were invited to identify features of teaching that were making a difference in students' learning. To date, the study has yielded only small correlations between any of the observational measures and student value-added scores (Kane et al., 2013). Most of the variance in learning is still left unexplained. In fact, the MET project is not alone in this. Other researchers have used some of the same observational measures to investigate the relationship between teacher practice and VAMs, and have found similarly low correlations (Bell et al., 2012; Grossman, Loeb, Cohen, & Wyckoff, 2013).

Because most of the variance in students' learning is unexplained, we question whether VAM + MET can improve teaching, at least by itself. Does it make sense to recommend (or hold teachers accountable for implementing) a practice that accounts for, say, only 5 percent of the variance in student learning? Where should educators and researchers go from here? One option is to stay the course – to keep trying. Perhaps researchers just haven't succeeded *yet* in understanding how features of teaching impact student learning. Once researchers find the features of teaching that make a big difference in students' learning across classrooms (i.e., that account for a higher percentage of variance), this approach will work fine.

Another option, and the one we favor, is to acknowledge that VAM + MET cannot, by itself, improve teaching. The approach, by design, is not structured to *improve* teaching. But, evidence-based approaches do exist that are designed explicitly to improve processes and outcomes. Before exploring such an approach, we explain why we think VAM + MET is not the most promising approach if the goal is to improve teaching.

Although the MET project is certainly the largest attempt to crack the teaching-learning code, it is by no means the first. Nuthall (2005), for example,

recounted his own long and tortuous journey to find observational measures of teaching that were strongly correlated with learning, a journey that, for the most part, did not succeed. Low correlations, by themselves, are not necessarily a bad thing, or even unexpected, from a research perspective. After all, teaching is only one of a number of factors that might influence student learning, and any one aspect of teaching would be expected to have low predictive power with respect to end-of-year measures of student achievement. Even low correlations can be very useful when developing theories of how teaching affects learning – they provide clues that might warrant further investigation or revisions in researchers' theories.

But from a policy for improvement perspective, the low correlations represent a significant challenge. For one thing, variables identified using primarily correlational research methods might be misleading, representing an average across many occurrences, some positive, of course, but some negative as well. Even though a variable is positively correlated with student learning outcomes, asking a teacher to *do more* of the variable (whatever it is) might have no effect, or even a negative effect, on student learning. For example, consider something as simple as the pace of instruction. It might be the case that many teachers who move more rapidly through the curriculum have students who, on average, learn more by the end of the year (Leinhardt, 1986). A statistically significant correlation is found. But, what if students in these teachers' classes are well prepared to learn? Teachers with students who are less prepared could actually hurt their students' learning by moving too quickly.

Interpreting correlations between teaching and learning is difficult, in part, because classrooms are extremely complex places. Teaching is not just a collection of variables but a dynamic system in which teaching actions can have different effects depending on the context in which they are embedded (Gallimore, 1996; Nuthall, 2004; Stigler & Thompson, 2009). The reason many reported individual correlations are so low might be due to the systemic, contextual nature of teaching. A teaching action that works well in some situations might be the wrong thing to do in another situation. Figuring out what will work well, for which specific purpose, in which situation, will require much more detailed theories than currently are available. And without such theories, researchers are unlikely to identify features of teaching they can recommend for all teachers across all contexts.

Because teaching is a system, affected by the context in which it operates, an approach like VAM + MET, based on searching for statistically significant averages, and intentionally ignoring the sources that produce variations, seems to us an approach ill-suited to yield information useful for improvement. So, we now turn to a different approach, one we will call improvement science. Connecting back to our opening story, this research tradition, though developed initially in the west, found its most fertile ground for development in Japan after World War II.

Improvement Science: An Alternative Research Approach for Improving Teaching

Because it has rarely been applied to education and is thus likely to be less familiar to readers, it is important to note that improvement science has a long history in other fields (Gabor, 1990; Juran & DeFeo, 2010; Langley et al., 2009; Kenney, 2008). For example, it has been used for years in industry (Rother, 2009) and medicine (Gawande, 2007; Kenney, 2008) as a research-based approach for improving system processes and outcomes. Only recently have researchers begun exploring application of these principles to improving education (Bryk, Gomez, & Grunow, 2011; Bryk, Gomez, Grunow, & LeMahieu, 2015; Morris & Hiebert, 2011; Sparks, 2013).

Two Key Features of Improvement Science

The main features of improvement science were set forth by Edward Deming in the middle of the last century (Gabor, 1990). Deming was deeply involved in creating systemic, evidence-based methods to help organizations improve their processes and outcomes. A first feature of improvement science, promoted by Deming, was a shift in the goal of research, from the traditional goal of testing large-scale theories to the goal of improving complex systems. Basic research has been about building general theories, and testing hypotheses in order to revise these theories. Improvement research, on the other hand, is about *improving the performance of a system*. "Every system is perfectly designed to achieve the results it achieves," writes Paul Batalden, a pioneer in the application of improvement science to health care (Berwick, 1996, p. 619). For instance, if only 60% of community college students nationwide are capable of passing a placement test to gain entrance to a college-level mathematics course – even though the vast majority are high school graduates – then that result is exactly what the system is designed to achieve. If educators want a different outcome, Deming would say that they must study the system, understand how it works, and then design changes that will improve the outcome.

A second feature of improvement science is a focus on understanding and reducing variation in outcomes. In the education research tradition, showing that the average effect of a teaching intervention is unlikely due to chance variations in sampling alone ($p < .05$) is often the end of the investigation, even though most of the variance remains unexplained. In improvement research, the goal of improving systems requires reducing variation in the outcomes of the system. Rather than accepting variance as measurement error or an inevitable consequence of the complexity of education, improvement science tries to understand the reasons for the variance and reduce it through incremental changes to the system. Variability, observed Deming, is produced by the *system* and reducing variability requires changes to the system. In the case of teaching, the variation of most concern is the variation in student learning, both within and among classrooms. Educators share the goal of having all students learn to some satisfactory level in a repeatable way.

A Framework for Improvement

Deming and his followers proposed a framework for improving systems that consists roughly of the following five components (see Gabor, 1990; Juran & DeFeo, 2010; Langley et al., 2009; Rother, 2009):

- Clear, shared goals
- Sensitive measures to chart progress
- Deep understanding of the problems/barriers that impede success – "seeing the system" in the way it functions currently
- Sources of innovations, grounded in explicit theories of the problem
- Mechanism for comparing/researching innovations

The last component – the mechanism for testing and improving innovations – is most commonly implemented as the PDSA cycle: Plan, Do, Study, Act (Langley et al., 2009). This methodology is an iterative process in which improvements are developed, tried, studied, and then refined, over and over, until the average level of outcomes is improved and variability is reduced to acceptable limits. The first step is to *plan* the test. The plan involves creating a potential solution based on a clear hypothesis about what will lead to improved outcomes and/or reduced variation. The second step is to *do* or carry out the test – that is, to execute the plan and assess the results of it. Next comes *studying* the result. What was learned from the test? How do actual results compare to expected ones? Finally, based on the results of the analysis, a decision is made on how to *act*. What next step is warranted based on what was learned: Should the change be adopted, adapted, or abandoned?

It is important to note that building knowledge – in the form of useful, *local* theories – is a critical part of improvement science. In fact, Rother (2009) points out that the first step in identifying a change in the system to be tested is a deep understanding of the system as it currently works – what Rother (2009) calls the "current condition." Often, just understanding the system as it works today – creating a local theory of how the current system is working – is enough to solve the problem that prompted the inquiry.

Based on a thorough understanding of the current condition, one then defines a "target condition" (Rother, 2009), which amounts to detailed hypotheses about what the system will look like if the problem is solved. PDSA cycles are the method used to progress from the current condition to the target condition, testing hypotheses and building knowledge along the way. By building knowledge of the system, researchers not only are improving the outcomes that the system is designed to produce, but also are figuring out the *active ingredients* – those elements and processes that are essential to improve the system's performance. Knowing the active ingredients in a system allows researchers to adapt the system to better fit new contexts.

The Relationship Between People and the System in Which They Work

Applying improvement science to education requires thinking about the role of people – the practitioners who implement the process – in the improvement of the system. Deming carefully distinguished between the system and the workers, arguing that it was grossly unfair to punish and reward individuals when it was the system that was producing too much variation in outcomes (Gabor, 1990). All systems produce variation. But if one relies on individuals, working in isolation, to overcome the variation, one will gain, at best, only temporary and isolated improvements. Rother (2009) makes the same point when describing Toyota's improvement *kata*, a concept based on Deming's framework. Even though individual workers will, heroically, pull out the stops to solve a problem – staying late, working harder, and so on – Toyota frowns on this, calling it a "workaround." Workarounds may solve the problem in the short term, but often forestall efforts to make the lasting, long-term, and large-scale improvements that come from designing processes that are immune to the naturally varying actions of individuals. This requires designing systems that work well in the context – human and otherwise – in which they need to perform.

In education, a similar point – differences between people and the system that shapes their work – is made by distinguishing between teachers and teaching (Hiebert & Morris, 2012). In the education policy tradition, the strategy has been to hold individual teachers accountable for variations in student outcomes, rewarding those teachers who produce higher performing students, and penalizing teachers who do not (Gitomer, this volume). Because of the high stakes involved, educators need highly objective and reliable measures of teaching effectiveness that are perceived as fair for all teachers.

The need for reliable measures means that the work of producing these measures falls to a highly specialized group of researchers and psychometricians. Because a large investment is required to produce such measures, they must be designed to work across a large variety of grade levels and contexts. But, this need for general-purpose measures makes them, by definition, less relevant to the specific goals and local contexts within which teachers work. Teachers can perceive such measures as irrelevant to their work, imposed from the outside (Gitomer, this volume; Goe, Bell, & Little, 2008; Martinez, Taut, & Schaaf, 2016; Pianta, this volume).

In the improvement science tradition, practitioners play a central role in figuring out how to improve the system of processes that produce the current outcomes. Teachers (in the case of education) are the people with the most detailed information about how the system of *teaching* works, and, although they might not be the people best positioned to create innovations, they are best able to test and adapt innovations, translating them into daily lessons and implementing the lessons effectively. Because the success of an innovation is determined by implementation, teachers are at the heart of the system.

Implementation of teaching innovations is considered a major challenge for education researchers. Teachers might resist adopting the innovations and, if they adopt them, they might implement the innovations differently than intended by the researchers. In improvement science, implementation is seen as a core part of the research process. Implementation is not just the enactment of research innovations, but part of researching the innovations themselves. And, in education, teachers are the people with expertise in teaching, in implementation. Because teaching is what must improve, close involvement of teachers throughout the research process is a key principle in the improvement science strategy.

As alluded to earlier, the aim of improvement science is not teach*er* assessment. That doesn't mean, however, that teach*ing* is not assessed. Measurement plays a different role in the improvement science tradition than it does in the education research tradition, but it is no less critical. Langley et al. (2009) describe three types of measures that are required for improvement: measures of outcomes, to detect when progress is being made, both in improved average outcomes and in reduced variation in outcomes; measures of process, to validate the local theories and hypotheses proposed as the mechanisms for improvement; and balancing measures, to make sure that improvements on one outcome are not making some other valued outcome worse. In education, this means that students' learning is assessed to detect when progress in teaching is being made, teaching is assessed to detect whether the changes hypothesized to make a difference in students' learning actually occurred, and student outcomes along with teaching activities are monitored to check for unwanted side effects (e.g., teaching became overly punitive to motivate students to perform better). The important point here is that the data are used in a direct way, to guide the improvement process.

In the education research tradition, measures are used to validate innovations, which are then disseminated to teachers who are expected to implement them. Reliability of measures is critical because the single test of an innovation's effectiveness must be trustworthy. In the improvement science tradition, measures must only be *good enough* to provide feedback to the improvement team as they work through multiple PDSA cycles. Measures don't need to provide reliable scores for individual teachers. The role of measures in improvement science is to provide information about teaching that is directly actionable during that cycle. Measurements will be taken again, and again, as the cycles are repeated. Repeated small measurements replace a single large-scale measurement because repeated measurements are better suited to guide an iterative improvement process.

An Example from Japan of Studying Teaching within the Improvement Science Tradition

Deming worked out many of the details of improvement science in Japan, where he worked to rebuild Japan's post-war industrial complex. His ideas and his focus on continuous improvement found fertile ground in Japanese culture, and his influence

in Japan has been great (e.g., Rother, 2009). Although the historical connection is not clear, lesson study in Japan (Lewis, 2002; Stigler & Hiebert, 1999) can be seen as an early application of Deming's ideas to educational improvement. Clearly evident in lesson study are the PDSA cycles used in improvement science.

Lesson Study and Improving Teaching

Japanese teacher groups develop, implement, and test improvements in teaching methods. Some methods are more effective than others. But, because teaching is complex, the effectiveness of methods depends on particular learning goals and local conditions. Consequently, teachers must not only improve the methods, but also improve their ability to select the best methods for particular conditions. The long track record of success for lesson study is due to the fact that it helps teachers accomplish both goals: it facilitates incremental improvements in teaching methods and it allows teachers to grow their expertise in selecting and implementing the methods (Huang & Li, 2009; Lewis, Perry, & Hurd, 2004, 2009; Lewis & Tsuchida, 1998).

Lesson study begins with examining the curriculum and formulating goals. This is akin to what Rother (2009) described as the target condition. Teachers work with district specialists, and with national curriculum standards, to specify, precisely, the desired long-term learning goals for students. Then, in the Plan phase of the cycle, teachers study methods of teaching to generate hypotheses about how particular methods, using particular instructional activities, are likely to help students achieve these learning goals. Students' learning goals can be described at different levels of detail but the descriptions are most helpful if they include a level that is specific enough to guide the design and assessment of a daily lesson. A lesson is designed that aims to help increase the mean level with which students achieve the learning goals for that lesson *and* reduce the variation among students in their performance. Assessments of students' learning that are directly tied to specific lesson goals are created to reveal whether the predictions of effectiveness were correct.

The Do phase of the cycle involves implementing the lesson. Assessments of teaching are a critical aspect of this phase, but they are substantively different than the more formal, large-scale, highly reliable measurements sought by the education research tradition. As noted earlier, this is largely due to differences in goals. Improvement science (and lesson study) focuses on continuous cycles of data collection for improvement whereas basic education research often counts on one-time tests of hypotheses to revise theories. Because cycles are repeated, assessments of teaching in any single Do phase of the cycle require collecting just enough data to know how the lesson was implemented in that cycle. Repeated assessments of teaching will occur as lessons are changed and implemented again to continually test for improvements.

Assessments of teaching embedded in the Do phase are designed so teachers can apply them, either through peer observations or self-report, and use the

results to compare outcomes across classrooms. Imagine teachers are teaching toward the same learning goals. Then the results of these assessments of teaching allow teachers to move naturally to the Study phase, where they compare the student outcomes in different classrooms and use their assessments of teaching to formulate some hypotheses about which set of actions (teaching methods) produced the best outcomes. By repeatedly connecting observed teaching actions with student learning, teachers can begin to build knowledge about which combinations of teaching actions or methods work best for a specific set of learning goals. This is improvement science at work in education, conducted directly within the context that matters the most – ongoing classroom teaching.

During the Study phase of the cycle, unintended consequences are also taken into account as new hypotheses are tested. Did the changes in teaching have their desired effect? Was the mean level of students' performance improved? What might have produced the variance in student learning and how can it be reduced? It is the formation, and reformation, of hypotheses about cause-effect connections between teaching and learning, along with documenting unintended effects, that builds the knowledge needed for continuing, steady, improvement?

Finally, the Act phase requires adopting, adapting, or abandoning the teaching actions now thought to more (or less) effectively produce the desired student learning. This sets into the motion the next PDSA cycle.

Lesson Study and Assessing Teachers

Earlier, we argued that improvement science focuses its assessment on processes, not people – on teaching, not teachers. Although this is true, the way in which its key features can guide the assessment of teachers is instructive. We illustrate this aspect of improvement science because it stands in stark contrast to strong movements in the United States toward teacher evaluation and accountability.

At the heart of the difference between improvement science assessments and teacher evaluations is the nature of the data collected, and how these data are used. In lesson study, our example of improvement science at work in education, student learning is assessed to test specific predictions made in the lesson plan. Teaching is observed, likewise, to test current hypotheses about how students' thinking and learning is affected by the teaching specified in the lesson plans. Formal measures are not required in this situation for there is no need to use them beyond the work of the lesson study group. What matters is just that the student assessment measures and the teaching observation rubrics be good enough to guide the work of improvement.

This approach to data and measurement appears to carry over into the Japanese system for teacher accountability (Stigler, 2010). Unlike in the United States, Japanese teachers are not formally assessed, either for their instructional practices or for their students' learning. Instead, lesson study and related protocols are designed to feed data back to teachers that will be useful for their own improvement. Students in

K-8 Japanese schools are typically assessed by monthly exams. These exams are created by the teachers who teach each particular subject and grade level. The combination of a national curriculum, large classes, and random assignment of students to classes sets up an ideal context for identifying and improving teachers whose students are underperforming.

Because the teachers construct the exam questions, and because they share among themselves the instructional goals for that month, teachers have no basis to complain about the exam itself: the scores their students achieve are meaningful and informative relative to teachers' goals. If one teacher's students consistently underperform relative to the other classes, there are rich sources of information that can be used to track down the cause. Students' performance on the monthly exam gives far more information about what students do and don't understand than does the typical standardized test. And, the professional culture in which teachers' observations in each other's classrooms is supported means there are natural avenues for figuring out which specific teaching actions might account for specific deficiencies in students' average performance.

Discussions about why a particular teacher's class shows lower scores can be difficult, socially, especially if the same teacher's students consistently underperform. Efforts to understand the reasons for the teacher's ineffectiveness can become the subject of a lesson-study-type investigation. Observations of teaching play a central role in comparing teaching actions that might be making the difference in students' learning. The lower-performing teacher is helped to see what could be changed to improve students' performance. What if the teacher chooses not to use the feedback to improve? Often, such teachers leave the profession. Imagine how it must feel to consistently produce lower levels of students' learning and have these results examined by one's colleagues.

This example from Japan shows a simple accountability system in which the first-hand observation of teaching plays a critical role. Observational *measures*, on the other hand, are not needed. Because the observations are intended only to provide information for improvement, and because they are accessible only among a small group of colleagues, there is no need for the kinds of reliable quantitative observation measures envisioned under the VAM + MET model. Both for student measures (the monthly exams) and for teacher measures (observations), psychometric robustness is traded for the timely and useful feedback that can be provided directly to those in the best position to act on it – the teachers themselves. The intent is not to establish statistically significant relationships but to improve classroom teaching. And teachers report finding it useful and effective (Lewis, 2002).

A Large-Scale U.S. Example of Teaching Assessment within the Improvement Science Tradition

To further elaborate the forms that assessments of teaching can take if they are conducted within the improvement science tradition, we could describe several projects

currently underway in the United States that explicitly aim to improve teaching by building in assessments of teaching for purposes of improvement rather than evaluation. One such project is the improvement of mathematics teacher preparation for K-6 prospective teachers at the University of Delaware (Hiebert, Wieman, & Berk, this volume). Another project, Advancing Quality Teaching, is part of a larger effort by the Carnegie Foundation for the Improvement of Teaching to improve the mathematics learning opportunities for community college students (www.carnegiefoundation.org/developmental-math). This is the project we describe here.

Begun in 2009 and 2010, respectively, the Statway™ and Quantway™ Math Pathways programs were designed, with support from the Carnegie Foundation, to help developmental education students succeed in a college credit-bearing mathematics course (Clyburn, 2013). Work began by designing new curricula, one focused on statistics, and the other on mathematical literacy. Then came a roll-out of the curriculum materials in 30 colleges across 13 states. In 2012, after two revisions to the materials, work expanded to include improving implementation of the curricula.

The problem faced by the Advancing Quality Teaching (AQT) team, of which we were members, was to help instructors improve the nature of teaching being used to implement the new curricula. A solution to this problem, common within the education research tradition, and one our AQT team was tempted to take, was to develop a rubric to measure the quality of teaching for every lesson in the curriculum. The rubric would specify the teaching actions our team believed were warranted by previous research to facilitate students' achievement of the learning goals for the lesson (generally, understanding concepts rather than just executing procedures). Feedback to the project directors, and the instructors, would be scores on the rubric with indicators showing which teaching actions were and were not observed. However, because our AQT team wanted to improve teaching, rather than evaluate it, it looked instead to the improvement science tradition to guide its work. This was possible because the learning goals for students are specified in the curricula and shared by all instructors teaching the courses.

One tool our AQT team borrowed from improvement science, not described previously, was a driver diagram (Langley et al., 2009) (see Figure 4.1). It serves as a way of recording backward engineering (Wiggins & McTighe, 2011). Users begin by identifying a desired outcome. In the case of the Statway™ and Quantway™ courses, that outcome, or primary learning goal, is to help students understand mathematics deeply and flexibly. A review of literature led the team to hypothesize that the best way to reach that outcome was to create particular learning opportunities for students. Focusing on learning opportunities as the way in which teaching affects learning served two purposes. First, because teaching is a cultural activity (Gallimore, 1996; Stigler & Hiebert, 1999), the same teaching actions can produce different learning outcomes in one context to another, and different teaching actions can produce the same opportunities in different

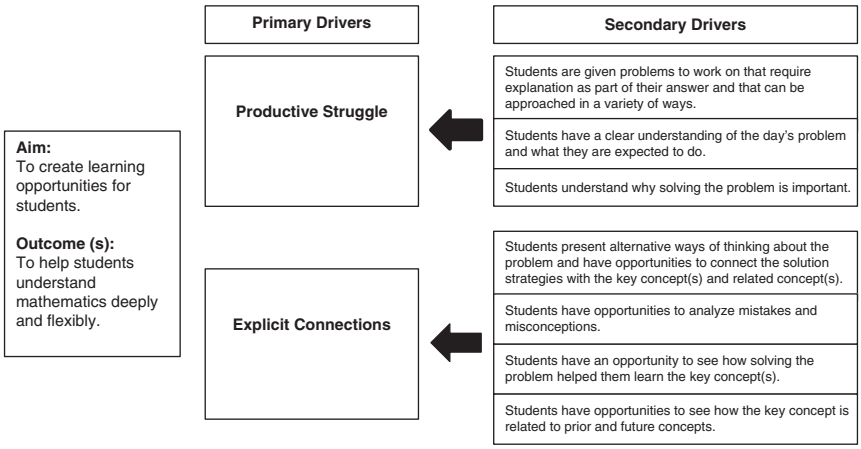

FIGURE 4.1 Cause-effect connections between teaching and learning using a backward mapping form

contexts. Consequently, our AQT team needed a construct that more stably links with learning than individual teaching actions. Second, learning opportunities offer a construct about which testable hypotheses could be developed. The hypotheses can be specified at a grain size that crosses individual instructional activities and even individual lessons, but they are specific enough that one can observe their instantiation.

Based on previous research, our AQT team identified two kinds of learning opportunities that seem to facilitate the goal of conceptual understanding in mathematics: (1) opportunities for students to struggle productively with the main mathematics concepts in a lesson; and (2) opportunities to make connections between mathematics facts, procedures, and/or concepts (Hiebert & Grouws, 2007). These opportunities, in the language of improvement science, are primary drivers (see Figure 4.1). Through additional reviews of the literature and discussions with course instructors, the team identified a series of secondary drivers – events that instructors can make happen in a mathematics class and which we hypothesize create the desired learning opportunities.

Following improvement science PDSA cycles (plan-do-study-act) described earlier, teaching assessments were used directly to improve teaching. Given the hypotheses about learning opportunities, this means refining teaching to create the appropriate learning opportunities at the right time in the lessons. In the case of our AQT team, the cycles played out in the following way.

The first step was to *plan* the test or experiment. The plan involved creating a potential solution based on a clear hypothesis about what will lead to improved outcomes and/or reduced variation. This meant identifying key teaching actions that would lead to the secondary drivers (see Figure 4.1). For instance, if our team

wanted to see students present alternative ways of thinking about the problem so they have opportunities to connect solution strategies with key concepts, an instructor might observe and study students as they work, so that she or he can call on students to present their solution strategies in an order that shows a building toward a main idea.

The second step in the PDSA cycle is to *do* the experiment – that is, to execute the plan and collect the results. In this case, our team asked instructors to observe students' work and call on students in a particular order to present strategies to the class (Smith & Stein, 2011). We asked instructors to document the process, paying particular attention to what facilitated and hindered their efforts and what they witnessed with respect to students' understanding of key concepts. Observers might also watch the lesson, live or on video, and assess whether and how these teaching activities occurred. Both self-report and observers' judgments contribute to teaching assessment and provide the data for the next step.

The third step is *studying* the result. What was learned from the experiment? How do actual results compare to expected ones? Two members of our team held calls every two to three weeks with a small group of instructors to review the results of their efforts. It's worth noting that the study phases, at least in one case, helped establish an additional line of work – one that involved a series of PDSA cycles of its own. As our team focused on helping students present alternative ways of thinking about a mathematics problem, we noticed that the lesson materials from which instructors were working often lacked a problem that was open-ended enough to allow for the application of alternative ways of thinking. The materials served as an obstacle to enacting the key teaching actions we were attempting to encourage. Those actions are, of course, just one part of a system – a system that also includes lesson plans. Thus, our team's PDSA cycles on teaching actions led to a set of PDSA cycles on materials improvement.

Finally, based on the results of the reviews, a decision is made on how to *act*. What next action is warranted based on what was learned? Should the team adopt, adapt, or abandon? Often, instructors offered suggested adaptations to the teaching actions or ideas about how to implement them more effectively. If adaptations were suggested, our team created hypotheses to guide the *plan* for another PDSA cycle. We continue until participating instructors are comfortable with the new routine in their classrooms and are ready to adopt a new action.

All the while, our team was guided by three questions that, together with the PDSA cycles, round out the improvement science model (Langley et al., 2009). First, what are educators trying to accomplish? In our case, the immediate goal was the secondary drivers, which in turn created the learning opportunities, which in turn were designed to produce the desired learning outcomes. Assessments of teaching play a key role at this point because the critical data are whether the teaching actions associated with the secondary drivers actually occurred. Second, how will educators know that a change is an improvement? At this point, instructors (and/or observers) need to collect (sometimes informal)

evidence about what actually happened in their interactions with students and how students responded. Did the teaching actions appear to create the kind of learning opportunities hypothesized and did these lead toward the desired learning outcomes? Third, what change can educators make that will result in an improvement? The initial change ideas come when creating the driver diagram, but these changes are adapted based on data collected during the PDSA cycle.

Conclusions

In this chapter, we compared two different approaches to improving teaching. The first approach, which is most common among U.S. education researchers, relies on general theories of teaching, measures of students' learning valid across multiple curricula, rubrics to observe teaching that can be used for different subjects and grade levels, and sophisticated statistical techniques that look for significant relationships between teaching and learning. All of this is captured by the simple equation: VAM + MET = Improved Teaching. It would be wonderful if this approach worked. It would make researchers' efforts directly applicable to improving teaching. It also fits well with U.S. educators' preference for "loose coupling" in the management of education systems (Elmore, 2000). The direct and relatively simple relationships on which this approach rests might help explain why it has persisted for so long as the dominant approach. If only the science would support it, researchers and policy makers who work in this tradition could continue their current activities with the confidence that classroom teachers will learn from their findings and implement their recommendations to yield more effective teaching and improved student learning.

Unfortunately, despite a century of work, the traditional approach has not succeeded in producing lasting, research-based improvements in teaching (Cuban, 1993; Hoetker & Ahlbrand, 1969; Nuthall, 2004, 2005). As it turns out, John Dewey (1929) warned the research community of this outcome nearly 90 years ago. Due to the complexity of teaching, Dewey wrote, it will never be possible to translate traditional research results directly into rules to guide practice: "No conclusion of scientific research can be converted into an immediate rule of educational art" (Dewey, 1929, p. 19). The reason is that scientific findings, as commonly conceived, could never address the multiple conditions and other factors that impact teaching and learning in schools.

On the other hand, Dewey (1929) did see a clear and valuable role for scientific research:

> The value of the science, the history and philosophy of education acquired in the training school, resides in the enlightenment and guidance it supplies to observation and judgment of actual situations as they arise…. [Scientific results] direct his attention, in both observation and reflection, to conditions and relationships which would otherwise escape him.
>
> *(Dewey, 1929, pp. 30–31)*

So, traditional scientific findings are useful, but only as they sharpen the eyes of the practitioners. It is the practitioners who must figure out when and how such findings might be useful for achieving educational goals.

Moving research and innovation into the classroom to improve teaching calls for something different, an alternative approach. A promising candidate, in our view, comes from the considerable literature on quality improvement, sometimes called improvement science. We are not the only educators, or even the first, to recognize the need for such an approach (e.g., Bryk et al., 2015; Cicarella, 2014; Darling-Hammond, 2014; Minnici, 2014).

Although improvement methodologies are not easy to implement, they do, as we have described, have some significant advantages. First, they shift researchers' focus from the teacher to teaching – from the people to the system that produces the outcomes. This is important, because long-term improvements will only come about if we can change the fundamental pedagogical routines that govern teaching (Stigler & Hiebert, 2009).

Second, these improvement methodologies push researchers to go beyond the average effects of teaching and focus on the huge variation that currently defines the output of the U.S. educational system. Just raising the mean performance of students will not suffice if the goal is to develop an educated citizenry and workforce. Educators must find ways of educating all students. The improvement system we have described seems well-suited to both raising the mean level of student learning *and* reducing the variation in learning among students.

Scientific inquiry will play a critical role in improving teaching. But VAM + MET is not the only means of applying science to this task. Quality improvement approaches are equally scientific, driven by data, but from a different tradition. Researchers must measure student outcomes, and develop ways of understanding the links between instruction and learning. But researchers must do these things in a way that will lead to sustainable improvements in practice.

We understand why the Japanese TIMSS representative was more interested in watching videos of teaching from other countries than collecting probability samples of lessons, coding features of teaching, and statistically comparing occurrences and frequencies. Watching teaching, with an eye toward understanding how the system works, how the interactions within the classroom create key learning opportunities for students, generates hypotheses for improving elements in the system. These hypotheses contain exactly the kinds of ideas that can be tested and improved through multiple cycles of planning, doing, studying, and acting. Watching videos of teaching was the lever our Japanese friend realized could set this improvement process into motion.

References

Bell, C. A., Gitomer, D. H., McCaffrey, D. F., Hamre, B. K., Pianta, R. C., & Qi, Y. (2012). An argument approach to observation protocol validity. *Educational Assessment*, *17*(2/3), 62–87.

Berwick, D. M. (1996). A primer on leading the improvement of systems. *British Medical Journal*, *312*, 619–622.

Briggs, D. C. (2012). Making value-added inferences from large-scale assessments. In M. Simon, K. Ercikan, & M. Rousseau (Eds.), *Improving large-scale assessment in education: Theory, issues, and practice* (pp. 186–201). London, England: Routledge.

Bryk, A. S., Gomez, L. M., & Grunow, A. (2011). Getting ideas into action: Building networked improvement communities in education. In M. T. Hallinan (Ed.), *Frontiers in sociology of education* (pp. 127–162). Dordrecht, the Netherlands: Springer.

Bryk, A. S., Gomez, L. M., Grunow, A., & LeMahieu, P. G. (2015). *Learning to improve: How America's schools can get better at getting better*. Cambridge, MA: Harvard Education Press.

Cicarella, D. (2014). The professional educator: A fine balance. *American Educator*, *38*(1), 18–21.

Clyburn, G. M. (2013). Improving on the American dream: Mathematics pathways to student success. *Change: The Magazine of Higher Learning*, *45*(5), 15–23.

Cuban, L. (1993). *How teachers taught: Constancy and change in American classrooms, 1890–1990* (2nd ed.). New York: Teachers College Press.

Dewey, J. (1929). *The sources of a science of education*. New York, NY: Liveright.

Darling-Hammond, L. (2014). One piece of the whole: Teacher evaluation as part of a comprehensive system for teaching and learning. *American Educator*, *38*(1), 4–13, 44.

Elmore, R. (2000). *Building a new structure for school leadership*. Washington, DC: The Albert Shanker Institute.

Gabor, A. (1990). *The man who discovered quality: How W. Edwards Deming brought the quality revolution to America*. New York, NY: Times Books.

Gallimore, R. G. (1996). Classrooms are just another cultural activity. In D. L. Speece & B. K. Keough (Eds.), *Research on classroom ecologies: Implications for inclusion of children with learning disabilities* (pp. 229–250). Mahwah, NJ: Erlbaum.

Gawande, A. (2007). *Better: A surgeon's notes on performance*. New York, NY: Picador.

Gitomer, D. H. Promises and pitfalls for teacher evaluation. This volume.

Goe, L., Bell, C., & Little, O. (2008). *Approaches to evaluating teacher effectiveness: A Research Synthesis*. Washington, DC: National Comprehensive Center for Teacher Quality.

Grossman, P., Loeb, S., Cohen, J., & Wyckoff, J. (2013). Measure for measure: The relationship between measures of instructional practice in middle school English language arts and teachers' value added scores. *American Journal of Education*, *119*(3), 445–470.

Hanushek, E. A., & Rivkin, S. G. (2010). Generalizations about using value-added measures of teacher quality. *American Economic Review*, *100*, 267–271.

Hiebert, J., & Grouws, D. A. (2007). The effects of classroom mathematics teaching on students' learning. In F. K. Lester, Jr. (Ed.), *Second handbook of research on mathematics teaching and learning* (pp. 371–404). New York, NY: Information Age.

Hiebert, J., & Morris, A. K. (2012). Teaching, rather than teachers, as a path toward improving classroom instruction. *Journal of Teacher Education*, *63*, 92–102.

Hoetker, J., & Ahlbrand, W. P., Jr. (1969). The persistence of the recitation. *American Educational Research Journal*, *6*, 145–167.

Huang, R., & Li, Y. (2009). Pursuing excellence in mathematics classroom instruction through exemplary lesson development in China: A case study. *ZDM Mathematics Education*, *41*(3), 297–309.

Juran, J.M., & DeFeo, J.A. (2010). *Juran's quality handbook: The complete guide to performance excellence* (6th ed.). New York, NY: McGraw-Hill.

Kane, T. J., McCaffrey, D. F., Miller, T., & Staiger, D. O. (2013). *Have we identified effective teachers? Validating measures of effective teaching using random assignment*. Seattle, WA: Bill & Melinda Gates Foundation.

Kane, T. J., & Staiger, D. O. (2010). *Learning about teaching: Initial findings from the measures of effective teaching project*. Seattle, WA: Bill & Melinda Gates Foundation.

Kane, T. J., & Staiger, D. O. (2012). *Gathering feedback for teaching: Combining high-quality observations with student surveys and achievement gains*. Seattle, WA: Bill & Melinda Gates Foundation.

Kenney, C. (2008). *Best practice: How the new quality movement is transforming medicine*. New York, NY: Public Affairs.

Langley, G. J., Moen, R., Nolan, K. M., Nolan, T. W., Norman, C. L., & Provost, L. P. (2009). *The improvement guide: A practical approach to enhancing organizational performance* (2nd ed.). San Francisco, CA: Wiley.

Leinhardt, G. (1986). Expertise in math teaching. *Educational Leadership*, *43*(7), 28–33.

Lewis, C. C. (2002). *Lesson study: A handbook of teacher-led instructional change*. Philadelphia: Research for Better Schools, Inc.

Lewis, C., Perry, R., & Hurd, J. (2004). A deeper look at lesson study. *Educational Leadership*, *61*(5), 18–23.

Lewis, C. C., Perry, R. R., & Hurd, J. (2009). Improving mathematics instruction through lesson study: A theoretical model and North American case. *Journal of Mathematics Teacher Education*, *12*(4), 285–304.

Lewis, C., & Tsuchida, I. (1998). A lesson is like a swiftly flowing river. *American Educator*, *22*(4), 12–17.

Martinez, F., Taut, S., & Schaaf, K. (2016). Classroom observation for evaluating and improving teaching: An international perspective. *Studies in Educational Evaluation*, *49*, 15–29.

McCaffrey, D. F., Lockwood, J. R., Koretz, D. M., & Hamilton, L. S. (2003). *Evaluating value-added models for teacher accountability*. Santa Monica, CA: RAND Corporation. Retrieved from www.rand.org/pubs/monographs/2004/RAND_MG158.pdf

McKnight, C. C., Crosswhite, F. J., Dossey, J. A., Kifer, E., Swafford, J. O., Travers, K. J., & Cooney, T. J. (1987). *The underachieving curriculum: Assessing U.S. school mathematics from an international perspective*. A national report of the second international study. Washington, DC: U.S. Department of Education, Educational Resources Information Center (ERIC).

Minnici, A. (2014). The mind shift in teacher evaluation: Where we stand – where we need to go. *American Educator*, *38*(1), 22–26.

Morris, A. K., & Hiebert, J. (2011). Creating shared instructional products: An alternative approach to improving teaching. *Educational Researcher*, *40*(1), 5–14.

Nuthall, G. (2004). Relating classroom teaching to student learning: A critical analysis of why research has failed to bridge the theory-practice gap. *Harvard Educational Review*, *74*(3), 273–306.

Nuthall, G. (2005). The cultural myths and realities of classroom teaching and learning: A personal journey. *Teachers College Record*, *107*, 895–934.

Rother, M. (2009). *Toyota Kata: Managing people for improvement, adaptiveness, and superior results*. New York, NY: McGraw-Hill.

Smith, M. S., & Stein, M. K. (2011). *5 practices for orchestrating productive mathematics discussions*. Reston, VA: The National Council of Teachers of Mathematics.

Sparks, S. D. (2013). 'Improvement science' seen as an emerging tool in K-12 sphere. *Education Week*, *33*(6), 4–6.

Stigler, J. W. (2010). Needed: Fresh thinking on teacher accountability. *Education Week*, *29*(33), 30.

Stigler, J. W., Gonzales, P., Kawanaka, T., Knoll, S., & Serrano, A. (1999). *The TIMSS videotape classroom study: Methods and findings from an exploratory research project on eighth-grade mathematics instruction in Germany, Japan, and the United States*. NCES 1999–1074. Washington, DC: U.S. Department of Education, National Center for Education Statistics.

Stigler, J. W., & Hiebert, J. (1999). *The teaching gap: Best ideas from the world's teachers for improving education in the classroom*. New York, NY: Free Press.

Stigler, J. W., & Thompson, B. J. (2009). Thoughts on creating, accumulating, and utilizing shareable knowledge to improve teaching. *The Elementary School Journal*, *109*(5), 442–457.

Wiggins, G., & McTighe, J. (2011). *The Understanding by Design guide to creating high-quality units*. Alexandria, VA: ASCD.