Stigler, J. W., & Givvin, K. B. (2017). Online learning as a wind tunnel for improving teaching. In C. A. Christie, M. Inkelas & S. Lemire (Eds.), *Improvement Science in Evaluation: Methods and Uses. New Directions for Evaluation*, *153*, 79–91.

6

Online Learning as a Wind Tunnel for Improving Teaching

James W. Stigler, Karen B. Givvin

Abstract

Attempts to improve teaching through research have met with limited success. This is, in part, due to the fact that teaching is a complex cultural system that has evolved over long periods of time—multiply determined and inherently resistant to change. But it is also true that research on teaching is difficult to carry out. Using traditional educational research methodologies, testing new methods of teaching requires, first, that teachers be able to implement the method at a scale sufficient for study, that random assignment of teachers to conditions can be feasibly carried out, and that ecological validity of the treatment can be preserved. In this chapter, we propose an alternative approach that combines the affordances of online learning with the methodologies of systems improvement. Using an analogy from the development of the airplane, we discuss how online learning might be a wind tunnel for the study and improvement of teaching. © 2017 Wiley Periodicals, Inc., and the American Evaluation Association.

Processes of teaching and learning lie at the very core of education. Yet, improving teaching has proven to be one of the most difficult challenges facing education researchers and reformers. In this chapter, we reflect on why something so pervasive and seemingly straightforward as teaching has been so resistant to change, and even to research. We then discuss a new approach—grounded in improvement science and supported

New Directions for Evaluation, no. 153, Spring 2017 © 2017 Wiley Periodicals, Inc., and the American Evaluation Association. Published online in Wiley Online Library (wileyonlinelibrary.com) • DOI: 10.1002/ev.20232

by education technology—that we believe has great potential for improving teaching.

Improving Teaching: Why It's Hard

There is a long tradition of research on teaching. Each new generation, it seems, seeks to reinvent or reinvigorate what appears to be a straightforward approach: measure various aspects of teaching, identify those that are associated with desired student outcomes, then focus improvement efforts on those critical variables. Why is this so hard to do?

More Than the Sum of the Variables: Teaching Is System

One reason it's hard is that teaching is more than an assemblage of variables. It is a complex system in which the impact of one variable may depend on the others in complicated ways and the causal impact on learning is neither simple nor straightforward. We are led to this conclusion mainly for one reason: attempts to identify the critical variables that define teaching quality have been largely unsuccessful, as have reform efforts based on such variables.

These attempts have emerged from two different research traditions. One of these is classroom research. In this tradition, researchers have sought to describe classroom teaching by measuring variables hypothesized to affect student learning outcomes. Sometimes the variables are rooted in theories of learning, for example, behaviorism (in the 1960s and 70s) or cognitive psychology (in the 90s and today), whereas at other times, they emerge from detailed qualitative observations of classrooms. In both cases, the results have been disappointing. Nuthall (2005), for example, describes his own personal journey to crack the code of teaching and learning, a journey largely marked by disappointments, but an interesting read in any case. Through a series of studies employing a range of methodologies over a 40-year period, Nuthall failed to find anything he could measure about teaching that was significantly correlated with student learning outcomes. A more recent attempt is the large Gates Foundation-funded project, Measures of Effective Teaching. In this study, which is perhaps the largest and most highly funded study ever conducted, only a few small correlations were found (Kane & Staiger, 2012).

Another research tradition starts not with observations in classrooms but with theories developed in the laboratory. These researchers tend to come from the cognitive and learning sciences, and they have produced some fascinating results—in the lab. For example, Bjork and colleagues have shown that spacing and interleaving of items to be studied produces greater learning than does blocking of items (i.e., grouping items of a type together), even though learners themselves do not perceive this to be the case (Bjork & Bjork, 2011; Kornell & Bjork, 2008). As robust as these

effects are in laboratory studies, they have rarely been effectively transferred into the classroom, mainly because there is a lot more going on in a classroom than can be captured by a single variable.

Regardless of where the variables come from—whether from learning theory or empirical observation—it is not easy to create an effective lesson out of a list of variables, even if those variables have been shown to be effective "all other things being equal." Teaching and learning make up a complex system of interacting parts, and it is very hard to change one part without affecting the others. At a macro level of analysis, the system of classroom teaching and learning includes the teacher, the students, curriculum content, teaching routines, materials, district and state policies, assessments, physical layout of the classroom, parents, homework, and so on. Even in the best-case scenario, no single variable is likely to have a large effect on student outcome.

Teaching Is a Cultural Activity

So yes, teaching is a system, and a complex system. But that's not all. It is a cultural system—a set of routines, supported by widely held beliefs and values, that have evolved over long periods of time and that represent a hard-fought compromise between the desired and the possible. Why do we think that teaching is a cultural activity? We aren't the first to make this assertion (e.g., Gallimore, 1996). But we found the idea a compelling one as we worked through our analyses of data from the Trends in International Mathematics and Science Study (TIMSS) video studies (Stigler & Hiebert, 1998, 1999/2009).

In these studies, national probability samples of videos of classroom instruction—eighth-grade mathematics and science lessons, to be specific—were compared across eight different countries, some high performing and others (such as the United States) not. Like other classroom studies, we failed to find clear observational correlates of cross-national differences in mathematics and science achievement. But more important for our purposes here, we found large discrepancies in teaching routines across, but not within, countries, even among the high-achieving countries (Givvin, Hiebert, Jacobs, Hollingsworth, & Gallimore, 2005). Thus, to a greater extent than we would have predicted, teaching routines within a country—even one as diverse as the United States—appear to vary little when viewed from a cross-national perspective.

Cultural activities are learned implicitly, through participation from an early age. Even though we might wish that teachers would learn how to teach from teacher education programs, the evidence suggests that teachers largely just teach the way they, themselves, were taught. Cultural activities are hard to see. Because the routines are widely shared within a culture, we tend not to notice aspects, even those that may prove critical for student learning. And cultural activities are hard to change. They are hard to

change, first, because they tend to lie outside our awareness. Cultural activities are also hard to change because they are multiply determined. We may try to change some aspect of our teaching. But when we do, we almost certainly will get pushback from the rest of the system: students will complain, parents will question the change, textbooks become difficult to use because they are not aligned with the change, and so on. The difficulty of putting the Common Core standards into place in classrooms and the often vitriolic opposition to them is just one example of pushback from the larger system.

The cultural nature of teaching presents a methodological challenge to researchers seeking to understand the relationship of teaching to learning. In order to conduct an experimental test of a new teaching method, we first must get a sufficient number of teachers to adopt the change and be able to implement it faithfully in their classrooms. Education is a human-made institution, which means we are free to innovate and create something fundamentally different from what existed before—*in theory*. On the other hand, we don't see a lot of natural variation in teaching within a culture, and you cannot study what you cannot implement on a large enough scale (Gallimore & Santagata, 2006). Changing teaching is notoriously hard, even for research purposes.

Why Labs Settings and Randomized Controlled Trials Aren't the Answer

As articulated in this chapter, laboratory models have their limitations. Even though they may help us to avoid the tricky challenges inherent in changing teaching, they suffer from a lack of ecological validity, which is heightened by the fact that teaching is a complex system. Laboratory models, typically focused on one or a small number of variables, yield interesting theoretical results but are unlikely to transfer easily to the complex system of teaching in schools.

But these are not the only challenges to our traditional research methodologies. The "gold standard" of education research—the randomized controlled trial (RCT)—has some serious limitations, even when the challenges of implementation have been met. The studies are expensive, largely because of the challenges already outlined. Furthermore, even though reaching a statistical criterion of $p < .05$ may qualify a study as publishable, it is still true that it is a measure of average effects. Much of the variance is left unexplained and most researchers who conduct RCTs do little to learn from the variability within conditions, even though the intervention being studied may be helpful for some students and harmful to others. And, the interventions studied tend not to be interpretable in the context of theory—what Lipsey (1993) refers to as "small theories"—which makes them very difficult to adapt to new students and contexts. There must be a better way.

Improving Systems

We have argued thus far that teaching is a complex cultural system and also that the nature of teaching presents methodological challenges to those wishing to study and improve it. But there is another research tradition—one we will call improvement science—that has developed explicitly for the purpose of improving complex systems. In this section, we discuss this tradition and assess its applicability to the problem of improving teaching.

Roots of Improvement Science

The pioneers of improvement science—Deming, Shewhart, Juran, and others—developed their methodologies largely in industrial and manufacturing contexts. Recently, however, great strides have been made in applying the principles of improvement science to health care and some impressive results have been achieved (Gawande, 2007, 2010; Kenney, 2008). Educators often resent the analogy of education and manufacturing, but then, healthcare professionals have voiced similar objections. Although we don't want to gloss over the differences, we do want to explore the methodologies, especially because they have led to some impressive accomplishments in the medical world.

Deming posited four pillars of improvement science: appreciation of a system, understanding variation, human psychology, and the theory of knowledge development. Systems thinking is perhaps the most important. Deming observed that we often fail to see the system that produces the outcomes we are interested in; instead, we tend to zero in on a single variable. In manufacturing, for example, variations in quality result from a number of factors, including random ones. Yet, we mistakenly (in Deming's view) blame the worker for low-quality products, failing to see the system that led to the result. Paul Batalden, one of the pioneers of improvement science in health care, reportedly said: "Every system is perfectly designed to achieve the results that it gets." The first step in improving a system is to see and understand the system the way it works now.

Whereas traditional education researchers are typically satisfied when the variance between an intervention and a control group is greater than that within, the improvement scientist seeks to understand and reduce variation to within acceptable limits. If a system produces great variation—as is true of educational outcomes in general—it is not enough just to know that the average of one group is greater than another. It is important to understand the root of the variation and then make improvements in the process to both reduce variation (i.e., by bringing up the low achievers to acceptable levels) and improve the level of outcomes overall. Improvement scientists have developed statistical techniques for analyzing variability that are specifically designed to help understand and improve the outcomes of complex systems. As discussed by Berwick (2015) in a recent talk, R. A. Fisher developed statistics that target improvement of simple systems;

Walter Shewhart developed the statistical techniques that are the foundation for modern improvement science.

Psychology is important, of course, because human actors are part of many of the complex systems we care about most and that are most difficult to improve. If a better method of teaching is discovered that does not mean, it will be adopted in schools. Teachers would need to believe it is better for their students and that it is feasible to implement the method in the contexts in which they work. Psychology is also important for reasons that go beyond the role envisioned by Deming. Theories of psychology are a primary source of hypotheses to guide the development of new ideas for teaching and learning.

Finally, the tradition of improvement science specifies a disciplined methodology for iterative improvement and knowledge development, a methodology that includes the idea of Plan–Do–Study–Act (PDSA) cycles. This theory of knowledge development has been described in a variety of ways, but all can be encompassed in a common framework called the Model for Improvement (Langley, 2009).[1] It includes two components, the first of which is a series of three questions that guide the work:

1. What specifically are we trying to accomplish?
2. What change might we introduce, and why?
3. How will we know that a change is actually an improvement?

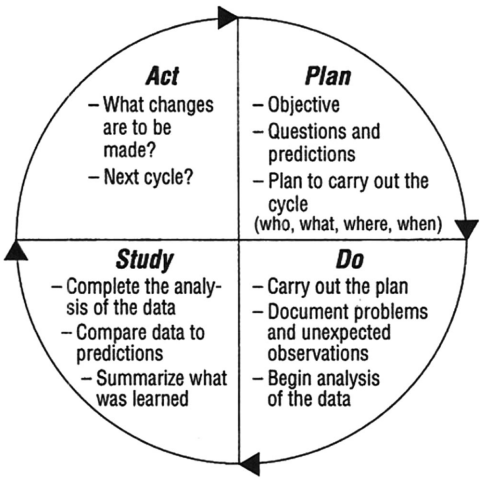
Answering these questions involves some important and often difficult pieces of work. Questions one and three go together, three being the question of measurement. As Bryk, Gomez, Grunow, and LeMahieu (2015) point out, "We cannot improve at scale what we cannot measure" (p.111). Langley et al. (2009) propose that we need at least three kinds of measures to do the work of improvement: measures of outcomes, measures of process, and balancing measures (to make sure that a change to improve one outcome does not accidentally make some other valued outcome worse).

Measures of process are important because they help to validate the theory behind a change. Theories are important because only if you *understand* the system will you be able to both reduce variation to within acceptable limits and successfully adapt improvements to new settings. Thus, it is important not only to develop a change idea but also to have some idea of why you expect the change idea to result in an improvement.

The second component of the improvement framework is the iterative methodology for testing changes, the PDSA cycle. PDSA cycles are small tests of change that use the scientific method (see Figure 6.1).[2] Importantly, PDSA cycles are conducted on an appropriate scale and in the

[1] This model has been introduced into education most prominently by Bryk et al. (2015).
[2] Much has been written elsewhere about the PDSA cycle. See Langley et al. (2009) and Rother (2009) for two alternative yet complementary formulations.

Figure 6.1. Overview of the PDSA Cycle



Source: Langley, Nolan, & Nolan, 1994

actual site where the system operates (Langley et al., 2009; Moen & Norman, 2006). Thus, from the very beginning, changes in the system that succeed have some ecological validity. If the costs of failure are high, or readiness for change is low, it's wise to start with very small-scale tests. Later, as the knowledge base grows, it will be possible to scale up to more sites (Parry, 2014; Rossi, 1987). The goal is to avoid going out in a Hail Mary flame of defeat based only on minimal evidence.

Each PDSA cycle is typically carried out over a brief period—the smallest test possible to enable the team to learn from the work. Cycles end with a decision to adopt, adapt, or abandon the change. Usually, the decision is to adapt; enough is learned to suggest how further modifications might lead to better outcomes with lower variation.

Applications to Teaching

The improvement methodology described here has been successfully used in education, though not widely. The most famous example is Japanese lesson study (Lewis, 2000, 2015; Stigler & Hiebert, 1999/2009). In lesson study, the unit chosen as the focus of the improvement work is a single classroom lesson. The lesson is chosen, we surmise, because it is the smallest authentic unit that includes all relevant aspects of the system of teaching that is to be improved. Thus, it has ecological validity, ensuring that the changes developed and the generalizable mechanisms discovered in the context of a single lesson might be applicable to other sites and to other lessons.

But lesson study is difficult to implement. In the United States, the culture of teaching and the organization of the teacher's work week make it difficult for teachers to work together on improving teaching. The lack of a national curriculum (which they have in Japan) means that the work done at one site may be less applicable to other sites, reducing the efficiencies of the improvement process. And even in Japan, lessons only partly meet the requirement of being a repeatable process—something fundamental to the assumptions of improvement science. If something cannot be repeated continuously, it is difficult to engage in the repeated tests (i.e., PDSA cycles) needed to yield results over time. It is true that teachers teach lessons every day. But the content of the lessons changes over the course of a year. So although it is possible to test more general aspects of pedagogy, there are limitations on what can be tested in an iterative fashion, at least in the short term, by a single teacher.

Online Learning as a Wind Tunnel

Much has been written about the development of the airplane (e.g., Baals & Corliss, 1981; Bradshaw, 2005; McCullough, 2015). Without getting into the controversies of interest to historians, it is interesting to note that at least some historians see the development of the wind tunnel as a critical event in aviation history. Prior to the advent of the wind tunnel, which was invented at the end of the 19th century, aviators would build flying machines and then launch them, usually with themselves attached. The cost of failure was high: each time a plane crashed it would take a long time to rebuild it, not to mention the effect failure had on test pilots. One could describe these trials and errors as a sequence of PDSA cycles. But it was a slow process. With a wind tunnel, a model plane could be built, tested, and modified within a relatively short period of time and with a substantial reduction in risk and expense. Advances after the wind tunnel were rapid.

We propose that online learning provides a wind tunnel for the improvement of teaching. If classroom teaching is implemented face to face, one teacher with many students, teaching a particular curriculum, it is hard to make iterative changes and test them with students. If one is testing a change in a particular lesson, which is part of a particular unit, it generally isn't possible to test a new change until the next time that lesson comes along, which might be a semester or even a year later.

Online education, on the other hand, presents us with new opportunities for research and improvement. Teachers participating in the improvement project can collectively design online lessons (e.g., videos with interactive prompts) and study students' learning outcomes. Based on what they find, they can design changes, incorporate them into the lesson material, and provide the revised lesson to a new group of students, immediately. Individual students can be randomly assigned to get different instruction, and thus teachers can study variation both within and across groups, on a

continuous basis. And, because online lessons can be accessed from anywhere at any time, the number of students who could be included in improvement research could, conceivably, be quite large. Like an airplane in a wind tunnel, iterative testing of instructional methods and materials can be conducted rapidly. We don't deny that the fine-tuning of the lessons will need to take place in live classrooms, but much of the work can be completed online. What the wind tunnel provides is a mechanism for implementing more rapid PDSA cycles—more rapid than can be implemented in regular classrooms, alone—enhancing and augmenting the improvement process.

Improvement methodologies are easily applied to online teaching. The Model for Improvement, PDSA cycles, and the statistical techniques of improvement science, discussed previously, might all be brought to bear on the study of online teaching. And if all we learn is how to make better online courses, that would be a worthwhile pursuit. In contrast with aviation, in the case of online learning, the wind tunnel is itself a meaningful end point. All that is learned there can be directly applied to the explosion of online instructional resources. But as a wind tunnel, online learning can also provide us with a laboratory model to use for understanding and improving the more general processes of teaching and learning, whether they be implemented in a live classroom or in a virtual environment. An example of how this might work comes from a recent project we have been developing in our lab.

Example: Learning from Instructional Conversations

An important goal of education is understanding: We want our students not only to learn the facts and procedures of a domain but we also want them to understand the core concepts that underlie the procedures and organize the domain. Research comparing novices with experts reveals that experts see problems differently than novices. Chi, Glaser, and Rees (1982), in a classic study, found that novices tend to classify physics problems based on their surface features (e.g., "rotational things," "pulleys and weights," or "objects on an inclined plane"), whereas expert physicists classify problems according to the physics principles at work (e.g., "conservation of energy law" or "Newton's Second Law"). Connecting concepts to problems makes the experts' knowledge more flexible and powerful in novel problem situations.

Producing students who understand turns out to be highly challenging. We know from extensive research in cognitive psychology that practices such as self-explanation that engage learners in actively connecting problems and procedures to concepts do promote learning with understanding. But we also know that attempts to create these kinds of practices in classrooms have proven extremely difficult. Gallimore and colleagues (Tharp & Gallimore, 1991), for example, identified a classroom discourse pattern they called the "instructional conversation." They also found that through

intensive work with teachers, they could create this kind of discourse pattern from scratch. But they failed, in the end, to find a way to implement instructional conversations at scale (an obstacle discussed also by Rossi, 1987). The cultural nature of teaching makes it nearly impossible to change something as deeply internalized as the routines of talk that define teacher–student interactions.

The fact that instructional conversations cannot be created at scale is problematic in two ways. First, if we cannot create instructional conversations in a large number of classrooms, it will be very difficult to research how such discourse patterns affect students' processing of, and learning from, classroom instruction (Gallimore & Santagata, 2006). Second, even if we assume that such instructional forms are highly effective for producing deep understanding, unless we can create such instructional conditions at scale we still will not be able to produce the kind of learning we desire. Which leads us to ask: Can we, by using online learning technologies, build a model of the instructional conversation (similar to the model airplane that one might build) and a wind tunnel in which to test it?

This has been a focus of recent work in our lab. First, we are attempting to study how students learn from instructional conversations by simulating their participation in such conversations online. Our first step in this work is to create and implement classroom lessons that exemplify the kind of instructional conversations we are interested in studying. In one project, carried out by Belinda Thompson, we created a series of lessons on algebraic expressions and equations. These lessons were designed for community college students taking a beginning algebra class—a developmental mathematics course designed to prepare students for college-level mathematics. The lessons were taught by an expert teacher to a group of community college students and videotaped. The teacher engaged students in a series of rich instructional conversations that focused on core concepts of beginning algebra.

The videos of the class were next uploaded to the cloud and then turned into online instructional modules by embedding interactive prompts and questions into the video. By having a different group of students engage with the online modules, we thus have created a simulation model that creates at least a semblance of what the experience of engaging in an instructional conversation might be like. No, it is not a perfect model, just as a tiny airplane placed in a wind tunnel would not have room for passengers! But it is an experience that can be more easily studied than a live classroom experience.

We are just beginning our studies using these videos. In the studies, we can have students study the videos as if they were participants in the live class. We can, using interactive software, have the video stop, for example, just as a student in the class makes a comment. And then we can have the student watching the video respond with the comment they would make were they in the class. Students can be randomly assigned to watch

different video clips and to respond to different prompts and questions in the same video. For example, we might ask one group of students to solve a problem posed in the video, another to explain its solution to a hypothetical student, and a third to represent the solution graphically. In this way, we (i.e., our small team of researchers and community college instructors) can rapidly develop and test hypotheses about the factors that govern students' thinking as they process the contents of a rich mathematical discussion and we can adapt lessons to reflect what we learn. Applying the techniques of improvement science, we can, over time, optimize students' learning from such instructional conversations. The results of this work can be applied in two settings: first, to the development of better online learning experiences that can more easily be deployed at scale; second, to the design of better curricula to guide live classroom discussion.

Concluding Thoughts

The methods of improvement science have great potential for improving teaching and learning. But realizing that potential has been difficult, in part, because of the nature of education systems. Settings in which a single teacher works with the same group of students over an extended period of time are not easily subjected to improvement methodologies, which require, above all, that there be a repeatable process that can be iteratively studied and improved. Simply put, the U.S. school system, as it currently exists, makes it difficult for improvement science to scale and spread as an internalized learning system. Online learning, to a large extent, offers a partial solution to this problem, making it possible to define repeatable instructional routines that are subject to experimental control. Teachers, the process owners, can be as involved in the study of online teaching as they would have been the study of their own classroom lessons. Because they will still need to adapt the online lessons to work in the context of their classrooms, online learning isn't a final solution, but because online learning can be used to more efficiently identify and test possible improvements that can be adapted to individual classrooms, it seems to us a very good start.

In this sense, online learning, finally, provides us with a wind tunnel that, though it cannot teach us everything we need to know, can provide us a way to advance knowledge and optimize learning without the risks of crashing. Its use is scalable, with at least some degree of ecological validity, and it offers opportunities for random assignment without sacrificing consistency in implementation. When combined with the processes of improvement science, online instructional modules can support the study of the complex system that is teaching. We are only just beginning to understand the potential of this work.

It is important to note that this kind of research and improvement cannot be done by researchers alone. It is true that researchers are a fertile

source of change ideas, mainly because they can bring theories to bear on understanding the mechanisms that produce learning. But as John Dewey (1929) pointed out long ago, research itself—especially research carried out in laboratories—will never produce ready-made rules to guide the improvement of teaching. Research, Dewey said, can make us sensitive to the factors that interact to produce learning. But simple rules will never be enough to tame the complexities of a system as complex as education.

References

Baals, D. D., & Corliss, W. R. (1981). *Wind tunnels of NASA*. Washington, DC: National Aeronautics and Space Administration.

Berwick, D. (2015). *Keynote address*. Carnegie Foundation Summit on Improving Education, San Francisco, CA.

Bjork, E. L., & Bjork, R. A. (2011). Making things hard on yourself, but in a good way: Creating desirable difficulties to enhance learning. In M. A. Gernsbacher (Ed.), *Psychology and the real world: Essays illustrating fundamental contributions to society* (pp. 56–64). New York, NY: Worth Publishers.

Bradshaw, G. (2005). What's so hard about rocket science? Secrets the rocket boys knew. In M. Gorman, R. Tweney, D. Gooding, & A. Kincannon (Eds.), *Scientific and technological thinking* (pp. 259–275). Hillsdale, NJ: Erlbaum.

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.

Chi, M. T. H., Glaser, R., & Rees, E. (1982). Expertise in problem solving. In R. Sternberg (Ed.), *Advances in the psychology of human intelligence* (Vol. 1, pp. 7–76). Hillsdale, NJ: Erlbaum.

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

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

Gallimore, R., & Santagata, R. (2006). Researching teaching: The problem of studying a system resistant to change. In R. R. Bootzin & P. E. McKnight (Eds.), *Strengthening research methodology: Psychological measurement and evaluation* (pp. 11–28). Washington, DC: APA Books.

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

Gawande, A. (2010). *The checklist manifesto: How to get things right*. New York, NY: Metropolitan Books.

Givvin, K. B., Hiebert, J., Jacobs, J. K., Hollingsworth, H., & Gallimore, R. (2005). Are there national patterns of teaching? Evidence from the TIMSS 1999 video study. *Comparative Education Review*, *49*(3), 311–343.

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

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

Kornell, N., & Bjork, R. A. (2008). Learning concepts and categories: Is spacing the "enemy of induction"? *Psychological Science*, *19*, 585–592.

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*. San Francisco, CA: Jossey-Bass.

Langley, G. J., Nolan, K. M., & Nolan, T. W. (1994). The foundation of improvement. *Quality Progress*, *27*(6), 81–86.
Lewis, C. (2000). *Lesson study: The core of Japanese professional development*. Paper presented at the Annual Meeting of the American Educational Research Association, New Orleans, LA.
Lewis, C. (2015). What is improvement science? Do we need it in education? *Educational Researcher*, *44*(1), 54–61.
Lipsey, M. W. (1993). Theory as method: Small theories of treatments. In L. B. Sechrest & A. G. Scott (Eds.), *New Directions for Program Evaluation: No. 57. Understanding causes and generalizing about them* (pp. 5–38). San Francisco, CA: Jossey-Bass.
McCullough, D. (2015). *The Wright brothers*. New York, NY: Simon & Schuster.
Moen, R., & Norman, C. (2006). *Evolution of the PDCA cycle*. Retrieved from http://pkpinc.com/files/NA01MoenNormanFullpaper.pdf
Nuthall, G. (2005). The cultural myths and realities of classroom teaching and learning: A personal journey. *Teachers College Record*, *107*(5), 895–934.
Parry, G. J. (2014). A brief history of quality improvement. *Journal of Oncology Practice*, *10*(3), 196–199.
Rossi, P. (1987). The iron law of evaluation and other metallic rules. *Research in Social Problems and Public Policy*, *4*, 3–20.
Rother, M. (2009). *Toyota kata: Managing people for improvement, adaptiveness and superior results*. San Francisco, CA: McGraw-Hill Professional.
Stigler, J. W., & Hiebert, J. (1998). Teaching is a cultural activity. *American Educator*, *22*(4), 4–11.
Stigler, J. W., & Hiebert, J. (1999/2009). *The teaching gap: Best ideas from the world's teachers for improving education in the classroom*. New York, NY: Simon and Schuster.
Tharp, R. G., & Gallimore, R. (1991). *Rousing minds to life: Teaching, learning, and schooling in social context*. New York, NY: Cambridge University Press.

JAMES W. STIGLER is professor of psychology at the University of California, Los Angeles, director of the TIMSS video studies, and founder of LessonLab Inc.

KAREN B. GIVVIN is a researcher and adjunct professor at UCLA, in the Department of Psychology. Her research focuses on better understanding students' mathematical knowledge—especially their misunderstandings—and how teachers might use that information to improve instruction.