Lemire, S., Christie, C. A., & Inkelas, M. (2017). The methods and tools of improvement science. In C. A. Christie, M. Inkelas & S. Lemire (Eds.), *Improvement Science in Evaluation: Methods and Uses. New Directions for Evaluation*, *153*, 23–33.

2

The Methods and Tools of Improvement Science

Sebastian Lemire, Christina A. Christie, Moira Inkelas

Abstract

Rooted in ideas from operations research in the 1930s, improvement science bloomed in the healthcare literature during the 1990s and has since then spread rapidly across fields such as management, social work, behavioral economics, and most recently education (Lewis, 2015). So what is thing called "improvement science"? What is the intellectual foundation of improvement science? And what does it look like in real-world applications? What, if anything, might we, as evaluators, learn from the techniques and tools of improvement science? These are questions that will be addressed in this chapter. © 2017 Wiley Periodicals, Inc., and the American Evaluation Association.

Toward a Definition of Improvement Science

Improvement science means many different things to many different people. Perhaps because of the rapid cross-field fertilization, the term "improvement science" has often been used interchangeably with terms such as "science of improvement," "continuous improvement," "system improvement," and even "scientific quality improvement," to name but a few (Health Foundation, 2011). Despite this rich and diverse terminological landscape, or perhaps as a result thereof, harvesting an explicit definition of improvement science is not an easy task. As noted by Marshall, Provost, and Dixon-Woods (2013), the lack of consensus on a definition may just

indicate the preparadigm phase in which improvement science currently resides, despite its growing popularity.

The label "science of improvement" emerges with Langley and colleagues' publication of *The Improvement Guide* in 1996 (Perla, Provost & Parry, 2013). Without offering an explicit definition of the term, Langley et al. (2009) identify William E. Deming's "system of profound knowledge" as the intellectual foundation for improvement science (p. 75). Following Deming, a system of profound knowledge is structured around four types of knowledge:

1. Knowledge of systems
2. Knowledge of psychology
3. Knowledge of variation
4. Knowledge of how knowledge grows

Given the foundational role of these four types of knowledge, brief consideration of what is meant by each is called for. Knowledge of systems refers to an understanding of systems as "an interdependent group of items, people, or processes working together toward a common purpose" (Langley et al., 2009, p. 77). For the improvement scientist, consideration of these interdependencies is central when designing, testing, and implementing changes. As noted by Langley et al. (2009), "considering interdependence will also increase the accuracy of our predictions about the impact of changes throughout the system"—a central aim of improvement science (p. 78).

In tandem with knowledge of systems, knowledge of psychology, understanding the human side of change, speaks to the importance of understanding how and in what way interpersonal and social structures influence system processes and performance when designing and implementing changes. Individuals may react or commit to, integrate or expunge, reject or support changes to a system. As such, deploying methods and tools that support the human aspect of change are more likely to lead to successful and sustained improvement.

Another central knowledge component, especially in relation to the measurement of change, is knowledge of variation. As noted by Langley et al. (2009), knowledge of variation involves a distinction between variations in system performance stemming from designed change (special cause variation) versus variations stemming from naturally occurring change (common cause variation). Separating the two types of variation, as well as determining whether a system is influenced by one or the other (or both), is central to testing change.

Finally, knowledge of how knowledge grows is central to ensure successful improvements. Central to this end is the role of predictions about which changes will result in improvements. As Langley and colleagues (2009) remind us, "The more knowledge one has about how the

particular system under consideration functions or could function, the better the prediction and the greater the likelihood the change will result in improvement" (p. 81). Building knowledge then, relies on the ability to compare predictions about changes with empirical results (Langley et al., 2009).

Returning to the topic of how to define improvement science, Deming's system of profound knowledge may still fall short of a satisfying definition. This is in large part because it does not specify what improvement science is or is not. The four types of knowledge are simply too broad in their potential application to serve well in this regard. Moreover, the purposes and functions of improvement science are also left unstated. From this perspective, Deming's system of profound knowledge is perhaps better viewed as the intellectual foundation for improvement science, the pillars on which improvement science is grounded.

Resting on the intellectual foundation from Deming, we may consider the two core features of improvement science provided by Langley and colleagues:

1. The idea that improvement emerges from developing, testing, implementing, and spreading change, and
2. The recognition that subject matter experts play a lead role in defining and informing each of those four steps (cited in Perla et al., 2013).

Stated differently, improvement science is about developing, testing, implementing, and spreading change informed by subject matter experts. The orientation toward change is echoed in the observation that improvement science is "a type of practical problem solving, an evidence-based management style, or the application of a theory-driven science of how to bring about system change" (Margolis, Provost, Schoettker, & Britto, 2009, p. 832). From this perspective, improvement science is situated somewhere between change management and research (Health Foundation, 2011).

Informed by the contributions cited here, a working—or at least workable—definition of improvement science for the purpose of this special issue may be offered. Inspired by Langley et al. (2009), among others, we define improvement science as:

> A data-driven change process that aims to systematically design, test, implement, and scale change toward systemic improvement, as informed and defined by the experience and knowledge of subject matter experts.

Admittedly, this is a working definition of improvement science for the purpose of this special issue—to compare and contrast improvement science with evaluation. As is evident from the proposed definition, several aspects of improvement science resonate with key aspects of traditional definitions of evaluation, including the systematic application of data, the use

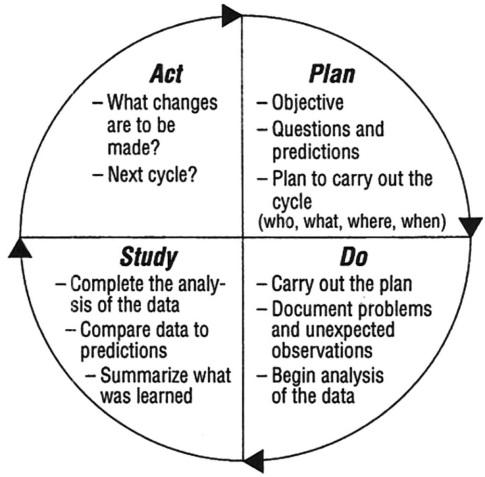
of data to test and scale up changes, the focus on systemic improvement, and the involvement of stakeholders. In this way, much of what improvement science *is* is evaluation. In this volume, we take up these issues and draw the connections necessary for the evaluation field to consider and use improvement science more widely as a strategy for improving program processes and outcomes.

The Model for Improvement—An Operational Framework for Practice

Another fruitful way of understanding improvement science is to examine the principles and operational framework that structure and support the practical application of improvement science. Painting in broad strokes, and expanding on the conceptual grounding laid out in the preceding section, Perla, Provost, and Parry (2013) identify seven propositions that provide the methodological foundation for the science of improvement:

1. The science of improvement is grounded in testing and learning cycles—an approach that in its practical application is structured around repeated Plan–Do–Study–Act (PDSA) cycles (see the subsequent section on the Improvement Science Toolbox).
2. The philosophical foundation of the science of improvement is conceptualistic pragmatism—an understanding of the importance of combining existing subject matter and theory to make predictions about changes to be implemented and tested.
3. The science of improvement embraces a combination of psychology and logic (i.e., a weak form of "psychologism")—an acknowledgment that psychology paired with analytical philosophy, logic, and mathematics provides the grounding for a stronger understanding of multiple dimensions of change.
4. The science of improvement considers the contexts of justification *and* discovery—an understanding that improvement emerges from the interplay between inductive and deductive logic, procedures of discovery and justification (see subsection on logic of PDSA cycles).
5. The science of improvement requires the use of operational definitions—a belief in the importance of conceptual clarity and shared understanding of what improvement is.
6. The science of improvement employs Shewhart's theory of cause systems—a focus on distinguishing between stable and unstable systems, special and common cause variation.
7. Systems theory directly informs the science of improvement—an appreciation that all change takes place in the context of a dynamic and adaptive system, why understanding the system's composition is a fundamental condition for improvement.

Figure 2.1. The Plan–Do–Study–Act Cycle



Source: Adapted from Moen, Nolan, & Provost, 2012

The strong association these principles have with the social sciences more generally and evaluation more specifically is considered by Christie, Lemire, and Inkelas (Chapter 1). Collectively, the principles serve to inform the nature of improvement science and in effect to guide improvement science practitioners and theorists (Perla et al., 2013).

In its real-world application, improvement science is framed by the Model for Improvement and structured around PDSA cycles. The Model for Improvement specifies three framing questions for improvement projects:

1. What are we trying to accomplish?
2. How will we know that a change is an improvement?
3. What change can we make that will result in improvement?

The primary function of these questions is to develop changes that will lead to sustained improvement within a system. As Langley et al. (2009) remind us, "Not all changes lead to improvement, but all improvement requires change"—central to improvement science then is to recognize and bridge the difference between the two (p. 357). Toward this aim, the Model for Improvement is realized through the Plan–Do–Study–Act cycle, a reiterative trial-and-learning process that connects empirical learning with redesign (Langley et al., 2009; Morris & Hiebert, 2011). A generic PDSA cycle is provided in Figure 2.1.

The first step in the cycle is to clearly state the objective of the PDSA cycle as well as the corresponding questions to be answered. Toward this

aim, the first step also involves the development of an operational plan that details where, when, and by whom the cycle will be implemented. A key component of the plan is a specification of the data collection to be carried out.

Step two in the PDSA cycle revolves around the implementation of the plan. To ensure a systematic and transparent process, documentation of challenges or issues emerging as part of the implementation of the PDSA cycle should be documented. These include any issues related to the data collection.

In the third step of the cycle, attention is awarded the results of the data collection. More specifically, observed patterns in the data are compared with the predicted patterns to identify similarities and contradictions. The aim is to determine whether the data support or undermine the predictions made based on past knowledge and experience.

Informed by this new knowledge, step four provides the opportunity to make additional changes or modifications to the designed change, before (re)running the PDSA cycle. The modifications to be made should be grounded on whether or not the previous steps promoted improvements (however, these are defined under step one). By doing so, a "learning loop" is created, in which iterative rounds of developing, testing, and implementing changes can take place (Langley et al., 2009).

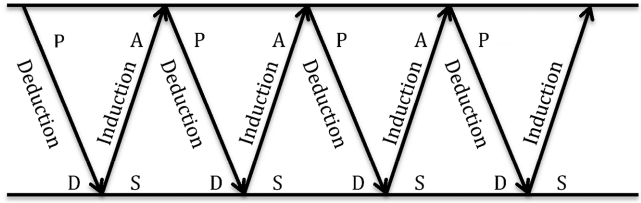
The PDSA cycle can be implemented in many different ways, depending on the specific purpose, context, and conditions of the project. Perhaps needless to say, there is no single right way to carry out PDSA cycles. That being said, Langley et al. (2009, p. 145) highlight three principles for the rigorous "testing of change":

Principle 1: Test on a small scale and build knowledge sequentially
Principle 2: Collect data over time
Principle 3: Include a wide range of conditions in the sequence of tests

Structuring improvement science projects around sequential PDSA cycles is compelling for several reasons. For one thing, the cycle involves a both inductive and deductive reasoning. The interplay is illustrated in Figure 2.2. A deductive approach is deployed when articulating predictions (the "Plan" step) and departures from these are observed (the "Do" step) as part of the PDSA cycle. Subsequently, inductive learning emerges in the "Study" and "Act" steps when divergences between the predictions and the observed outcomes are translated into a set of revised predictions.

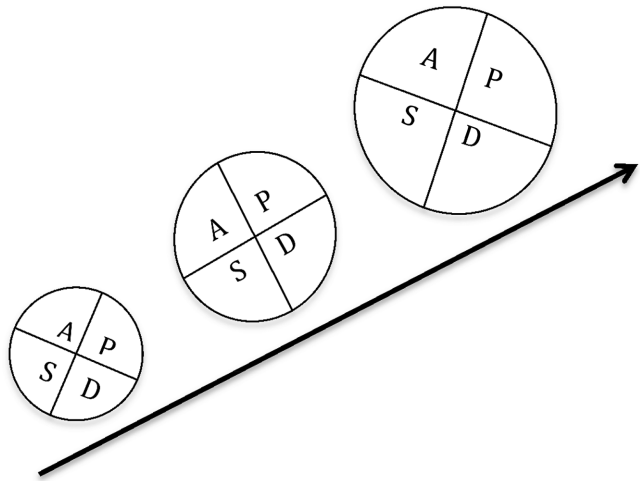
Second, the sequential use of PDSA allows for a trail of evidence; advancing a string of small, mutually informed experiments. These trails of evidence may sometimes even develop from small to increasingly larger changes and more formal tests. In this way, the underlying logic of sequential testing aligns closely with Campbell and Stanley's framework for designs (as preexperimental, experimental, or quasiexperimental) and can be traced

Figure 2.2. The Interplay of Inductive and Deductive Logic



Source: Adapted from Langley et al., 2009

Figure 2.3. Sequential PDSA Cycle



Source: Adapted from Moen, Nolan, & Provost, 2012

even further back to the fundamental principles of the scientific method as articulated by Aristotle and Copernicus, among others (Perla et al., 2013). A generic illustration of sequential experimentation using the PDSA cycle is provided in Figure 2.3.

Third, the progression from smaller to increasingly larger changes also serves to dampen the fear of failure. This is because the piecemeal introduction of small changes constricts the potential adverse consequences of harmful changes. After all, only changes that have proven successful during small-scale testing are further developed and scaled up for the purpose of affecting the system as a whole.

Fourth and finally, and as noted by Langley et al., "Satisfactory prediction of the results of tests conducted over a wide range of conditions is the means to increase the degree of belief that the change will result in improvement" (2009, p. 141). The credibility of the tests can be enhanced

by manipulation (e.g., by removing or alternating the change), factorial design strategies, theory (e.g., grounding the change in theory of change), and replication across diverse settings. Again, the sequential nature of the PDSA cycles lends itself well to this incremental testing and confidence building.

Leaving all these compelling features aside, the PDSA cycle is not without its shortcomings. One criticism raised by Langley et al. (2009) is that the small-scale cycles tend to fail to produce impact at the systemic level, which after all, is what improvement science aims for. As noted by Langley et al. (2009), the "small-scale" refers to the testing and not necessarily to the change introduced; the latter may represent a significant departure from practice as usual (p. 102). Another approach to lessen the concern is to coordinate multiple PDSA cycles that collectively promote changes at the system level.

Another issue relates to the varied use and real-world application of PDSA cycles, is that of different degrees of compliance with guidelines for good PDSA practice and reporting, resulting in a lack of transparency about the iterative cycles of improvement, among other things (Taylor et al., 2014). As Taylor and colleagues point out, studies that use PDSA as a "black box" intervention should be cautioned against.

The Control Chart—A Central Tool in the Improvement Science Toolbox

A third way of understanding improvement science is by considering the core tools that support and characterize improvement science. A plethora of tools and methods have been developed to support different stages of the improvement science (and even specific steps of the PDSA cycle). These include tools for developing a change, testing a change, implementing a change, and spreading a change. Many of these are illustrated in the case chapters comprising this volume. Interested readers are also encouraged to find inspiration in the comprehensive list of improvement science tools provided in the appendix of the improvement guide (Langley et al., 2009). However, given the purpose and page limits of the present volume, this is not the place to consider all of these. Instead, consideration is given to the use of control charts—a useful (yet relatively rare tool) in the context of evaluation.

Control Charts—What Are They?

Control charts, also referred to as process behavior charts or Shewhart charts, comprise a central statistical tool in improvement science. Developed by Walter Shewhart in the context of improving production lines, control charts graphically depict outcome patterns over time. Control charts

typically consist of a centerline (e.g., a mean or media) and a set of corresponding control limits (e.g., ±2 standard deviations of the centerline).

At root, control charts are about analyzing process variation over time. As noted by Moen, Nolan, and Provost (2012, p. 286), control charts offer a formal approach for distinguishing between:

- *Common cause variation (noise)*. This is variation stemming from causes that are inherent in the system (process or product) over time, affect everyone working in the system, and affect all outcomes of the system.
- *Special cause variation*. This is variation rooted in causes that are not part of the system (process or product) all the time or do not affect everyone, but arise because of specific circumstances.

The upper and lower control limits in control charts reflect the boundary between these two types of variation. Outcome patterns within the boundaries are considered expected, natural variation in the system, whereas outcome variations outside of these boundaries are considered signals of special cause variation, often subject to further analysis. In the context of improvement science, this demarcation is of central importance because it allows for the identification of improvements stemming from system changes. In this way, the upper and lower control limits in control charts are pivotal in distinguishing between random variation (i.e., noise) and special cause variation, reflecting "true" signals of change.

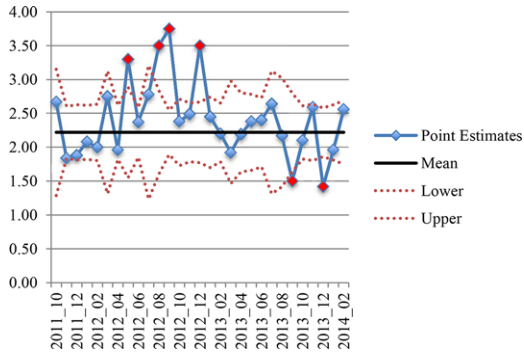
For the purpose of establishing upper and lower control limits in control charts, a baker's dozen of guidelines has been suggested, including:

- ±3 sigma from the centerline
- ±2 standard deviations of the centerline
- ±3 standard deviations of the centerline

These guidelines are arbitrary in the sense that their application tends to be based on equal parts convention and subjective preference. As just one example, and as noted by Murray and Provost (2011), the three-sigma guideline is grounded on "experience" rather than statistical theory (p. 160). The statistically oriented reader will probably recognize the other two guidelines' reliance on statistical theory. In real-world practice, the use of ±3 standard deviations of the centerline appears to be the most prevalent among improvement science practitioners, at least within the context of improvement science projects in the health sciences.

An illustrative example of a control chart is provided in Figure 2.4. The dataset supporting the estimation of the control charts stems from the Magnolia Community Project—an improvement science project in the social welfare sector. The outcome variable of interest is a coverage score, representing the degree of care-related concerns covered during meetings between service providers and clients. The coverage score ranges from 0

Figure 2.4. Control Chart for Coverage Scores (by Month)



to 4 (with 4 representing complete coverage) for each of these meetings. A total of 850 individual meetings were scored in the period from October 2011 to February 2014.

As Figure 2.4 shows, several point estimates indicate special cause variation: four point estimates indicate variation above the expected common cause variation and two point estimates indicate variation lower than expected common cause variation. In the context of improvement science, these point estimates would motivate further analyses to identify the special cause(s) that produced the observed pattern. For example, the two figures might prompt an analysis of any systematic changes that were made to the organization in the summer of 2012, aiming to identify the cause that generated the pattern in the data. This would involve allocation of human resources and time and potentially lead to conclusions and decisions on future systematic changes to be made in the organization.

In summary, improvement science cannot be defined by any one process or procedure. For the purpose of this volume, we define improvement science as a data-driven change process that aims to systematically design, test, implement, and scale change toward systemic improvement, as informed and defined by the experience and knowledge of subject matter experts. In its practical application, improvement science is framed by the Model for Improvement and structured around PDSA cycles. A central tool for identifying special cause variation is the control chart.

References

Health Foundation. (2011). *Report: Improvement science*. Retrieved from http://www.health.org.uk/publication/improvement-science

Langley, G. J., Moen, R. D., Nolan, K. M., Nolan, T. W., Norman, C. L., & Provost, L. P. (2009). *The improvement guide* (2nd ed.). San Francisco, CA: Jossey-Bass.

Lewis, C. (2015). What is improvement science? Do we need it in education? *Educational Researcher*, *44*(1), 54–61.

Margolis, P., Provost, L. P., Schoettker, P. J., & Britto, M. T. (2009). Quality improvement, clinical research, and quality improvement research—Opportunities for integration. *Pediatric Clinics of North America*, *56*, 831–841.

Marshall, M., Provost, L. P., & Dixon-Woods, M. (2013). Promotion of improvement as a science. *Lancet*, *381*, 419–421.

Moen, R. D., Nolan, T. W., & Provost, L. P. (2012). *Quality improvement through planned experimentation*. New York, NY: McGraw Hill.

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

Murray, S. K., & Provost, L. P. (2011). *The health care data guide: Learning from data for improvement*. San Francisco, CA: Jossey-Bass.

Perla, R. J., Provost, L. P., & Parry, G. J. (2013). Seven propositions of the science of improvement: Exploring foundations. *Quality Management in Health Care*, *22*(3), 170–186.

Taylor, M. J., McNicholas, C., Nicolay, C., Darzi, A., Bell, D., & Reed, J. E. (2014). Systematic review of the application of the plan-do-study-act method to improve quality in healthcare. *BMJ Quality and Safety*, *23*, 290–298.

SEBASTIAN LEMIRE is a doctoral candidate in the Social Research Methodology Division in the Graduate School of Education and Information Studies, University of California, Los Angeles.

CHRISTINA A. CHRISTIE is professor and chair of the Department of Education in the Graduate School of Education and Information Studies, University of California, Los Angeles.

MOIRA INKELAS is associate professor in the Department of Health Policy and Management in the Fielding School of Public Health, University of California, Los Angeles, and assistant director of the Center for Healthier Children, Families and Communities.