PART 1

Seeking Change in a Complex World

In the mid-1970s, Ernö Rubik created a puzzle that he could not easily solve. He was a professor at the Academy of Applied Arts in Budapest Hungary and wanted a tool to help his students learn three-dimensional geometry and movement (Simpson, 2015; Wallop, 2014). Using materials found in his mother's home, he designed a cube-shaped model with nine primary color squares on each side. He named it the "Magic Cube." Now, he was stumped. As he moved one side, another side would move. The movement of one colored square influenced the movement of another square. With each move, the brightly colored squares became more jumbled. There were 43 quintillion possible permutations. It took him a month to solve it.

If you are a child of the 1980s, you probably remember trying your hand at Rubik's Cube. If you were like me, you couldn't do it. A few of our fellow Generation Xers got creative and tried a different approach: They simply pried off the jumbled color pieces and re-glued them to match on each side. Problem solved! Or was it?

What those industrious Generation Xers might not remember, or perhaps never knew, was that by taking the cube apart and rebuilding it, you could actually render the puzzle unsolvable. Rubik's Cube purists will tell you that scrambling the pieces throws off the orientations of the edge and corners pieces. It is a system of interconnected pieces.

Three decades later, communities across the United States were facing a puzzle of a different, more dangerous sort. In 2018, there were approximately 47,000 opioid overdose deaths in the United

States (Centers for Disease Control and Prevention [CDC], 2020a). Opioid use was not a new epidemic. Between 1999 and 2018, the number of opioid-related deaths had increased drastically (CDC, 2020a). Numerous federal, state, and local efforts had been undertaken for years to improve the problem. Promising solutions were known. For instance, the state of Washington passed prescription reform legislation that resulted in a "27 percent reduction in the number of overdose deaths between 2008 and 2012" (Martin et al., 2016, p. 6).

However, as legislative efforts addressed the over-prescription of opioids, new problems emerged, creating three waves of opioid overdose deaths (CDC, 2020b). The first wave of deaths occurred because of the rise of prescription opioids in the 1990s. The second wave began around 2010 due to increased heroin use. And the third wave, beginning in 2013 with the sharpest increase, was the result of synthetic opioids such as fentanyl. As prescription abuse decreased, illicit use increased. As law enforcement battled heroin trafficking, fentanyl was introduced. Like a Rubik's Cube with life and death consequences, a change in one area led to a consequence in another. The problem was always changing and thus, remained persistent, and the solution remained elusive.

How does one approach problems in today's complex society? Solving persistent and difficult problems requires a systematic and systemic approach that helps us make continuous moves toward improving the problem. One intervention, one new program, one policy change is unlikely to solve the problem on its own.

During the first wave of opioid crisis, there were numerous interventions that were shown to be effective; however, many were not systemic or continuous. That is, they did not account for the next problem that would arise, whether due to an unintended consequence or another change in the system's landscape. According to an Institute for Healthcare Improvement, the lack of a systems-wide view of the problem was one of the most significant drivers of the crisis (Martin et al., 2016). Among others, they identified these reasons for why the crisis continued:

- Lack of coordination of approaches and resources: They noted that many of the intervention initiatives were siloed and only addressed one part of the problem.
- Lack of effective implementation of promising practices: They suggested that the continued crisis was not due to the lack of knowledge, evidence-based strategies, and recommendations for action, but rather the lack of support for executing these strategies and recommendations.
- Failure to engage necessary communities and stakeholders: Importantly, they acknowledged that improvement efforts needed to include those they intended to help—members of the local communities, families, individuals—along with law enforcement, faith-based organizations, and schools.

As evaluators seeking to lead change, we can easily fall victim to the above failures by focusing on individual interventions, programs, and policies rather than the problem as it exists in a multifaceted complex system.

Part 1 of this book provides a foundational grounding by introducing formative evaluation, continuous improvement, and systems and complexity science theory. Evaluators commonly look to formative evaluation as a guiding framework for change or improvement. Yet to lead change in complex systems, evaluators need to distinguish formative approaches that embrace and provide methods for understanding and being responsive to emerging challenges and complexity.

What do we really mean by formative evaluation? It holds a different meaning for different people. Some consider formative evaluation the first step toward a more conclusive summative evaluation. Others use the term as a catch-all phrase for any evaluation aimed at improvement. Chapter 1 of this book provides the history of formative evaluation and unpacks its different meanings in an effort to help evaluators better specify which model best fits their needs. I further posit that evaluators hoping to improve persistent problems in complex systems should consider continuous improvement approaches grounded in improvement science.

Many of us in evaluation are familiar with the idea of continuous improvement. It takes various forms but is generally considered to be any ongoing endeavor to improve products, processes, practices, or services. While continuous improvement is often regarded as a form of formative evaluation, primarily labeling it so can lead to a potential oversight of its defining characteristic; that continuous improvement is, well, *continuous*. There is no end point to the process. More than a broad formative approach for improving one particular program, policy, or practice, continuous improvement lends itself to an evaluative strategy for driving change in complex systems because it is responsive to emergent challenges.

In Chapter 2, I further consider the concept of complex systems by providing background on complexity theory and systems thinking. Like the social systems in which the opioid crisis exists, today's organizations continually confront challenges, whether it be a school district addressing high rates of chronic absenteeism, or a hospital investigating why a high number of patients are acquiring new infections in their care. These organizations exist in complex systems with many interrelated and dependent parts. The actions of one individual, department, or policy influence other people and processes elsewhere in the system. Furthermore, these actions and related outcomes may be unforeseen because system actors adapt to changes and new contextual conditions resulting from this interconnectedness.

Conditions are rarely stable or predictable. Those hoping to solve persistent problems need to engage and honor the experiences and perspectives of those within the system and build their capacity for ongoing disciplined inquiry. By doing so, those on the front lines are empowered to continually identify and respond to emerging issues. Leading change in complex systems requires a philosophical shift in how people work and learn. This book provides practical guidance in how to build this capacity.

Through an in-depth case study in Part 2, I examine the use of continuous improvement as an evaluative strategy in practice and describe how one approach grounded in improvement science was integrated into an improvement network consisting of five schools who hoped to improve mathematics instruction. By doing so, I provide an empirical example of implementing some of the concepts introduced in Part 1, and importantly, share struggles and difficulties that emerged through the process.

Improvement science is a disciplined inquiry process by which the subject matter and the improvement experts (often the evaluator) collaborate to find the root cause of a problem within a system, develop a theory of change for improving it, and rapidly experiment with changes to determine if they lead to improvements. This book's case study addresses real-world improvement science challenges and complexities rather than solely sharing the ideal situation. Instead of directing how the process should be in the ideal context, I discuss what it looks like in an actual context then encapsulate the case study's significant learnings and offer additional lessons learned since researching the case study.

My hope is that this book provides practical guidance to other evaluators seeking to effect positive change in the world. It is a call to action.

Change is hard. Change is messy. Change is hardest and messiest in complex systems where one can't simply force the pieces to fit the solution like a re-glued Rubik's Cube. But change is not impossible if evaluators and practitioners add the right tools to their toolbox. This book offers some of those tools.

1

What Do We Mean by Formative Evaluation?

I became interested in improvement science early in my graduate school career. As a program evaluator in a PhD program, others often asked me how improvement science differed from formative evaluation and developmental evaluation. At the time I did not have good answer. Now, I do.

Purpose distinguishes the approach. This foundational chapter begins our journey of understanding how to lead change through evaluation by delving into the meaning of formative evaluation and how it has evolved over the years. It sets the stage for why continuous improvement methods (e.g., improvement science as a form of formative evaluation) are a valuable approach for driving change in complex systems.

In this chapter, I cover:

- The Evolution of Formative Evaluation
- Continuous Improvement and Improvement Science
- Other Continuous Improvement Approaches

The Evolution of Formative Evaluation

Evaluators often find themselves in one of two broad roles: providing a credible and balanced summative judgment about a program or policy, or formatively supporting the development or improvement of some program, policy, process, or organizational practice. Evaluators conducting summative evaluations engage in systematic inquiry to provide a conclusion about an entity's worth or effectiveness and typically provide findings to stakeholders at the end of an evaluation. Those involved in formative evaluation also engage in systematic inquiry, but the process is designed to deliver timelier data and findings to inform stakeholders about what or how to improve. Both roles are important and can be used for improvement, but this chapter focuses on the latter, using formative evaluation as an evaluative strategy to lead change.

Michael Scriven introduced the term "formative evaluation" in the late 1960s. In *The Methodology of Evaluation* (1966), he importantly distinguished the *goals* of evaluation from the *roles* of evaluation. His paper focused on curricular evaluation, although his points also applied to other kinds of evaluation. Scriven stated that the *goals* of evaluation were to answer questions about how well "certain entities" perform, either on their own, or compared to another (p. 2). The *roles* of evaluation, however, could take various forms. To use Scriven's examples, evaluation could play a role in the development of curriculum or in determining the worth of that curriculum. By making this argument, he distinguished "formative" evaluation aimed at improving a program during its development, from "summative" evaluation that sought to provide final conclusions about a program's value or worth.

Notably, Scriven was arguing a counterpoint to Cronbach (1963), who viewed course improvement as a *primary* purpose of evaluation. Cronbach also focused on education and curriculum, and in this context, improvement meant "deciding what instructional materials and methods are satisfactory and where change is needed" (p. 236).[1] Cronbach posited that "the greatest service evaluation can perform is to identify aspects of the course where revision is desirable" (p. 238) and that "evidence must become available midway in curriculum development . . ." (p. 239). He stated that "[e]valuation, used to improve the course while it is still fluid, contributes more to improvement of education than evaluation used to appraise a product already placed on the market" (p. 239). While Scriven (1966, 1996) believed that formative evaluations were valuable and necessary, they were not a substitute for a final summary judgment about a course or program. Both were important roles of evaluation. For Scriven, and subsequently many other evaluators, improvement was formative evaluation, and its primary purpose was as a step toward preparing a program for a subsequent summative evaluation, not for the sake of improvement itself. Thus, he coined the terms formative and summative evaluation to make a distinction between the two.

However, the meaning of formative evaluation has evolved over the years. Many evaluators use the term to mean any type of evaluative activity aimed at improving a policy, program, or process, and not solely as a step toward preparing them for a summative evaluation. Yet for some evaluators, it is Scriven's initial viewpoint of formative evaluation that prevents them from embracing

[1] Cronbach also listed two other types of decisions for which evaluation is used: (1) "Decisions about individuals: identifying the needs of the pupil for the sake of planning his instruction, judging pupil merit for purposes of selection and grouping, acquainting the pupil with his own progress and deficiencies." (2) "Administrative regulation: judging how good the school system is, how good individual teachers are, etc." (p. 236).

the term more broadly. For example, in 1996, Michael Quinn Patton, a well-known evaluator and former president of the American Evaluation Association, acknowledged that the meaning of formative evaluation "has been enlarged to include any evaluation whose primary purpose is program improvement" (p. 135). However, in that same article, he stopped short of claiming that developmental evaluation was a form of formative evaluation because its purpose was not to prepare for a summative evaluation. Developmental evaluation is an evaluation model that seeks to *develop* innovative programs, products, policy reforms, and organizational changes in complex environments (Patton, 2011). Scriven's definition at the time was too limiting.

Even since then, the concept of formative evaluation has continued to progress and become a catch-all term synonymous with evaluation for improvement. This is, at the very least, partially due to the expanding discipline of evaluation in which there are multiple models and approaches of evaluation that extend beyond Scriven's original dichotomy of formative and summative. However, the concept of what formative evaluation means is still vague.

For some, formative evaluation's purpose is to improve the program, typically as part of understanding implementation or processes before evaluating whether the program is achieving intended outcomes.

In *Evaluation Essentials*, Alkin and Vo (2018) provide this definition of formative evaluation:

> *Formative evaluation* generally takes place during the early stages of program implementation. Formative evaluation is conducted in order to provide information for program improvement, which generally means that the evaluation information would provide an indication of how things are going. (p. 12, emphasis in original)

While Alkin and Vo (2018) seem to subscribe to the more traditional definition of formative evaluation, they also acknowledge that formative evaluation may "take place over extended periods of time" and label this as *continuous* formative evaluation, which is "focused on the ongoing development of a program or innovation in more complex settings" (p. 12). They note that some call this practice "developmental evaluation," thereby either acknowledging a broader conception of formative evaluation or a potential divergence from Patton (1996) on whether developmental evaluation falls under the formative umbrella.

Further, Alkin and Vo (2018) provide another important distinction within the category of formative evaluation: Evaluators only occasionally conduct *final* summative evaluations. Rather, evaluators more often practice what they call "summary formative evaluation" (p. 13). There may be a period of time where formative activities occur, and then the evaluator will summarize findings at the end of that period. Additionally, Alkin and Vo remind us that both processes and interim outcomes (versus end-of-evaluation outcomes) may be included in summary formative evaluations. Many use summary results for program improvement, and many summarize formative results to reach conclusions. And, in fact, it is the use of the information that determines its categorization.

There is a classic maxim in evaluation attributed to Robert Stake:

- When the cook tastes the soup, that's formative.
- When the guest tastes the soup, that's summative.

Alkin and Vo (2018) further this idea by considering that (1) when the cook tastes the soup, they are interested in whether it tastes good (interim summary outcome), and (2) when the guest tastes the soup, the cook may also be interested in the guest's feedback for the purpose of improving the soup the next time they serve it (summary formative evaluation).

Expanding on their ideas, now consider: Should the cook ever stop caring whether the guest likes the soup? Even if the recipe does not change, the context might. Restaurant menus of the 1970s commonly offered pea soup, but today, the humble pea soup has been replaced by carrot ginger, cucumber gazpacho, miso, and other recipes preferred by today's palates. Also, consider the scenario where the restaurant's management changes. The old cook is replaced by a new chef. Might this change how the soup recipe is implemented and subsequently tastes in this new context? Summary conclusions can be ongoing and responsive to the latest needs.

What is considered formative or summative depends on the purpose, use, and context. Scriven (1996) himself made this point. Therefore, extending the concept of formative evaluation more broadly makes practical sense. As Alkin and Vo mention, only occasionally do program designers, staff, and evaluators stop at a final assessment of the program's value or worth.

Many of today's evaluation needs are formative and *ongoing*. That is, programs aimed at improving particular societal problems can rarely afford to remain static. What once worked, or worked in a particular context, may not work again. Let us return to the persistent problem of opioid overdoses discussed earlier. As progress was made in the first wave of deaths due to prescription painkillers, another challenge soon emerged in the form of heroin abuse. When progress was made combatting heroin use, a new problem arose: fentanyl. The Institute for Healthcare Improvement (Martin et al., 2016) suggested that this problem does not persist due to a lack of knowledge, evidence-based strategies, and recommendations. Rather, one of the reasons is a lack of effective implementation of these strategies and recommendations. Consistent implementation is often a challenge in any context, yet in a complex environment where people, policies, and places are never static, the challenge of implementation is multifaceted.

If we return to Alkin and Vo's (2018) definition of formative evaluation—"formative evaluation generally takes place during the early stages of program implementation"—responding to the emergent challenges arising while implementing a potential solution is formative. And if an evaluator is *continually* addressing emergent needs in the pursuit of improving a program, policy, or problem, they are engaged in formative evaluation regardless of whether the intervention ever reaches the summative state because in complex environments with persistent problems, one can rarely reach the stable environment conducive to a conclusive summative evaluation. Instead, evaluators are often engaged in continuous formative evaluation to respond to program development in complex settings (Alkin and Vo, 2018). Complex persistent problems require agility and responsiveness to emergent evidence by program designers *and* evaluators.

Patton (2011) provided an example of this necessity in the opening pages of his book *Developmental Evaluation: Applying Complexity Concepts to Enhance Innovation and Use*. He described the moment when developmental evaluation, as a specific model of evaluation, was born. As an evaluation consultant working with a community leadership program in Minnesota, his 5-year contract specified a need for 2 1/2 years of formative evaluation services, followed by 2 1/2 years of summative evaluation services. The first phase went well. Patton explained how the program made major programmatic and operational changes as part of the formative evaluation, and program staff enthusiastically sought feedback to continually make improvements. Then came the moment to transition from the formative evaluation to the summative.

As Patton detailed in his book, he began to see a shift in the group:

> We've had a couple of years changing and adapting the program. I've been impressed by your openness and commitment to use evaluation feedback to make improvements. But now, in the next phase of the evaluation, called summative evaluation, the purpose is to make an overall judgment about the merit and worth of the program. Does it work? Should it be continued, perhaps even expanded? Have you come up with a model that others might want to adopt? This means that from now on you can't make any more improvements or changes because we need the program—the model—to stay stable in order to conduct the summative evaluation. Only with a fixed intervention, carefully implemented the same way for each new group of leaders in training, can we attribute the measured outcomes to your program intervention in a valid and credible way.

Mouths fell open. Staff was aghast. They protested:

> We don't want to implement a fixed model. In fact, what we've learned is that we need to keep adapting what we do to the particular needs of new groups. Communities vary. The backgrounds of our participants vary. The economic and political context keeps changing. New technologies like the Internet are coming into rural Minnesota and creating new leadership challenges. Small communities are becoming parts of regional networks. We need to get more young people into the program. Immigrants are moving into rural Minnesota in droves, creating more

> diverse communities. We need to reach out and adapt what we do to Native Americans. No! No! No! We can't fix the model. We can't stand still for 2 years. We don't want to do the summative evaluation. (p. 2)

Patton described the rest of the conversation with the group and the eventual agreement that they would never conduct a summative evaluation on this program. Rather, it would remain in a constant *developmental* stage and they would continually report their activities, developments, and learnings to stakeholders. And there, according to Patton, he coined the term "developmental evaluation."

In addition to the program staff recognizing the need for a new evaluation approach to respond to their community's complexity, it is notable that Patton was still bounded by traditional notions of evaluation. That is, to justify a continuous formative approach to evaluation, the program must be considered to be in a perpetual state of development.

Today, formative evaluation has progressed into the broader catch-all term for program improvement. This book subscribes to that broader definition, and thus, considers developmental evaluation as a form of formative evaluation. Yet not all formative evaluation is the same. Different approaches can be further specified by their intended improvement purpose and use. For example, in the case of developmental evaluation, its intended purpose is program development.

Furthermore, broader ideas have evolved around what we consider to be evaluation. Preskill and Torres (1999) extended our notion of what can be evaluated when they advanced Evaluative Inquiry for learning in organizations. They envisioned "evaluative inquiry as an ongoing process for investigating and understanding critical organizational issues" and "an approach to learning that is fully integrated with an organization's work practices" (p. 1). Russ-Eft and Preskill (2009) further cemented these ideas with their characterization of evaluation:

> First, evaluation is viewed as a systematic process. It should not be conducted as an afterthought; rather it is a planned and purposeful activity. Second, evaluation involves collecting data regarding questions or issues about society in general and organizations and programs in particular. Third, evaluation is seen as a process for enhancing knowledge and decision making, whether the decisions are related to improving or refining a program, process, product, system, or organization, or determining whether to continue or expand a program. In each of these decisions, there is some aspect of judgment about the evaluand's merit, worth, or value. Finally, the notion of evaluation use is either implicit or explicit in each of the above definitions. (p. 4)

Their third point is especially germane here: Evaluation is seen as a process for enhancing knowledge and decision making, *whether the decisions are* *related to improving or refining a program, process, product, system, or organization.* Thus, they promote the idea that it is the entity upon which decisions are made that defines the evaluand (i.e., the entity being evaluated). This is broader than more common notions of program, process, product (or policy) and includes a system or organization. Evaluators are no longer limited to Scriven's early proclamation that the goals of evaluation were to answer questions about how well "certain entities" perform. Now, we can embrace the idea that evaluation includes enhancing knowledge and decisions about improvement.

Recently, Rohanna and Christie (in preparation) further expanded the concept of the evaluand by advancing the idea that the evaluation entity can be the social problem within a complex system. By doing so, they build on Preskill and Torres' (1999) and Russ-Eft and Preskill's (2009) ideas that evaluation is a process for investigating organizational issues and enhancing knowledge and decision making to improve a system.

Continuous Improvement and Improvement Science

By subscribing to Russ-Eft and Preskill's description of evaluation, we can embrace continuous improvement as a form of evaluation, and in particular, a type of formative evaluation. However, it is important not to fuse all improvement-oriented approaches under the broad umbrella of formative evaluation. Formative evaluation approaches should be distinguished from each other by their specified use and purpose. Like developmental evaluation, which declares program development as its intended use, continuous improvement has an intended use to *continually* evaluate and improve some entity, whether program, process, product, system, organization, or problem. Its defining characteristic—it is continuous—makes it a promising evaluation strategy for leading change in complex systems, particularly when we consider persistent problems such as the opioid crisis. Notably, continuous improvement approaches can be further specified by their purpose, as discussed later in this chapter.

But first, let us consider another persistent social problem in a complex system: completion rates in California community colleges.

Persistent Social Problems

In her 2013 article titled *Improving on the American Dream,* Gay Clyburn shared the story of Mary Lowry, a student at Foothill College who remembered crying and blaming herself because she could not pass her college math class. An otherwise successful student in high school, Mary had difficulty with math and was placed in a non-credit remedial math course in community college. She struggled. "I thought something was wrong with me," she said. "No matter how hard I tried—and I had really tried hard—I could not pass a math class"

(Clyburn, 2013, p. 15). Mary feared she would not be able to earn her degree, after failing her math class three times.

Like many community college students, Mary was at risk for dropping out of college and not realizing her dreams. In 2013, fewer than half (48.5%) of California community college students completed a degree, certificate, or transferred to a 4-year college within 6 years.[2] Students like Mary who entered community college but deemed unprepared were placed in non-credit remedial courses, also called *developmental courses*. For those students, the completion rate was even lower at 41.1%.

This was a societal problem without an easy fix. California had, and still has, the largest community college system in the United States, serving approximately 2.1 million students across 116 colleges (California Community Colleges Chancellor's Office, 2020a). Most of the students (80%) were enrolled in at least one developmental course during their college experience (Mejia, Rodriguez, & Johnson, 2016). Furthermore, Latinx and African American students were disproportionately affected with higher enrollment in developmental courses, and lower than overall completion rates (CCC Student Success Scorecard; Mejia et al., 2016).

Community colleges are promoted as an affordable, accessible, and equitable path for students to achieve a higher education degree or vocational certificate. More than half of all undergraduate Latinx and African American students attend community college, and many are low-income and nontraditional students (Mejia et al., 2016). Developmental courses were designed to help students who were identified as underprepared for their pursuit of higher learning or a vocational career. The developmental sequence was supposed to provide foundational and basic skills in math or English, and thus help them complete their college-level courses.

The espoused vision of the community college system was "making sure students from all backgrounds succeed in reaching their goals" (California Community Colleges Chancellor's Office, 2020b). In actuality, the system was creating a roadblock by requiring a lengthy developmental sequence and adding multiple semesters of additional coursework for no college credit. The result: Fewer than half of these students actually completed community college. The system was having the opposite effect, and students like Mary were paying the price.

Why? Because requiring developmental courses was an attempt at a straightforward solution to a complex systemic problem. Students who were often identified as underprepared tended to be low-income or nontraditional. They were required to take more classes. More classes meant more tuition costs and greater financial burden. More classes also meant more chances of failure. More failure meant students like Mary blamed themselves rather than the system. Students were failing, giving up, and dropping out.

[2] Source: California Community Colleges Student Success Scorecard. Percentage of 2007–2008 cohort who were enrolled the first time and tracked for six years. 5-yr. Trends. https://scorecard.cccco.edu/scorecardrates.aspx?CollegeID=000#home

The problem was difficult to improve. For years, community colleges had been concerned about their low completion rates. State policymakers flowed funding into a multitude of initiatives, including $20 million annually since the 2007–2008 school year for its Basic Skills Initiative (Mejia et al., 2016). Still, the problem persisted.

The Carnegie Foundation for the Advancement of Teaching took a new approach. In 2010, they convened a network of researchers, practitioners, and community college faculty to jointly unpack, frame, and really understand the problem before they set out to solve it (Clyburn, 2013). They developed a new pathway for students placed into developmental math. Students could take a quantitative reasoning (Quantway) or statistics (Statway) course that was both developmental and college-level, allowing them to earn college credit. Additionally, the courses connected math concepts to real-world problems.

But the network did not stop there. Their team of scholarly and practical experts committed to continually studying, understanding, and improving this problem. They developed and tested activities designed to help students persist through the courses and increase their sense of belonging. In their first year, 51% of the 1,077 students enrolled in the Statway course completed the sequence in one semester, compared with 21% of their campus peers who took the traditional path to completing the sequence, which took a year (Silva & White, 2013). Mary was one of those successful Statway students (Clyburn, 2013).

With their success, the Carnegie Foundation continued to roll out the pathway options to more community colleges. Their network committed to continuing their inquiry as they scaled up the pathways. They learned from both what was working well and what could still be improved. By 2016–2017, approximately 7,500 students over 48 institutions were enrolled in either Statway or Quantway, with most of them successfully completing (62% and 72%, respectively) (Huang, 2018).

Although these innovative pathways were working, not all community college students had the opportunity to enroll. In 2017, California policymakers passed Assembly Bill 705, which required colleges to give students alternate options to remedial courses, allow them to enroll in transfer-level courses, and use high school records instead of less predictive placement tests. Early evidence suggests more students are enrolling in non-remedial courses and are succeeding (Mejia, Rodriguez, & Johnson, 2019). Thus, these changes have the potential to disrupt the old system and will require ongoing disciplined inquiry by policymakers, practitioners, researchers, and evaluators to learn from, and respond to, emergent challenges.

The community college example not only serves as an illustration of a persistent problem in a complex system, but it also shows the power of embracing continuous improvement approaches and multiple perspectives for understanding and improving a problem. The Carnegie Foundation for the Advancement of Teaching grounded their network in improvement science.

Improvement Science

The American Society for Quality (ASQ) defines continuous improvement as "the ongoing improvement of products, services or processes through incremental and breakthrough improvements."[3] There are many different models of continuous improvement, including improvement science (i.e., the "science of improvement") (Berwick, 2008, p. 1182). But what exactly does the science of improvement mean? Improvement science is a broad approach with various definitions and models. When it is applied *continuously* as part of an effort to integrate its tools and methods into an organization's everyday work, it falls under the umbrella of continuous improvement; yet, importantly, the two terms are not interchangeable.

Improvement science has been defined as the following:

> A field of study focused on the methods, theories, and approaches that facilitate or hinder efforts to improve quality in context-specific work processes, and centers inquiry on the day-to-day 'problems of practice that have genuine consequences for people's lives.' (Bryk, 2009, cited by Park, Hironaka, Carver, & Nordstrum, 2013, p. 598; Health Foundation, 2011).
>
> 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 (Lemire, Christie, & Inkelas, 2017, p. 25).
>
> Improvement science is a methodological framework that is undergirded by foundational principals that guide scholar-practitioners to define problems, understand how the system produces the problems, identify changes to rectify the problems, test the efficacy of those changes, and spread the changes (if the change is indeed an improvement) (Hinnant-Crawford, 2020, p. 29).

Improvement science is scientific: It is disciplined and systematic and grounded in a theoretical and methodological approach. It is systemic: It not only seeks to improve a problem but also the system in which the problem is situated. There is substantial overlap between continuous improvement and improvement science. In fact there is so much overlap that differences between the two may seem inconsequential. However, it is important to understand the distinguishing features of each when leading change in complex systems. Not all continuous improvement approaches are scientific or systemic, and not all improvement science is continuous. In my own experiences, I have heard organizations claim they are committed to continuous improvement and implement regular surveys to receive and respond to feedback. There is nothing wrong with this. It is a beneficial, systematic practice for continually improving and meeting stakeholder needs, but it does not address a specific problem situated within a broader system, nor it is grounded in theoretical principles for

[3] https://asq.org/quality-resources/quality-glossary/c

improvement. On the flipside, improvement science, while iterative and continuous in its approach to improving a specific problem, does not necessarily instill a continuous improvement culture throughout an organization. There may be one *project* team focused on solving one problem. Once the problem is vastly improved, the team may disband, and along with it, the continuous and collaborative mode for ongoing inquiry. Again, there is nothing wrong with this approach. But the sweet spot is found in the overlap, where inquiry is disciplined and ongoing, embraces a systemic view, and is grounded in theoretical and methodological principles of improvement. Therefore, this book advocates a continuous improvement model that is grounded in improvement science.

Improvement science was founded on much of the work of W. Edward Deming (Langley et al., 2009). Deming was an engineer and statistician who advanced production, management effectiveness, and quality improvement. His ideas shaped Japanese manufacturing and industrial practices after World War II (Walton, 1986).

Improvement science delineates two types of knowledge: subject knowledge and profound knowledge. Subject knowledge is considered the content knowledge within a particular area, often held by practitioners and/or researchers, while profound knowledge is the more systematic awareness of "how to make changes that will result in improvement in a variety of settings" (Langley et al., 2009, p. 75). Deming defined profound knowledge "as the interplay of theories of systems, variation, knowledge, and psychology" (Deming, as cited in Langley et al., 2009, p. 75).

BOX 1.1

Deming structured his system of profound knowledge around four types of knowledge (Christie, Inkelas, & Lemire, 2017):

1. **Knowledge of systems**: This type of knowledge refers to the interdependence of departments, people, and processes within an organization (Langley et al., 2009). Integration of these individual parts toward a common aim contributes to a successful organization (Deming, 1994; Langley et al., 2009).
2. **Knowledge of variation**: This component not only promotes the shift from analyzing averages to a deeper study of variation in data, but it also encourages an understanding of different types of variation and their implications for system performance (Christie, Inkelas, & Lemire, 2017; Langley et al., 2009).
3. **Knowledge of how knowledge grows**: This type of knowledge refers to learning by making predictions about potential changes, then actually making the changes and measuring the results (Langley et al., 2009).
4. **Knowledge of psychology**: This component reflects the human side of change and encompasses how attention to people's values, attitudes, and motivations can influence change (Langley et al., 2009).

Deming is also credited with the Plan, Do, Study, Act (PDSA) cycle, which was an evolution of Walter A. Shewhart's initial cycle of scientific testing (Moen & Norman, 2010). The PDSA cycle is formatted for rapidly experimenting with new practices and generating new knowledge (Langley et al., 2009). Its four stages—plan, do, study, and act—follow a dynamic, deductive, and inductive learning process. Experiment logistics are planned during the first stage (Plan), implemented during the second stage (Do). During the third stage (Study), the experimenter analyzes relevant data, reflects on the process, and determines the next steps. In the final stage (Act) next steps are put into action. Ideally, the PDSA cycle should occur within a short timeframe, on a small scale so ideas can be quickly tested, and either adapted and retested, gradually scaled up, or potentially abandoned as necessary.

From Plan to Do, the deductive approach is applied. A theory is tested with the aid of a prediction. In the Do phase, observations are made and departures from the prediction are noted. From Do to Study, the inductive learning process takes over. Gaps in the prediction are studied and the theory is updated accordingly. Action is then taken on the new learning in the last stage (Langley et al., 2009, p. 82)

Langley and his colleagues (2009) at Associates in Process Improvement expanded on Deming's work and developed the "Model for Improvement." The Model for Improvement encompasses three questions and the PDSA Cycle (Langley et al., 2009). The three questions are:

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

While there are different models of improvement science, the process typically encompasses three broad phases illustrated by Figure 1.2. These three stages loosely correspond with the Model for Improvement's three questions.

FIGURE 1.1 ● PDSA Cycle

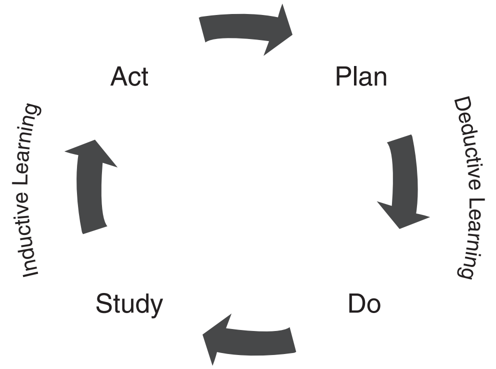
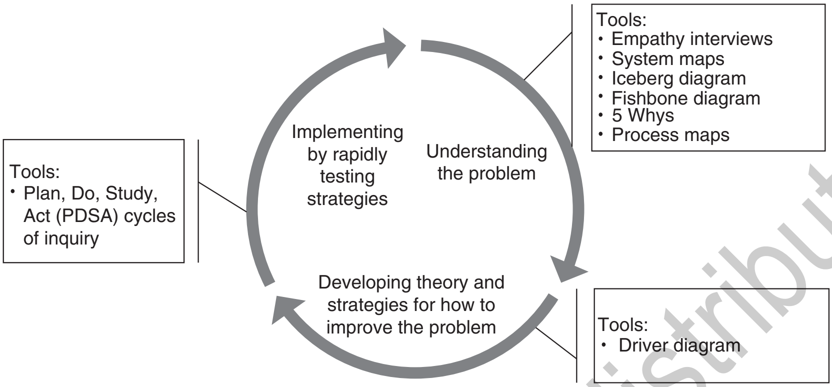


FIGURE 1.2 ● Three Broad Improvement Phases


Tools:
• Empathy interviews
• System maps
• Iceberg diagram
• Fishbone diagram
• 5 Whys
• Process maps
Tools:
• Plan, Do, Study, Act (PDSA) cycles of inquiry
Implementing by rapidly testing strategies
Understanding the problem
Developing theory and strategies for how to improve the problem
Tools:
• Driver diagram

Understanding the Problem

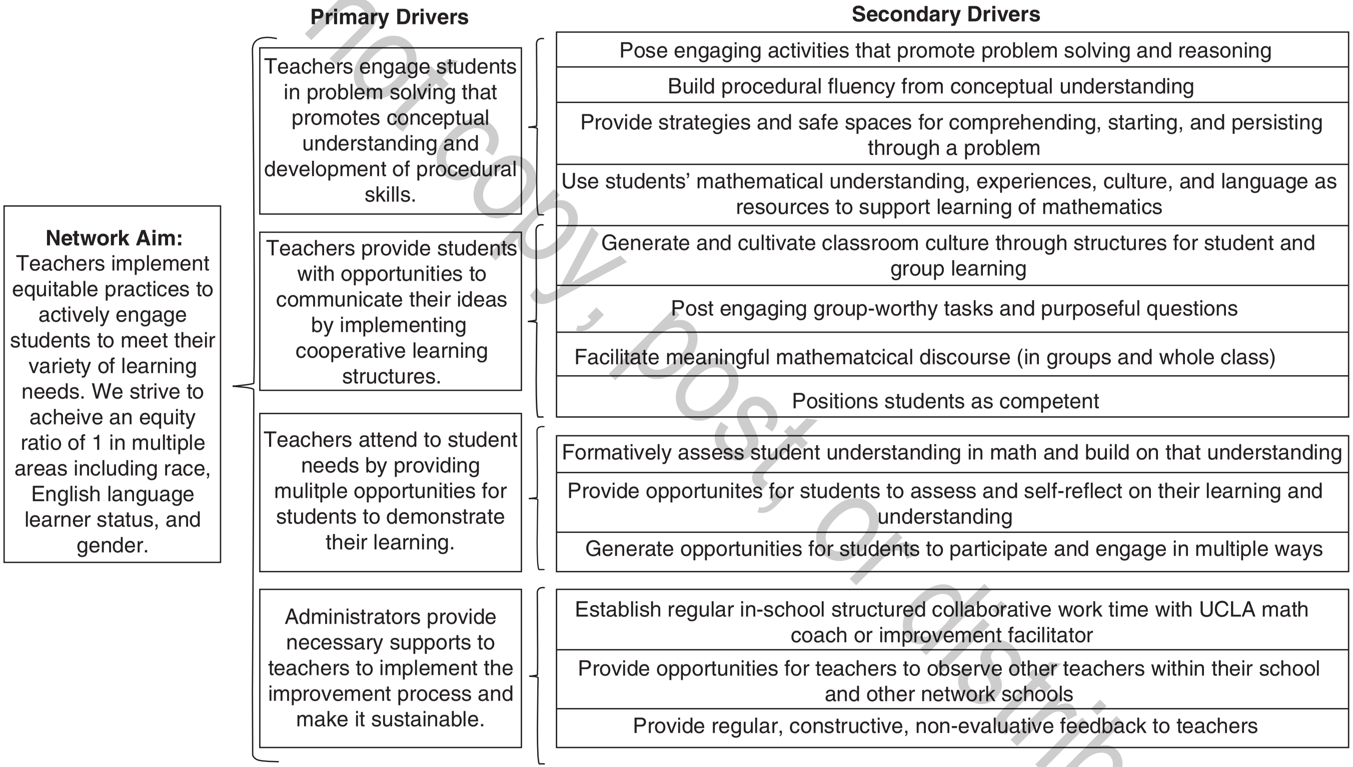
Before attempting to improve or solve a problem, evaluators first need to identify and understand the cause (or causes) of the problem. We live in an action- and solution-oriented world. It is easy to impatiently jump right to the solution without taking the time to fully understand why there is problem. Others have referred to this phenomenon as "solutionitus" (Bryk et al., 2015). Root cause and causal systems analysis tools such as the 5 Whys protocol and the Cause and Effect diagram, also known as the *fishbone* or *Ishikawa diagram*, help teams dig deep or go wide across the system, thus preventing solutionitus. There are also other improvement tools that are often used in the phase, including empathy interviews, system maps, iceberg diagram, and process maps.

Developing a Theory for Improving the Problem

Evaluators are very familiar with the concept of a theory of change or a theory of action. Improvement science has an analogous tool, known as a *Driver Diagram* or *Theory of Improvement*. The tool, shown in Figure 1.3, has a similar purpose as a logic model but with more of system-wide focus. The system can be bounded in different ways to have a narrower or broader focus. In Figure 1.3, the system is bounded more narrowly to the classroom.[4] The aim, shown in the box on the far left, is the measurable goal. The Primary Drivers identify the high-leverage focal areas in the system that would drive the change in the aim. High-leverage refers to the idea of focusing on those areas within the system that will provide the "most bang for your buck." Using the information gained in the Understanding the Problem phase, a team considers what causes they can address with the fewest resources (or currently available resources) and will have the biggest impacts.

[4] The case study, presented in Part 2 of this book, provides details about the decision to narrowly focus this driver diagram.

FIGURE 1.3 ● Driver Diagram Example



Network Aim: Teachers implement equitable practices to actively engage students to meet their variety of learning needs. We strive to acheive an equity ratio of 1 in multiple areas including race, English language learner status, and gender.

Primary Drivers

Secondary Drivers

Teachers engage students in problem solving that promotes conceptual understanding and development of procedural skills.

- Pose engaging activities that promote problem solving and reasoning
- Build procedural fluency from conceptual understanding
- Provide strategies and safe spaces for comprehending, starting, and persisting through a problem
- Use students' mathematical understanding, experiences, culture, and language as resources to support learning of mathematics

Teachers provide students with opportunities to communicate their ideas by implementing cooperative learning structures.

- Generate and cultivate classroom culture through structures for student and group learning
- Post engaging group-worthy tasks and purposeful questions
- Facilitate meaningful mathematcical discourse (in groups and whole class)
- Positions students as competent

Teachers attend to student needs by providing mulitple opportunities for students to demonstrate their learning.

- Formatively assess student understanding in math and build on that understanding
- Provide opportunites for students to assess and self-reflect on their learning and understanding
- Generate opportunities for students to participate and engage in multiple ways

Administrators provide necessary supports to teachers to implement the improvement process and make it sustainable.

- Establish regular in-school structured collaborative work time with UCLA math coach or improvement facilitator
- Provide opportunities for teachers to observe other teachers within their school and other network schools
- Provide regular, constructive, non-evaluative feedback to teachers

The Secondary Drivers unpack the broad Primary Drivers, thereby leading to more manageable and actionable steps. Change ideas—those strategies for improvement to be tested through the PDSA cycles—"plug in" to these secondary drivers; they should be developed to improve the secondary drivers. Notably, these are all hypothesized theories. Driver diagrams are dynamic and should always be updated with the most recent emergent learnings about a problem.

Implementing by Rapidly Testing Strategies and Generating New Knowledge

The PDSA is a format for rapidly testing potential strategies (change ideas) aligned with the drivers. Importantly, it is a vehicle for reflective practice. The prediction surfaces assumptions, what one expects to happen while implementing a change idea. Evidence is collected, and then the individual and/or team reflect on whether they met the assumption, and why or why not. With this new knowledge, the team can choose to scale up a change idea, adapt it and conduct another PDSA cycle, or abandon it all together. Ultimately, this knowledge provides more understanding around the problem, potentially altering the theory for how to improve it.

Although Deming's ideas were grounded in industry and manufacturing, they have expanded to other fields. Currently, improvement science is most prevalent in healthcare. Don Berwick, the founder of the Institute for Healthcare Improvement (IHI), was one of the early champions of improvement science. Improvement science has successfully cracked vexing healthcare challenges such as how to reduce the number of child asthma-related visits to the emergency room (C. E. Williams, 2015).

Improvement science has spread to education in recent years, as demonstrated by the aforementioned community college example. The Carnegie Foundation combined the improvement science model with the concept of networked improvement communities when they formed a network of practitioners and experts who collaboratively experimented with innovative common practices and consistent measures across the community colleges. K–12 schools are also embracing improvement science, as demonstrated by this book's case study.

Other Continuous Improvement Approaches

It should be noted that there are other improvement models. Below are brief descriptions. While not an exhaustive list, it does represent the more commonly known models in manufacturing, health care, and education.

Six Sigma: Six Sigma is another improvement science model focused on organizational quality improvement. The concepts behind it originated in the 1970s when an engineer at Motorola developed a new quality assurance methodology aimed at reducing process variation. The model incorporated

Walter Shewart's ideas around variation: Variation that fell outside of three standard deviations (sigma) from the mean, up or down, required a correction (Six Sigma Global Institute [SSGI], 2020). The goal was to improve consistency of processes, and thus, performance. Six Sigma practices can now be found in numerous and varying sectors, including manufacturing, retail corporations, government, and health care (American Society for Quality [ASQ], 2020a).

There are more specified models within the Six Sigma approach. The define, measure, analyze, improve, and control (DMAIC) is one of those. DMAIC is a data-driven quality improvement framework with the five phases representing the different stages of the process (ASQ, 2020a).

- Define the problem, improvement activity, project goals.
- Measure the process performance by developing process maps, capability analysis, and pareto charts.
- Analyze the process to determine root causes of problem or poor performance.
- Improve the process performance by addressing the root cause. May use approach where you rapidly introduce change, conduct controlled test, and apply statistical design of experiments (DOE) method.
- Control the improved process and future process performance by documenting and monitoring process behavior. The goal is to ensure process consistency within three sigma, up or down.

Lean: Lean is a set of management practices and techniques aimed at eliminating non-value-added activities, thereby, improving efficiency and effectiveness (ASQ, 2020b). Lean began in the automobile industry, dating back to Henry Ford's mass production intention of increasing efficiency (SSGI, 2020). Toyota is a well-known champion of the method. They have developed systems to reduce waste in processes and procedures so that all work adds value. Lean practices are also used in various sectors, including manufacturing, finance, and health care.

Lean Six Sigma: This approach combines ideas from Six Sigma and Lean to improve performance by improving consistency of process performance and increasing efficiency and reducing waste. Lean Six Sigma also incorporates the DMAIC model, with a broader focus on the problem rather than just on the process (SSGI, 2019) and can be found in use in numerous sectors, including government, health care, industry.

Data Wise: The Data Wise project developed out of the Harvard Graduate School of Education to support educators engaged in collaborative inquiry.

The aim of this model is to integrate data-driven inquiry into instruction and district improvement. Their Universal Data Wise Improvement Process includes the following steps (Lockwood et al., 2017):

- Organize for collaborative work by establishing structures and teams
- Build data literacy to increase comfort with data
- Create data overview to identify a focal question to address
- Dig into data to address the focal question and identify a learner-centered problem
- Examine own practice and identify a focal problem of practice
- Develop action plan
- Plan to assess progress
- Act and assess by implementing plan and gathering, reviewing, and reflecting upon evidence of improvement

Lesson Study: Lesson study is a cycle of inquiry process that originated in Japan as a collaboration among teachers to improve instruction. As such, it is a collaborative inquiry process whereby teachers plan a lesson with fellow teachers, collect data regarding the lesson, review and reflect on those data, and revise the lesson as needed. There are four broad steps (The Lesson Study Group at Mills College, 2018; Teacher Development Trust, 2015):

- Study: Teams of teachers collaborate to investigate a potential problem, review relevant research, and identify a goal for students.
- Plan: The teachers plan a lesson together that addresses the issue raised in the Study phase. They predict how students will react to the lesson and determine what student data to collect.
- Teach: One team member teaches the lesson, while the other teachers collect student data around student thinking and learning.
- Reflect: The team meets after the lesson to discuss the data, reflect on what they learned, and consider whether it met their prediction. They then decide how to revise the lesson.

Ultimately, all these approaches and models can be considered formative evaluation because they focus on improvement, be it a process, a problem, an organization, or a system, with the intention of enhancing knowledge and decision making (Russ-Eft & Preskill, 2009). Yet continuous improvement is a distinct form of formative evaluation because of its *continuous* nature. This feature lends itself to being a powerful vehicle for change. The evaluation team can uncover and address emergent needs by ongoing disciplined inquiry. Improving persistent problems requires agility and responsiveness. As the soup example has shown us, contextual conditions rarely remain static.

Conclusion

Improvement approaches fall under the broad umbrella of formative evaluation. The decision of which model to use will depend on your purpose, problem, and type of system in which you are situated. When attempting to improve a problem, diagnosing whether it exists in an organized system or complex system is a key first step (Rohanna & Christie, in preparation). Chapter 2 prepares the reader for this task by discussing the idea of systems and complexity in more depth, while Part 2 begins the case study of applying an improvement science framework within a complex system. By better understanding which approach to utilize, evaluators are better positioned in their quest for leading change.

Questions for Discussion

1. Why does the author consider improvement science to be an evaluative strategy for change?
2. What persistent problems have you encountered in your own settings or organizations?
3. How might improvement science or another continuous improvement model be applied to that problem?