Christie, C. A., Lemire, S., & Inkelas, M. (2017). Understanding the similarities and distinctions between improvement science and evaluation. In C. A. Christie, M. Inkelas & S. Lemire (Eds.), *Improvement Science in Evaluation: Methods and Uses. New Directions for Evaluation*, *153*, 11–21.

1

Understanding the Similarities and Distinctions Between Improvement Science and Evaluation

Christina A. Christie, Sebastian Lemire, Moira Inkelas

Abstract

In this chapter, we discuss the similarities and points of departure between improvement science and evaluation, according to use, valuing, and methods—three dimensions of evaluation theory to which all theorists attend (Christie & Alkin, 2012). Using these three dimensions as a framework for discussion, we show some of the ways in which improvement science and evaluation are similar and how they are different in terms of purposes, goals, and processes. By doing so we frame the illustrative cases of improvement science that follow in this issue.

Improvement science is an approach to increasing knowledge that leads to an improvement of a product, process, or system (Moen, Nolan, & Provost, 2012). Evaluation is a systematic process designed to yield information about the merit, worth, or value of "something" and, for the context of this journal issue, that something is assumed to be a program or policy. These two approaches have much in common, but little has been written about the ways in which they are similar, different, or how they can be used cooperatively. In what follows, we consider similarities and distinctions across three dimensions of evaluation theory: use, valuing, and

methods. Before advancing this comparison, however, an important distinction about the theoretical foundations of improvement science and evaluation is called for.

Theoretical Foundations of Evaluation and Improvement Science

Evaluation has a theoretical literature that is composed of an array of models for how best to practice evaluation. Simply consider the prescriptive models for theory-based and realist evaluation, utilization-focused and empowerment evaluation, and, perhaps more recently, the surge of interest in developmental evaluation. These theories have an ultimate goal, which articulates what is assumed to be the primary purpose of evaluation from the perspective of that theoretical approach. Common for all of these evaluation theories is their prescription of distinctive elements that help to delimit the approach and distinguish it from other approaches. For example, there are theoretical approaches that define evaluation as the science of valuing (Scriven, 2003) whereas others see the ultimate goal to be program improvement (Patton, 2008) or empowerment (Fetterman, 1994).

For a theory to be translated or applied in practice, it must be delineated in terms of the processes and procedures that lead to the actualization of the theory's goal. Not all prescriptions for practice are theories, however. A theory must have an identifiable "character"—specific goals and purposes that are articulated and complemented by a set of procedures for achieving these. Not all procedures or methods are unique to a particular theoretical approach. And although these approaches have different goals for evaluation, they do share common procedures, such as, for example, the use of qualitative data collection strategies. Unfortunately, and as observed by Miller and Campbell (2006), some of these evaluation theories prescribe what evaluators should do but rarely how to do it.

Improvement science differs from evaluation in this regard. Improvement science emerged from the study of production in the early 20th century, Deming's profound knowledge, emphasizing systems theory, analytical studies, the study of variation, and human psychology in production systems (Langley et al., 2009). Although differences about the precise definition of improvement science exist, there is general agreement on the ultimate purpose of improvement science, that being continuous improvement through systematic study. Improvement is accomplished through design or redesign of products and/or processes and improvement of the system that produces the products and processes (Moen et al., 2012). Painting in broad strokes, this unifying goal is arguably more focused than the many different goals prescribed by evaluation theory. And, perhaps as a result, improvement science is typically more operational, offering not just a unifying goal but also, and perhaps more important, guidance on how to achieve that goal.

There are, of course, different models for how to do improvement science, different procedures for practice. These models, however, are not grounded or motivated by epistemological or ideological positions. Rather, they are defined by a different set of procedures and techniques that are informed by a shared postpositivist epistemology. In general, it is this pragmatic approach to knowledge production that results in a broad array of concrete techniques and tools to be potentially applied in evaluative contexts. Before illustrating the applications of improvement science, we situate improvement science in the context of use, valuing, and methods dimensions of evaluation.

The Use Dimensions of Evaluation and Improvement Science

Use refers to the extent to which the practices and procedures of the evaluation are designed to promote the goal of using an evaluation's process or findings (Weiss, 1998). Use is also a dimension of practice where evaluation and improvement science share common ground. Whereas use is a central aim of evaluation, as evidenced by the rich literature on this topic, it has always been defined in relation to (and as a possible outcome of) different evaluation activities. Use defined as improvement is what defines improvement science. The development of improvement science techniques, methods, and concepts emerges from the idea of process, product, or systems change. Improvement is the driver of improvement science.

There are several key ideas and processes that are associated with evaluation models that have as the primary goal producing actionable evidence for improvement of programs and decision making that are shared with improvement processes. This is not surprising given that the goal of improvement science is to generate information that leads to testing a change. Thus, the primary goals and outcomes of use theories are well aligned with the goals of improvement science, as are many of the procedures associated with use approaches, such as stakeholder engagement and participation.

Improvement science is often highly collaborative with a specific role for people closest to the process, similar to participatory evaluation approaches that have grown out of evaluation use approaches, such as practical participatory evaluation (Cousins & Earl, 1992). Key here is that both provide a process for connecting practitioners to an evaluative process through engagement. Buy-in is a necessary condition in such participatory processes, which often serves to reduce fears and anxiety that are related to the conduct of evaluative processes (Donaldson, Gooler, & Scriven, 2002) and thus increases the likelihood that learning will occur as a result of the process. Accordingly, intimate knowledge of processes that are the target of improvement is fundamental to improvement science. In fact, in improvement science leaving the process owners out of the change process is looked down upon and associated with failure to improve (Kaplan, Provost, Froehle, & Margolis, 2012).

Another necessary condition for success in a participatory evaluation process is the capacity of stakeholders to think "evaluatively" (Vo, 2013). Evaluative thinking is a type of critical thinking that is specific to evaluation, where systematic evidence is used to construct arguments and value judgments that are contextually relevant (Vo, 2013). Although this particular type of cognitive process is not discussed in relationship to improvement science, it is easy to see how evaluative thinking is also necessary for good improvement processes to occur.

In contrast, improvement science emphasizes systems thinking and knowledge of the psychological processes that generate change (Deming, 2000). Whereas interest in systems thinking is growing among evaluators, it has yet to become a backbone of our practice. The concepts, tools, and techniques of systems thinking are arguably still far—both in conceptual and operational terms—from most evaluation practice. As such, it is probably more talked about than practiced. Similarly, and perhaps somewhat more surprising, the attention to the psychological processes of change have been less salient in evaluation. This is surprising given the intellectual roots of evaluation in social psychology. Yet, and in spite of the obvious relevance, the role of and attention awarded behavioral change theories, as just one example, are limited in evaluation, both in theory and practice. Central to the improvement scientists' practice is the knowledge of psychology of change. The explicit use of psychology as well as the learning method itself (i.e., adaptation of the scientific method for action-oriented learning) is core to improvement science as a change management method. The resultant array of change concepts speaks to this point and should be of wide interest to evaluators.

Improvement science also differs from evaluation use approaches in that improvement does not focus on social accountability. Rather, because the focus is local, emphasizing the implementation of small, rapid cycle tests of changes, data are often collected for use by service delivery providers (i.e., a physician), so that outcomes can be improved. Improvement science is intended to be practiced by "process owners" (people in systems), why the focus is usually quite concrete (leading to specific action), as opposed to being broad (i.e., we need a culture change).

While presenting on improvement science at a meeting of the American Evaluation Association, we fielded questions about how improvement science and developmental evaluation may or may not be similar, and we hope to clarify this for our readers. Michael Patton, a well-known evaluation use researcher and theorist, has recently written on and popularized an evaluation approach referred to as developmental evaluation (2010). Developmental evaluation "tracks and attempts to make sense of what emerges under conditions of complexity, documenting and interpreting the dynamics, interactions, and interdependencies that occur as innovations unfold" (Patton, 2010, p. 7). Formative evaluation focuses on improvements, like improvement science, rather than on developments, which is the focus of

developmental evaluation. In the first chapter of his book, Patton describes how in developmental evaluation, the evaluator brings evaluative thinking and data to the development of program approaches for different groups of recipients and policies. Accepting this definition, although improvement science and developmental evaluation share the focus on quick data collection cycles that take into account the system dynamics, Patton would argue that these processes are different in their purpose and goal. Moreover, developmental evaluation does not begin with the defined theory, rather outcomes emerge through the evaluative process. Improvement science, however, usually begins with identifying clear, specific, and measurable outcomes, exactly what Patton describes should not take place in a developmental evaluation.

The Valuing Dimensions of Evaluation and Improvement Science

Valuing describes approaches that emphasize the importance of how, by whom, and in what way value judgments about programs are determined in evaluation. Many evaluators view their work in terms of the ways in which it may contribute to social good. Paraphrasing from Ernest House (1980), evaluation determines who gets what. As such, social justice oriented evaluators, such as Ernest House (1980), Jennifer Greene (2016), and Donna Mertens (1999), argue that the evaluator should take a position on evaluation as a process for improving social conditions and disrupting systematic power imbalances. In this way, evaluation deliberately addresses the social, economic, and political systems in which programs and policies are developed and implemented, sometimes with the aim of challenging the status quo.

In improvement science, the good that comes of the process is the improvement in quality, which should lead to better outcomes and services for program beneficiaries. This point, however, is not taken up ideologically. Instead, quality improvement focuses on stakeholders identifying the needed areas of improvement in a system, toward identifying and testing relevant solution(s). This kind of systems thinking is aligned with the work of Meadows (2008) where a system is an "interconnected set of elements that is coherently organized in a way that achieves something" (p. 11). From this perspective, improvement science does not by definition focus on systems issues in an effort to improve the social, economic, or political system conditions (though these conditions may be influenced).

Related to social justice, evaluation scholars in recent years have focused on the role of culture in evaluation and the various ways in which culture, race, and class shape the evaluation process. Theoretical writings have addressed issues of cultural diversity and why and how evaluators need to reflect on and engage in thoughtful practices that respect the culture of the evaluation context. Culture refers not only to race, ethnicity, social class,

language, sexual orientation, age, and gender but also to organizational culture and institutions such as government, education, family, and religion (American Evaluation Association, 2011).

Issues of cultural competence are also addressed in quality improvement and the improvement science literature. When searching the literature, there are improvement science studies that focus on how services can be improved toward better alignment with the values of a particular group, such as HIV-positive men who have sex with men, or African American women with Type II diabetes. The literature does not, however, address issues related to the ways in which culture affects the process of improvement science itself, similar to what has emerged in the evaluation literature. Whereas the evaluation literature addresses the importance of evaluation taking into account the culture of the actors and contexts in which an evaluation takes place (as is also the case with improvement science), a growing body of literature also concerns the ways in which cultural competence in evaluation "is a stance taken toward culture" that "emerges from an ethical commitment to fairness and equity for stakeholders" (American Evaluation Association, 2011).

The sustained interest in valuing among evaluators should be viewed in the broader context of evaluation as a driver for social betterment. In this spirit, and especially in the era of accountability, program performance takes center stage. One unfortunate consequence emerging from this heavy focus on outcomes is the limited room for programs to fail, especially when aiming for large-scale, longer-term changes. Fear of failure pervades evaluation. In contradistinction, learning from error is an important aspect of improvement science. Improvement science is a process of testing change. In the spirit of this goal, failure is expected and accepted. By removing the high stakes associated with studying program impact and instead studying small iterative changes, mistakes are less consequential and change more manageable. A space for error has been carved out.

The Methods Dimensions of Evaluation and Improvement Science

Methods refer to those approaches that have as the primary goal methodological rigor and knowledge generation. Early evaluation practice was grounded in a positivist search for effective solutions to social problems (Shadish, Cook, & Leviton, 1991). From this perspective, stringent application of research methodology (e.g., Campbell's "Experimenting Society") was used to produce evidence of a program's success. Successful programs would then be replicated and transferred to other problems or contexts and those not proven successful would be terminated (Cronbach et al., 1980). These evaluation experiments often proved difficult to sustain and rarely provided contextually valid data (e.g., Cronbach et al., 1980; Patton, 2008; Shadish et al., 1991). And even when positive results were not

obtained, programs often continued (Patton, 2008; Shadish et al., 1991). As a result, evaluators have increasingly turned their attention toward outcome patterns and variations across different implementation settings, times, and contexts.

Identifying and understating variation in outcomes is also a very important part of the improvement science methodological toolbox. Part of what explains variation in program outcomes is the complexity of the contexts in which programs are delivered. Deming (2000) argued that a key to solving most quality problems is recognizing that often what we are attempting to change are complex systems. A common mistake is to assume that it is one component of a system, or one variable, that is causing the problem. If this were the case, innovations could always be tested using more traditional evaluation designs such as randomized controlled trials (RCT). But in fields such as health care or education, the effects of single variables are most often dwarfed by the complexity of the system in which they are embedded (Berwick, 2008). So, it is necessary to understand the component processes that make up the system and how they work together in order to understand the roots of the problem and generate innovative solutions. Sometimes quality can be improved by merely tweaking the system, that is, making small changes that enable the system to function in context the way it was designed to function. But other times the system must be redesigned from the ground up or major components changed. This framework consists of the following components: clear shared goals; sensitive measures to chart progress; deep understanding of the problems/barriers that impede success; sources of innovations, grounded in explicit theories of the problem; and mechanisms for comparing/researching innovations and systematically testing whether proposed changes are actually improvements (see Juran & DeFeo, 2010; Langley et al., 2009; Rother, 2009). These five components bring focus to the improvement process and highlight the kind of evaluative process necessary to establish that a hypothesized change is, in fact, an improvement.

Another crucial characteristic of the methodological approach taken by improvement scientists has to do with the intended outcome of their work. Whereas traditional evaluators have focused on demonstrating significant improvements in average outcomes compared with the status quo, the focus in improvement science is equally on reducing variability in outcomes. A true improvement is one that can be counted on to work for most everyone, not just for a few participants and not only with a subset of circumstances. Reducing variability requires that innovations be grounded in explicit theories of change, for it is critical to know not only that an innovation works but also why it works in a particular context. Having theories that explain variability across contexts is also critical for future scaling of an intervention to function in a wide array of contexts.

There are many evaluation models that Christie and Alkin (2012) classify as methods approaches, which focus on understanding how programs

work on average with a given sample, in a particular context. Thus, whereas improvement science shares epistemological "land" with methods-focused evaluation approaches, methods approaches such as the RCT and some quasi-experimental approaches do not have improvement as the primary goal; rather, these studies are taken up to study impact, with less attention on understanding variation and more on how the program works on average when implemented with fidelity. In contrast, improvement science places a heavy emphasis on learning from variation and learning about the specific conditions under which processes fail to work.

In evaluation, many of these designs, but in particular the RCT, assume that the program is already performing at its best, and the purpose of the study is to determine whether there is a causal link between the program and its intended outcomes. Improvement science uses many of the practices and procedures used by methods approach evaluation theorists. However, these approaches are used in different ways and for different purposes. For example, randomization (the scientific method) can be part of small-scale tests in improvement to address the question, "What happens when one group of people do this and another group does not? How does the outcome differ?" The distinction is that these RCTs are not full-scale field trials. Rather, randomization is used throughout the small-scale testing and implementation phases of improvement science in an attempt to isolate the change in an outcome if a program process is tweaked or changed in a particular way. Also, although the logic of randomization works the same way in improvement science and evaluation, an important distinction is that improvement scientists do not view randomized experiments as the "gold standard" for producing evidence because context is so essential to improvement; the goal is prediction not estimation so any methods that do not open the black box are suboptimal.

For improvement science, the scientific method is used for action-oriented learning. Thus, use of randomization is an admission of a lack of knowledge about something, and improvement is in pursuit of that knowledge, otherwise the ability to predict the impact of a change under varying conditions is compromised. In evaluation, random assignment is usually applied in the context of large-scale studies and the program beneficiaries are most often the unit at which random assignment is most desired as it is the purpose of these studies to address the question of overall program impact on a given set of outcomes. It is typically a design feature of large-scale studies, as opposed to a useful logic potentially applied in small-scale, iterative learning cycles.

Another key feature of improvement science that is shared with methods focused evaluation approaches is the use of program theories, both theories of change and theories of action. Theories about program processes and their connection to program outcomes are critical in improvement science as they frame the ideas for change in relationship between the aims of the improvement process. These theories are often depicted

in driver diagrams. Driver diagrams articulate measurable changes in outcomes, how the processes that are involved in the change process will advance measurable outcomes. They are very similar to logic models as they often incorporate elements of both the theory of change and theory of action. The processes articulated in the driver diagram are ideally evidence based. Like theories of change and action on evaluation, they also often incorporate the expert knowledge of those working within the context or system.

The theories in improvement science are then used to guide small tests of the changes in processes and procedures over a short period of time that will result in improvement. In improvement science, the scale of the test comes from degree of belief (confidence that the idea will lead to the desired outcome, readiness of the system for the change, cost of failure). This is also where the use of an RCT might come in. Randomization and replication are used in improvement science at the smallest scale and during implementation, whereby procedures are revised and refined based on the data generated from the improvement process, and the theory is revised along the way. This process is collaborative and participatory, with those leading the improvement process (akin to the evaluator) and those implementing the processes (akin to program staff) working together to collect and understand data.

Theories are often tested on a small scale and then brought to scale later in the process. Thus, it is critical that the immediate outcomes in the driver diagrams be well articulated, as they may be the only outcomes in the theory that are tested. In evaluation, we often focus on measuring the longer term outcomes of the theory or its impact. Consequently, theories in improvement science are often more dynamic that those that are developed in the context of an evaluation study that may focus on measuring outcomes 3, 5, or even 10 years after program implementation. Whereas the idea of incremental scale-up is nothing new in evaluation (see Weiss, 2010), the procedures and tools for sequential learning are much more formalized and empirically tested in the context of improvement science.

The Way Forward

Evaluation and improvement science intertwine across the use, valuing, and methods dimensions. Cutting across these three dimensions, the cases presented in this issue illustrate a broad range of improvement science applications, offering analytical strategies, data visualization techniques, and data collection strategies to potentially further and support the use of improvement science as an evaluation strategy in specific contexts.

References

American Evaluation Association. (2011). *Public statement on cultural competence in evaluation*. Fairhaven, MA: Author. Retrieved from http://www.eval.org/p/cm/ld/fid=92

Berwick, D. M. (2008). The science of improvement. *Journal of the American Medical Association*, *299*(10), 1182–1184.

Christie, C. A., & Alkin, M. C. (2012). An evaluation theory tree. In M.C. Alkin (Ed.), *Evaluation roots* (2nd ed.). Thousand Oaks, CA: Sage.

Cousins, J. B., & Earl, L. M. (1992). The case for participatory evaluation. *Educational Evaluation and Policy Analysis*, *14*(4), 397–418.

Cronbach, L. J., Ambron, S. R., Dornbusch, S. M., Hess, R. D., Hornik, R. C., Phillips, D. C., … Weiner, S. S. (1980). *Toward reform of program evaluation*. San Francisco, CA: Jossey-Bass.

Deming, W. E. (2000). *Out of the crisis*. Boston, MA: MIT Press.

Donaldson, S. I., Gooler, L. E., & Scriven, M. (2002). Strategies for managing evaluation anxiety: Toward a psychology of program evaluation. *American Journal of Evaluation*, *23*(3), 261–273.

Fetterman, D. M. (1994). Steps of empowerment evaluation: From California to Cape Town. *Evaluation and Program Planning*, *17*(3), 305–313.

Greene, J. (2016). *Advancing equity: Cultivating an evaluation habit. Evaluation for an equitable society*. Charlotte, NC: Information Age Publishing.

House, E. R. (1980). *Evaluating with validity*. Beverly Hills, CA: Sage.

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

Kaplan, H. C., Provost, L. P., Froehle, C. M., & Margolis, P. A. (2012). The model for understanding success in quality (MUSIQ): Building a theory of context in healthcare quality improvement. *BMJ Quality and Safety*, *21*, 13–20.

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.

Meadows, D. H. (2008). *Thinking in systems*. White River Junction, VT: Chelsea Green Publishing.

Mertens, D. M. (1999). Inclusive evaluation: Implications of a transformative theory for evaluation. *American Journal of Evaluation*, *20*(1), 1–14.

Miller, R. L., & Campbell, R. (2006). Taking stock of empowerment evaluation—An empirical review. *American Journal of Evaluation*, *27*(3), 296–319.

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

Patton, M. Q. (2008). *Utilization-focused evaluation* (4th ed.). Thousand Oaks, CA: Sage.

Patton, M. Q. (2010). *Developmental evaluation. Applying complexity concepts to enhance innovation and use*. New York, NY: Guilford Press.

Rother, M. (2009). *Toyota kata: Managing people for improvement, adaptiveness and superior results*. San Francisco, CA: McGraw-Hill Professional.

Scriven, M. (2003). Evaluation in the new millennium: The transdisciplinary view. In S. I. Donaldson & M. Scriven (Eds.), *Evaluating social programs and problems: Visions for the new millennium* (pp. 19–42). Mahwah, NJ: Erlbaum.

Shadish, W. R., Cook, T. D., & Leviton, L. C. (1991). *Foundations of program evaluation: Theories of practice*. Newbury Park, CA: Sage.

Vo, A. (2013). *Toward a definition of evaluative thinking* (Unpublished dissertation). University of California, Los Angeles.

Weiss, C. H. (1998). *Evaluation* (2nd ed.). Upper Saddle River, NJ: Prentice-Hall.

Weiss, C. H. (2010). Scaling impact. *The Evaluation Exchange, XV(1)*. Boston, MA: Author. Retrieved from www.hfrp.org

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.

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.

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.