Recommendations for Evaluation of Health Care Improvement Initiatives

Gareth J. Parry, PhD; Andrew Carson-Stevens, MBBCh, MPhil; Donna F. Luff, PhD; Marianne E. McPherson, PhD, MS; Donald A. Goldmann, MD

Institute for Healthcare Improvement, Cambridge, Mass (Drs Parry and Goldmann); Department of Pediatrics, Boston Children's Hospital, Boston, Mass (Drs Luff and Goldmann); Harvard Medical School, Boston, Mass (Drs Parry, Luff, and Goldmann); Institute of Primary Care and Public Health, Cardiff University, Cardiff, Wales, UK (Dr Carson-Stevens); and National Initiative for Children's Healthcare Quality, Boston, Mass (Dr McPherson)

The views expressed in this report are those of the authors and do not necessarily represent those of the US Department of Health and Human Services, the Agency for Healthcare Research and Quality or the American Board of Pediatrics Foundation.
The authors declare that they have no conflict of interest.
Publication of this article was supported by the Agency for Healthcare Research and Quality and the American Board of Pediatrics Foundation.
Address correspondence to Gareth J. Parry, PhD, Institute for Healthcare Improvement, 20 University Rd, 7th Floor, Cambridge, MA 02138 (e-mail: gparry@ihi.org).
Received for publication October 11, 2012; accepted April 12, 2013.

ABSTRACT

Intensive efforts are underway across the world to improve the quality of health care. It is important to use evaluation methods to identify improvement efforts that work well before they are replicated across a broad range of contexts. Evaluation methods need to provide an understanding of why an improvement initiative has or has not worked and how it can be improved in the future. However, improvement initiatives are complex, and evaluation is not always well aligned with the intent and maturity of the intervention, thus limiting the applicability of the results. We describe how initiatives can be grouped into 1 of 3 improvement phases—innovation, testing, and scale-up and spread—depending on the degree of belief in the associated interventions. We describe how many evaluation approaches often lead to a finding of no effect, consistent with what has been termed Rossi's Iron Law of Evaluation. Alternatively, we recommend that the guiding question of evaluation in health care improvement be, "How and in what contexts does a new model work or can be amended to work?" To answer this, we argue for the adoption of formative, theory-driven evaluation. Specifically, evaluations start by identifying a program theory that comprises execution and content theories. These theories should be revised as the initiative develops by applying a rapid-cycle evaluation approach, in which evaluation findings are fed back to the initiative leaders on a regular basis. We describe such evaluation strategies, accounting for the phase of improvement as well as the context and setting in which the improvement concept is being deployed. Finally, we challenge the improvement and evaluation communities to come together to refine the specific methods required so as to avoid the trap of Rossi's Iron Law.

KEYWORDS: evaluation; methods; quality improvement

ACADEMIC PEDIATRICS 2013;13:S23–S30

WORLDWIDE, INTENSIVE EFFORTS are underway to improve the quality of health care delivery. The credibility and effectiveness of these efforts, as well as their suitability for replication and spread, are difficult to measure and understand because of poor alignment between the aims of improvement initiatives and the evaluation design. Systematic reviews often find that published studies of improvement initiatives are of poor quality and do not meet current standards used to assess evidence-based health care.[1–3] Because improvement initiatives are complex and context sensitive, fixed-protocol randomized controlled trials (the gold standard for evidence-based medicine) are not well suited for the evaluation of improvement initiatives.

Here we describe how improvement initiatives are iterative in nature and can vary depending on whether they are at the initial innovation stage, a more developed testing stage, or at a wider spread stage. We argue that applying a research paradigm and asking the question "Does it work?" may not be helpful for improvement initiatives. Rather, applying an evaluation paradigm can frame the question as, "How and in what contexts does the new model work or can be amended to work?" To answer this question, we propose using theory-driven formative evaluations. Finally, we argue that the evaluation needs to consider the full path of an improvement intervention, from the activities used to engage participants and change how they act to the expected changes in clinical processes and outcomes.

To illustrate the issues central to the assessment of health care improvement, the story of penicillin is helpful. What can penicillin teach us about innovation, testing, scale-up and spread, and degree of belief in health care improvement? Alexander Fleming, Ernst Chain, and Howard Florey were awarded the Nobel Prize for Medicine in 1945 for their work in the development of penicillin. Of

the 3 scientists, Fleming is the best known. However, it is worth considering the roles played by Chain and Florey.

Fleming, born in Scotland, had investigated the properties of staphylococci for several years. In 1928, he went on vacation, leaving behind a discarded and contaminated petri dish. On his return, he found the substance in the petri dish appeared to have killed the bacteria he had been working on. Fleming isolated what he believed to be the active ingredient in the petri dish that had killed the bacteria and named it penicillin.[4] For the following 10 years, he worked with limited success to produce larger quantities of penicillin and test it for use as a surface antiseptic.

Then, in 1935, at Oxford University, a young German-born scientist, Chain, took an interest in the work of Fleming. Chain teamed up with Florey, an Australian-born professor of pathology. Chain and Florey identified a technique for developing greater quantities of penicillin, sufficient for testing in experiments in infected laboratory mice. They started by testing 3 infected mice, all of which recovered. With these data, their degree of belief in the potential of penicillin increased enough for them to enlist another young Oxford scientist, Norman Heatley. With Heatley, they developed methods of producing penicillin in greater quantities, enabling them to test penicillin in 50 mice. The tests proved so successful that their degree of belief in penicillin was sufficient for them to begin testing in humans. They tested penicillin on 3 people who were predicted to die from bacterial infection. All of them survived. With this evidence, they obtained funding to undertake larger clinical trials, which confirmed the effectiveness of penicillin.[5,6] As a result, penicillin was produced and administered on a large scale, saving millions of lives. In 1945, Fleming, Chain, and Florey were awarded the Nobel Prize for Medicine.

At the Florey Centenary lecture in 1998, Sir Henry Harris, professor of medicine at Oxford University, said,

> Without Fleming, no Chain; without Chain, no Florey; without Florey, no Heatley; without Heatley, no penicillin.

Thinking about Harris's statement in terms of improvement, the above sentence might be reworded to say:

> Without Fleming, no innovation; without Chain and Florey, no testing; without Heatley, no wide-scale use of penicillin.

Fleming's evaluation methods were very different from those used in subsequent testing and wide-scale production. During the innovation phase, the emphasis was on understanding the mechanisms of action, rapid iterative testing of hypotheses, and predictions—the essence of the scientific method. Later, during the testing phase, the emphasis was on increasing the degree of belief in penicillin through careful, small-scale testing in animals, testing in animals under more varied conditions, and then moving to humans. Finally, wide-scale testing (akin to clinical trials) explored whether penicillin could be used in many settings. At each phase, the scientists conducted numerous scientific experiments until they had built a sufficient degree of belief that the changes they made were leading to the outcomes they sought.

In the improvement field, it is assumed that people act according to their degree of belief that an intervention will be effective in their setting. Degree of belief is associated with Bayesian statistical perspectives (Appendix), and later, we discuss how adopting such a perspective is central to designing appropriate evaluations for health care improvement.[7,8] (Conceptual pragmatism, as described by C. I. Lewis in 1929, draws similarities between improvement methods and Bayesian statistical perspectives.) Although lacking a standard definition in the improvement field, degree of belief is influenced by expert knowledge and opinion, previous experience with this or similar interventions, and new data collected in their setting, indicating whether the intervention is at an innovation, pilot and feasibility testing, or scale-up and spread phase. The innovation phase can be viewed as a process of discovery and description, where descriptive theories are used to detail new models of care.[9] There is a limited degree of belief about whether the model will be generalizable across a broad range of settings (Fig. 1), and testing occurs within a small number of settings. The testing phase, in which degree of belief is moderate, involves testing models in increasingly varied contexts to determine additional settings in which they can be amended to work. In the scale-up and spread phase, the focus is on increasing the adoption of the model in settings in which the prior testing suggested it is likely to bring about improvement. As in the penicillin example, the appropriate evaluation approaches for improvement models will vary across the innovation, testing, and scale-up and spread phases.

WHY NEW IMPROVEMENT IDEAS FAIL SO OFTEN

Donald Campbell was a leading figure in program evaluation in the United States.[10] Since the 1960s, Campbell's championing of evaluation in public policy led to the wide

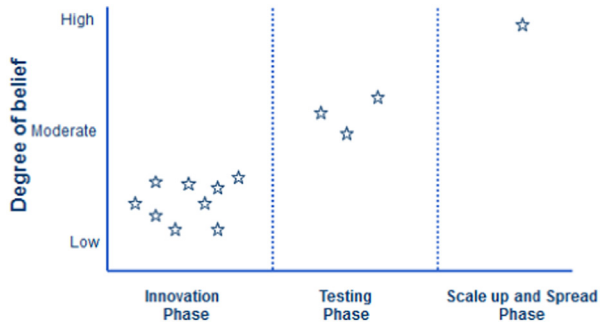

Figure 1. *Degree of belief in an idea and the three phases of improvement. The degree of belief in an idea or intervention will indicate the improvement phase. Ideas associated with a low degree of belief are likely to require thorough exploration and amendment in a small number of settings before they are considered ready for wider testing or spread in more settings. Only those ideas associated with a high degree of belief should be spread widely. Of the many ideas that are identified in the innovation phase, it is likely that only some of those will be considered ready for wider testing and even fewer will be developed and considered ready for wide-scale spread.*

establishment of formal evaluation methods. However, in health care, evaluation often entailed only an impact assessment of the overall intervention, with little focus on the processes involved or the context of the participants.[11] This narrow focus led to a perception that interventions that work in initial studies lose their effectiveness as they are implemented widely. This diminishing effect phenomenon has been described in the program evaluation field by Peter Rossi as the Iron Law of Evaluation[12,13]:

> The expected value of any net impact assessment of any social program is zero. This means that our best a priori estimate of a net impact assessment of a program is that it will have no effect.
> It also means that the average of net impact assessments of a large set of social programs will crawl asymptomatically toward zero.

In health care, Ioannidis found that of 34 interventions that had been replicated, 41% were found to have a smaller effect size or were not found to be effective.[14] In pediatrics, cardiac intensive care units (ICUs) were associated with improved outcomes when first established. As these cardiac ICUs became more widespread, studies reported a diminishing effect on outcomes.[15,16] One explanation for Rossi's Iron Law (Fig. 2) is that the effectiveness of an innovation is often based on studies in a small number of settings—for example, the introduction of a rapid response team for patients identified as deteriorating on the floor.[17] The full range of complexity of the innovation may not be fully understood, and a simple but intuitively appealing summary model of the change formed into a fixed protocol—for example, "When a patient has a physiologic early warning score above *x,* call for 2 ICU nurses and a physician."[18] Replicating the intervention in other organizations similar in context may be effective in 80% of organizations. At face value, this is a success, and several organizations, varying more in context from the earlier sites, may decide to replicate the model as a fixed protocol. After implementing this protocol, the intervention may now work in 70% of these organizations. Although a lower proportion of organizations found the intervention successful, there may be further attempts to replicate in an increasingly broad range of contexts. The intervention may work in only 50% of these organizations—implying an equal chance that it will or will not work. In terms of Rossi's Iron Law, the net impact has tended toward zero. This matters in health care improvement

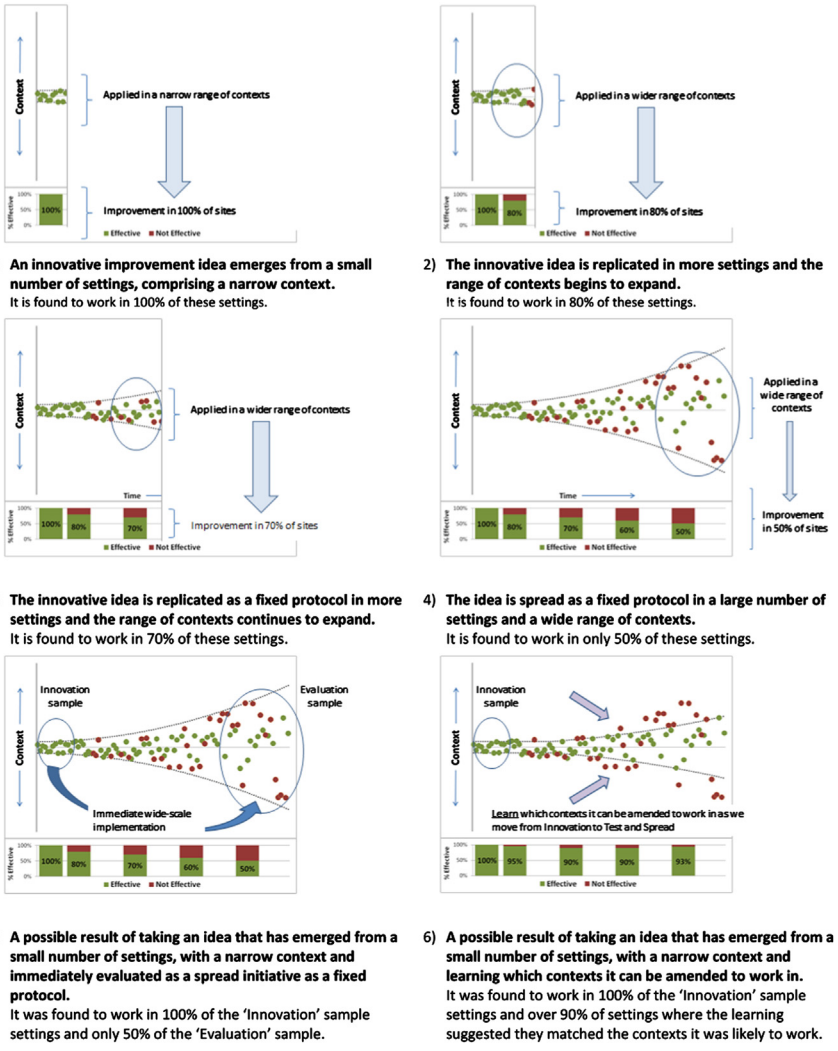

Figure 2. *Reduction in effectiveness of a new improvement idea as it is spread as a fixed-protocol intervention.*

because frequently a promising new model is found effective in a small number of settings, but when replicated as a fixed protocol across a broad range of contexts, it is found to be ineffective.[19]

A strategy of amending models has greater potential to enable spread to a larger number of organizations. This approach calls for evaluation that is theory driven and formative.[20] Evaluators must understand the core concepts that underpin the detailed tasks undertaken as part of a new model. The core concepts are more likely to be generalizable to other organizations than the detailed tasks.[21] For example, the core concepts associated with a rapid response team may include "use a reliable method to identify deteriorating patients in real time" or "when a patient is deteriorating, provide the most appropriate assessment and care as soon as possible." As organizations attempt to introduce a new model, they should start with the core concepts. By using examples of what has worked in other settings, they can test approaches for introducing detailed tasks tailored to their local context. In the rapid response team example, several organizations may start with similar core concepts, but they vary in how they apply the early warning scoring or staff the response team.

Improvement methods offer practical approaches for tailoring conceptual models to local contexts.[22,23] Improvement teams generally prioritize learning what works in their organization; often they are less interested in initiating a larger-scale quantitative evaluation to demonstrate that their adaptations to the concept are suitable for application across a broad range of settings (generalizability).[24] Even in a broader improvement initiative, such as the Breakthrough Series Collaborative, where teams from numerous organizations aim to achieve specific improvement goals, it can be challenging to understand whether success or failure of the improvement initiative related to the new model or to a variety of contextual factors, including the time frame of the collaborative, institutional leadership support and organizational will, local resources and culture, and the ability of organizations to scale up and spread the changes within their system.[25–27] Consequently, to increase the chances that findings will be generalizable, formative evaluation aligned with the improvement approach is required. The following section provides guidance for evaluation approaches for health care improvement.

SUGGESTED APPROACHES TO EVALUATION OF HEALTH CARE IMPROVEMENT

As described above, the guiding evaluative question for health care improvement is, "How and in what contexts does the new model work or can be amended to work?" To answer this question, we propose using theory-driven formative evaluations. The specific approach will be informed by 2 primary considerations. The first is the degree of belief in the new conceptual model and whether it is at the innovation, testing, or scale-up and spread phase. If the improvement work is part of a multiorganization improvement initiative, the second consideration is the rationale for the overall approach, including how and to whom to teach improvement methods.

The Kirkpatrick Framework, used to evaluate training programs, is helpful to consider here.[28] It describes 4 levels of learning opportunity: 1) experience—what was the participants' experience? 2) learning—what did the participants learn? 3) behavior—did they modify their behavior? and 4) results—did the organization improve its performance? (Table 1). For an improvement initiative, the Kirkpatrick Framework helps describe a program theory, or a chain of reasoning from the activities involved in an initiative through to the change in processes and outcomes expected. The program theory may also be broken down into an activity-focused execution theory and a clinical-focused content theory.[29]

Execution theory is defined as the rationale for how the experience provided by the improvement initiative (Kirkpatrick level 1), the teaching delivered (Kirkpatrick level 2), and the learning accomplished leads to improvement in the process measures (Kirkpatrick level 3). For example, consider an initiative aimed at increasing the use of prenatal steroids for pregnant women at risk of premature labor. Here a Breakthrough Series Collaborative may be chosen, with the rationale that it will provide a shared learning experience, where participants will learn about how prenatal steroids are used in similar settings so that they can apply and test these ideas in their own setting.

Content theory is defined as the rationale for how improvement in process measures associated with applying the new model (Kirkpatrick level 3) leads to improvement in organizational performance or patient outcomes (Kirkpatrick level 4). For example, in the above maternal steroids initiative, the theory is that increased use of maternal steroids for mothers at risk of preterm labor will lead to reduced neonatal mortality.

Every evaluation of improvement work should start by seeking clarity on the content and execution theories and the degree of belief in them. For this, an overall program theory, displaying the execution and content theories in the form of a logic model, is helpful and is standard practice in program evaluation.[19,30] A logic model provides a framework for clarifying and illustrating how the

Table 1. Kirkpatrick Framework for Evaluation and Application to Improvement Initiatives

Kirkpatrick Level	Evaluation	Applicability to Improvement Initiatives
1) Experience	What was the participants' experience?	Did the participants have an excellent experience working on the improvement project?
2) Learning	What did the participants learn?	Did the participants learn improvement methods and begin testing?
3) Process	Did they modify their behavior?	Did participants work differently and experience change in the process measures?
4) Outcome	Did the organization improve its performance?	Did the participants' organization improve its outcomes or performance?

activities and inputs associated with a project lead to short-term process improvement (execution theory) and then on to mid- and long-term outcome improvement (content theory).[31] The program theory guides evaluators in developing appropriate research questions and data collection methods, using both qualitative and quantitative methods. Qualitative data indicate what improvement teams are doing and why, where they meet barriers or facilitators to change, and the contexts where success is or is not being achieved and why. Quantitative data illustrate progress toward the goals. These results are used to update the initial program theories. This approach is similar to a Bayesian perspective, where the degree of belief of success of an initial theory is determined and then prospective data are used to update the degree of belief—though, in the case of improvement, the theory may also be updated.[17] Improvement initiatives vary in content, context, and approach, making it difficult to describe approaches to evaluation for every situation. Rather, below, we provide broad guidance on evaluation approaches by improvement phase of innovation, testing, and scale-up and spread.

PHASES OF MODEL TESTING

INNOVATION

The innovation phase aims to generate a new model of care or content theory (Table 2). Evaluation here should describe the new content theory, including the underlying concepts that inform it and the context in which the model was developed. In addition, an evaluation should estimate

Table 2. Summary of Evaluation Aims and Approaches by Improvement Phase

Improvement Phase:		
Innovation	Testing	Scale-up and spread
What is the aim of the improvement phase?		
To generate or discover a new model of care with evidence of improvement in a small number of settings. Example: One or 2 teaching hospitals.	To engage organizations and enable them to test whether a model works or can be amended to work in their context. Example: A broader number of teaching hospitals and some large general hospitals.	To engage organizations to adopt models associated with a high degree of belief in their applicability and impact in a broad range of contexts. Example: A broad number and type of hospitals in a state or country.
What is the aim of the evaluation?		
From a small group of organizations, with limited context, to: • Describe a new content theory. • Provide an estimate of the improvement achieved from applying the content theory. • The degree of belief that the content theory will apply in similar contexts from where it was developed.	From an initial content theory, with moderate degree of belief, to: • Describe an amended content theory. • Provide an estimate of the improvement achieved from applying the amended content theory in specific contexts. The degree of belief that the amended content theory will apply in specific contexts. From an initial execution theory, to: • Describe an amended theory for engaging with organizations in specific contexts to test and amend the new content theory in the future. • Provide an estimate of the likely application of testing and amendment of content theory in the future.	From an initial content theory, with high degree of belief that it will apply in specific contexts, to: • Describe any amendments identified in the spread phase. From an initial execution theory, to: • Describe an amended theory for engaging with organizations in specific contexts to apply the content theory in the future. • Provide an estimate of the likely uptake of the new content theory in specific contexts in the future.
What evaluation approaches may be helpful?		
A quantitative measurement system that focuses on Kirkpatrick levels 3 and 4, to provide estimates of the impact of variations in the development of the content theory. Clarification of the content theory through qualitative interviews with model developers and those who have tested the model, to draw out the underlying concepts, describe them and indicate how they impact on the results obtained. Regular, rapid-cycle feedback of the findings to the leads of the innovation phase.	A quantitative measurement system that focuses on Kirkpatrick levels 1 to 4, to provide estimates of the impact of amendments to the execution and content theories. Longitudinal quantitative data analysis, including control chart and interrupted time series methods, to provide an estimate of the improvement associated with amended content and execution theories. Randomized cluster and stepped-wedge designs. Recommendations for how to amend content and execution theories through qualitative methods to identify how teams did or did not learn and apply their learning, in their local context. Regular, rapid-cycle feedback of the findings to the leads of the testing phase.	A quantitative measurement system that primarily focuses on Kirkpatrick level 3 and secondarily on level 4, to provide estimates of the impact of amendments to the execution theory. Randomized cluster and stepped-wedge designs. Longitudinal quantitative data analysis, including control chart and interrupted time series methods, to provide an estimate of the improvement associated with an amended execution theory. Recommendations for how to amend the execution theory and point to issues with the content theory through qualitative methods to identify how teams did or did not learn and apply their learning, in their local context. Regular, rapid-cycle feedback of the findings to the leads of the scale-up and spread phase.

the measured improvement achieved as a result of the new model in this context and indicate the degree of belief that the model is likely to apply in other settings. For the rapid response team example, this involves describing the underlying concepts behind the model, the details of how a few organizations implemented it, and the impact on patient outcomes.

The rapid-testing approach to progress in innovation can be challenging to an evaluator. Expertise and training in qualitative methods are essential. It is important to interview model developers and those who tested the initial model in order to draw out the underlying concepts, describe them clearly, and indicate how they impact the results obtained. It is critical to allow the interview responses to be as unrestricted as possible, so that new or unexpected changes or experiences can emerge and be captured. It will also be important to develop a clear quantitative measurement system, to provide accurate estimates of the overall impact of the model and any subsequent amendments.

In this innovation phase, substantial learning generated from the rapidly undertaken testing must be fed directly back into the model building process on a regular basis. When Fleming was developing penicillin, he designed and undertook experiments himself; he did not rely on an external evaluator to perform additional data collection and analysis. In a similar way, innovators may be able to develop and evaluate new models themselves. In the innovation phase, the role of an independent evaluator may be analogous to that of an auditor, who checks that what was done and reported is accurate.

Testing

This phase aims to engage organizations to test whether a new model works or can be amended to work in their context (Table 2). Improvement methods and approaches are taught to staff at organizations, with the aim of tailoring a model to their local context. The aim of an evaluation here is to establish in which contexts the new model is likely to work or can be amended to work. In addition, the evaluation will aim to provide an estimate of the likely effect size and an updated degree of belief in the content theory. For the rapid response team example, this involves identifying and describing those contexts in which the model was found to work and with what impact.

If the improvement work is undertaken as part of a wider initiative, such as a collaborative, the aim should be to understand the execution theory and indicate whether it was successful in teaching people improvement methods and having them act on their learning. If not, the evaluation should indicate how the execution theory could be amended in the future. Moreover, it is important to assess at the outset the expected impact of the work and over what time period. To do so, a strong and reliable measurement system is needed. When assessing the content theory, an evaluator needs to measure whether the model, as implemented, resulted in the behavior changes expected in the time period expected (Kirkpatrick level 3). In addition, the evaluator needs to measure whether the model resulted in improvement in organizational performance predicted, in the time period predicted (Kirkpatrick level 4).[32] A qualitative approach, interviewing or observing improvement teams, will help understand how the changes were made, suggest updates to the content theory, and identify potential unexpected consequences.

When assessing the execution theory, it is important to measure whether teams learned and applied according to what was predicted in the initial execution theory (Kirkpatrick levels 2 and 3). Again, qualitative methods can identify how teams did or did not learn and apply their learning, in their local context, to inform updates to the initial execution theory.

To understand whether the new model is effective in a particular context, it is important to ask what would happen if the new model was not introduced (a counterfactual). An individual organization can use longitudinal data analysis, including control chart and interrupted time series methods, to answer this question. For example, the impact of a new model can be calculated from the difference between values predicted from existing trends in data over time and values observed after the new model is introduced. Moreover, with supporting qualitative data that describes how models were introduced, what underpins them, and the influence of unanticipated external factors, the degree of belief in a new model in a particular context can be ascertained. To assess the impact of the wider improvement initiative, an evaluator can estimate the relative impact of the intervention by comparing the results (Kirkpatrick level 3 and 4 measures) with a comparison group of organizations not impacted by the improvement work using a broad range of randomized approaches, for example, clustered randomized designs and stepped-wedge designs.[33]

Last, the evaluation and improvement teams must communicate openly and frequently while still protecting the independence of the evaluator in assessing results. This will allow the evaluator to understand the program and the improvement teams to incorporate rapidly what the evaluation recommends on a regular basis. Such a rapid-cycle evaluation approach will, if appropriate, allow for midcourse adjustments to the content and execution theories. Alternatively, the evaluation may reveal that the new model has little impact outside of the narrow context of organizations within which it was originally developed.

Scale-up and Spread

In the scale-up and spread phase, the focus is on adopting models where there is a high degree of belief in their applicability and impact in a broad range of contexts (Table 2). Here, the content theory is likely to be well tested and substantial amendment is unlikely. The primary focus is on the execution theory, specifically on the methods used to apply the model in local settings (Kirkpatrick level 3). For example, the Door-to-Balloon Alliance used a variety of approaches to attract interest in and engage organizations in the campaign, which resulted in substantial improvements in process measures associated with acute myocardial infarction care.[34]

The aim of the evaluation in this phase is to assess whether the improvement activities associated with the

execution theory worked as predicted and resulted in increased uptake of the model—and how, if necessary, the execution theory may be amended to increase uptake in the future.

Evaluators addressing the scale-up and spread phase of an improvement initiative will need to develop a strong measurement system focused on process measures (Kirkpatrick level 3) and on what people learned from the scale-up and spread methods (Kirkpatrick level 2). An evaluator should also utilize some measurement aimed at assessing overall impact (level 4), especially if monitoring unintended consequences is a concern.

Qualitative approaches are more challenging to undertake at the scale-up and spread phase. Substantial resources would be required to undertake in-depth interviews or to employ ethnographic approaches with all participating organizations. A purposive subgroup sampling approach is more appropriate to gain understanding of how the execution theory has or has not worked in a range of settings and, if necessary, suggest how it may be amended in future spread initiatives.

Similar to the testing phase, a rapid-cycle evaluation approach can provide regular feedback into the ongoing improvement work.

CONCLUSIONS

We recommend that the guiding question for those planning to undertake evaluation of health care improvement be, "How and in what contexts does a new model work or can be amended to work?" Evaluators seeking to answer this question will need to understand whether the improvement work is at the innovation, testing, or scale-up and spread phase. We recommend improvement initiatives clarify a program theory that comprises execution and content theories, illustrated by a logic model. Evaluators may wish to apply the Kirkpatrick Framework in understanding the interplay between the improvement phase and evaluation approach. Finally, we recommend that evaluations account for the iterative nature of improvement work and are undertaken prospectively, generating learning applicable to ongoing improvement efforts and enabling midcourse adjustments to the program theory. We also recommend that improvement initiatives are resourced to include a trained program evaluation researcher to conduct this work.

Here we provide a general approach to undertaking formative, theory-based evaluations of iterative improvement initiatives. To establish this as a more common approach to evaluate improvement initiatives, we must challenge the improvement and academic evaluation communities to come together to provide more details. For example, we need a more standard taxonomy for improvement concepts and context, defined degree of belief and refined specific methods that can provide us with a prediction of how and where new models will work best, and what the cost and quality benefits will be. With this input, there is a greater chance that funders will commission studies based on this approach and that research describing what it takes to achieve improvement in health and health care are more likely to be published. Overcoming this challenge will require innovation, but it is vital if the health system is to move forward and avoid the trap of Rossi's Iron Law.

SUPPLEMENTARY DATA

Supplementary data related to this article can be found online at http://dx.doi.org/10.1016/j.acap.2013.04.007. References 35–39 are cited in the Appendix.

REFERENCES

1. Conry MC, Humphries N, Morgan K, et al. A 10-year (2000–2010) systematic review of interventions to improve quality of care in hospitals. *BMC Health Serv Res*. 2012;12:275.
2. Versteeg MH, Laurant MG, Franx GC, et al. Factors associated with the impact of quality improvement collaboratives in mental healthcare: an exploratory study. *Implement Sci*. 2012;7:1.
3. Mittman BS. Creating the evidence base for quality improvement collaboratives. *Ann Intern Med*. 2004;140:897–901.
4. Fleming A. On the antibacterial action of cultures of a penicillium, with special reference to their use in the isolation of *B influenzae*. *Br J Exp Pathol*. 1929;10:226–236.
5. Florey HW. Penicillin: a survey. *Br Med J*. 1944;2:169.
6. Chain E, Florey HW, Hardner AD, et al. Penicillin as a chemotherapeutic agent. *Lancet*. 1940;239:226–228.
7. Spiegelhalter DJ, Myles JP, Jones DR, et al. Methods in health service research. An introduction to Bayesian methods in health technology assessment. *BMJ*. 1999;319(7208):508–512.
8. Lewis CI. *Mind and the World Order: Outline of a Theory of Knowledge 1929*. Reprint. New York, NY: Dover; 1991.
9. Christiansen C. The ongoing process of building a theory of disruption. *J Prod Innov Manag*. 2006;23:39–55.
10. Campbell DT, Stanley J. *Experimental and Quasi-Experimental Designs for Research. Chicago*. Ill: Rand-McNally; 1963.
11. Landon BE, Wilson IB, McInnes K, et al. Effects of a quality improvement collaborative on the outcome of care of patients with HIV infection: the EQHIV study. *Ann Intern Med*. 2004;140:887–896.
12. Rossi PH. The iron law of evaluation and other metallic rules. *Res Social Problems Public Policy*. 1987;4:3–20.
13. Pawson R, Tilley N. *Realistic Evaluation*. London: Sage; 1997.
14. Ioannidis JPA. Contradicted and initially stronger effects in highly cited clinical research. *JAMA*. 2005;294:218–228.
15. Burstein DS, Jacobs JP, Li JS, et al. Care models and associated outcomes in congenital heart surgery. *Pediatrics*. 2011;127:e1482–e1489.
16. Parry GJ. Replicating cardiac intensive care units: the importance of structure, process, and outcomes. *Pediatrics*. 2011;127:e1595–e1596.
17. Goldhill DR, McNarry AF, Mandersloot G, et al. A physiologically-based early warning score for ward patients: the association between score and outcome. *Anaesthesia*. 2005;60:547–553.
18. Dixon-Woods M, Bosk CL, Aveling EL, et al. Explaining Michigan: developing an ex post theory of a quality improvement program. *Milbank Q*. 2011;89:167–205.
19. Kaplan HC, Brady PW, Dritz MC, et al. The influence of context on quality improvement success in health care: a systematic review of the literature. *Milbank Q*. 2010;88:500–559.
20. Rossi PH, Freeman HE, Lipsey MW. *Evaluation: A Systematic Approach*. 7th ed. Thousand Oaks, Calif: Sage; 2004.
21. Ovretveit J, Leviton L, Parry GJ. Increasing the generalisability of improvement research with an improvement replication programme. *BMJ Qual Saf*. 2011;20:i87–i91.
22. Langley GJ, Moen RD, Nolan KM, et al. *The Improvement Guide: A Practical Approach to Enhancing Organizational Performance*. New York, NY: Jossey-Bass; 2009.
23. Deming WE. *Out of the Crisis*. Cambridge, Mass: MIT Press; 2000.

24. Provost LP. Analytical studies: a framework for quality improvement design and analysis. *BMJ Qual Saf*. 2011;20:i92–i96.
25. Institute for Healthcare Improvement. *The Breakthrough Series: IHI's Collaborative Model for Achieving Breakthrough Improvement*. Boston, Mass: Institute for Healthcare Improvement. Available at: http://www.ihi.org; 2003. IHI Innovation Series White Paper. Accessed June 18, 2013.
26. Benning A, Dixon-Woods M, Nwulu U, et al. Multiple component patient safety intervention in English hospitals: controlled evaluation of second phase. *BMJ*. 2011;342:d199.
27. Berwick DM. The question of improvement. *JAMA*. 2012;307: 2093–2094.
28. Kirkpatrick DL, Kirkpatrick JD. *Evaluating Training Programs*. 3rd ed. San Francisco, Calif: Berrett-Koehler Publishers; 2006.
29. Lipsey MW. Theory as method: small theories of treatments. *New Directions Progr Eval*. 1993;57:5–38.
30. Millar A, Simeone RS, Carnevale JT. Logic models: a systems tool for performance management. *Eval Progr Planning*. 2001; 24:73–81.
31. W. K. Kellogg Foundation. Logic model development guide. Available at: http://www.wkkf.org/knowledge-center/resources/2006/02/wk-kellogg-foundation-logic-model-development-guide.aspx. Accessed March 26, 2013.
32. Lilford RJ, Chilton PJ, Hemming K, et al. Evaluating policy and service interventions: framework to guide selection and interpretation of study end points. *BMJ*. 2010;341:c4413.
33. Brown CA, Lilford RJ. The stepped wedge trial design: a systematic review. *BMC Med Res Methodol*. 2006;6:54.
34. Bradley EH, Nembhard IM, Yuan CT, et al. What is the experience of national quality campaigns? Views from the field. *Health Serv Res*. 2010;45:1651–1669.
35. Lilford RJ, Braunholtz D. Who's afraid of Thomas Bayes? *J Epidemiol Commun Health*. 2000;54:731–739.
36. Chaloner K, Church T, Louis TA, et al. Graphical elicitation of a prior distribution for a clinical-trial. *Statistician*. 1993;42:341–353.
37. Parmar MKB, Spiegelhalter DJ, Freedman LS. The chart trials: Bayesian design and monitoring in practice. *Stat Med*. 1994;13:1297–1312.
38. Lilford R, Fetal Compromise Group. Formal measurement of clinical uncertainty: prelude to a trial in perinatal medicine. *BMJ*. 1994;308: 111–112.
39. Parry GJ, Tucker J. The effect of the United Kingdom Neonatal Staffing Study results on the prior views of neonatal doctors: a Bayesian analysis. The role of Bayesian analysis in health policy decision making. In: Tavakoli M, Davies H, eds. *Health Care Policy, Performance, and Finance: Strategic Issues in Health Care Management*. Aldershot, UK: Ashgate; 2004.