ISQua®

International Journal for Quality in Health Care, 2018, 30(S1), 29–36
doi: 10.1093/intqhc/mzy021
Methodology Paper

OXFORD

Methodology Article

Practical recommendations for the evaluation of improvement initiatives

GARETH PARRY[1,2], ASTOU COLY[3], DON GOLDMANN[1,2,4], ALEXANDER K. ROWE[5], VIJAY CHATTU[6], DENEIL LOGIUDICE[7], MIHAJLO RABRENOVIC[8,9], and BEJOY NAMBIAR[10,11]

[1]Institute for Healthcare Improvement, 53 State Street, 19th Floor, Boston, MA 02109, USA, [2]Harvard Medical School, 25 Shattuck St, Boston, MA 02115, USA, [3]USAID Applying Science to Strengthen and Improve Systems (ASSIST) Project, University Research Co., LLC, Chevy Chase, MD, USA, [4]Harvard School of Public Health, 677 Huntington Ave, Boston, MA 02115, USA, [5]Division of Parasitic Diseases and Malaria, Center for Global Health, U. S. Centers for Disease Control and Prevention, Building 24, Room 03-217, Mailstop A06, 1600 Clifton Road, Atlanta, GA 30329-4027, USA, [6]Public Health and Primary Care Unit, Faculty of Medical Sciences, The University of the West Indies, Trinidad and Tobago, School of Global Health & Bioethics, EUCLID University Champ Fleurs, Trinidad, West Indies, [7]Quality and Process Improvement Consultant, [8]Faculty of Business Economics and Entrepreneurship, Belgrade, Serbia, [9]Chairman of Management Board, The Institute of Virology, Vaccines and Sera 'Torlak', Belgrade, Serbia, [10]Institute for Global Health, UCL, 30, Guilford Street, London WC1N 1EH, UK, and [11]Academy of Medical Sciences, Malawi University of Science and Technology (MUST), Limbe, Malawi

Address reprint requests to: Gareth Parry, Institute for Healthcare Improvement, 53 State Street, 19th Floor, Boston, MA 02109, USA; Harvard Medical School, 25 Shattuck St, Boston, MA 02115, USA. Tel: +1 617 301 4974; E-mail: gparry@ihi.org

Editorial Decision 24 January 2018; Accepted 5 February 2018

Abstract

A lack of clear guidance for funders, evaluators and improvers on what to include in evaluation proposals can lead to evaluation designs that do not answer the questions stakeholders want to know. These evaluation designs may not match the iterative nature of improvement and may be imposed onto an initiative in a way that is impractical from the perspective of improvers and the communities with whom they work. Consequently, the results of evaluations are often controversial, and attribution remains poorly understood. Improvement initiatives are iterative, adaptive and context-specific. Evaluation approaches and designs must align with these features, specifically in their ability to consider complexity, to evolve as the initiative adapts over time and to understand the interaction with local context. Improvement initiatives often identify broadly defined change concepts and provide tools for care teams to tailor these in more detail to local conditions. Correspondingly, recommendations for evaluation are best provided as broad guidance, to be tailored to the specifics of the initiative. In this paper, we provide practical guidance and recommendations that funders and evaluators can use when developing an evaluation plan for improvement initiatives that seeks to: identify the questions stakeholders want to address; develop the initial program theory of the initiative; identify high-priority areas to measure progress over time; describe the context the initiative will be applied within; and identify experimental or observational designs that will address attribution.

Key words: improvement, learning, complex adaptive systems, implementation, delivery

Background

Recently, progress has been made in the improvement field to develop clearer guidance on how to describe the methods and results of improvement initiatives. For example, the Standards for Quality Improvement Reporting Excellence (SQUIRE) 2 provides guidelines on how to describe the problem an improvement initiative aimed to address, the rationale for the improvement approach, relevant contextual issues, what was found and what the findings mean [1]. Building on SQUIRE, guidance developing the evaluation design of improvement initiatives will be helpful. The current lack of clear guidance for funders, evaluators and improvers on what to include in evaluation proposals can lead to evaluations that do not answer the questions stakeholders want to know, designs that do not match the iterative, adaptive nature of improvement, or designs imposed onto an initiative that are impractical from the perspective of improvers and the communities with whom they work [2, 3]. Consequently, the results of evaluations are often controversial, and attribution remains poorly understood [4].

Improvement initiatives often identify broad change concepts and provide tools to tailor these to local conditions [5, 6]. For example, teams may be encouraged to place handwashing signs in places that make most sense to them, rather than in some pre-specified location. If teams find, as a result of their testing, that improvement occurs, they are encouraged to start testing in other settings, and move towards implementation. Moreover, across several sites, with varying capability and contexts the improvement work is likely to move at varying speeds. In some sites entirely, different approaches may be needed to bring about improvement. As such, improvement initiatives seldom follow a fixed protocol. The optimal evaluation will time specific activities such as data collection to the timing of the improvement activities. For example, there will be no point in collecting data from all settings in a site, if they are only testing in one or two places. Consequently, the timing of evaluation activities will need to follow the timing of the improvement activities. Thus, the activities and focus of improvement initiatives are likely to change over time as those undertaking local testing at the point of care learn what does and does not work with their context.

A major aspect of any intervention and its evaluation is the design. Much literature exists on the importance of including the evaluation team from the start of any project or program design. One approach, described by Leviton and colleagues, based on Evaluability Assessment, offers guidance relevant to improvement initiatives [2]. An evaluability assessment features components that include involving the intended users of evaluation information, clarifying the intended program, exploring the likely impact of the program, reaching agreement on needed changes in activities or goals, exploring alternative evaluation designs and agreeing on evaluation priorities and intended users of the information. Leviton describes an iterative process that funders and evaluators can follow to understand the program and develop the most appropriate evaluation design given available resources and time.

The Evaluability Assessment approach can be used to guide the development of an evaluation design for improvement projects and programs by:

- Agreeing among all key stakeholders, including the funder, improvement and evaluation teams, on the Theory of Change
- Agreeing among all key stakeholders, on the evaluation design, including:
  - The evaluation questions
  - Formative and/or summative approaches
  - Availability and use of data to assess attribution
  - Available human and financial resources

In this paper, we use the Evaluability Assessment approach to guide the design of evaluations for understanding attribution in improvement initiatives. We identify tools and approaches commonly used in the improvement field to provide practical guidance and recommendations for developing a proposal to evaluate improvement initiatives.

Agreeing on the Theory of Change

For improvement initiatives, the theory of change includes three interdependent pieces: the 'What', the 'Context' and the 'How', and below we summarize and recommend existing tools and approaches for describing these, and how they may be included in an evaluation plan:

The What

The 'what' describes the rationale for what changes, if made locally, may lead to improved outcomes. In improvement, Driver Diagrams are often used to illustrate the primary drivers the improvement team has identified and predict, if implemented, will lead to improvement [7]. From these primary drivers, detailed secondary drivers are identified and used by improvement teams to generate specific change ideas for local testing. Figure 1 shows an example of a Driver Diagram for an initiative that aims to reduce newborn mortality across six health care districts. The three primary drivers of the intended outcome are prioritizing increased access, activating local champions, and reliably delivering a clinical pre-natal care bundle. Developing Driver Diagrams requires clinical knowledge, critical appraisal skills, quantitative data skills and facilitation skills (Box 1).

Context

There are many considerations when seeking to understand the context within which an improvement initiative is conducted. In addition, context can change and be an active component of an initiative. There are growing options available for characterizing some aspects of context. To fully understand the impact of context, an evaluation plan must allow for the possibility that additional contextual themes will emerge [8, 9]. Contextual exploration likely requires social science input, including qualitative methods that consider the following:

- How context may interact with the 'what' and 'how' over the course of the improvement initiative
- An approach to describe how the context varies across participating sites
- A description of the process the improvement and evaluation teams will use to identify and measure the likely impact of context on the improvement initiative (Box 2).

The How

The evaluation plan must describe the entire intervention, including how people at the point of care are expected to test, implement or scale-up the changes depicted in the Driver Diagram. For example, if an IHI Breakthrough Series is used, the evaluation plan must describe the rationale for how the activities (learning sessions, action periods) will result in local change in the timescales available [10].

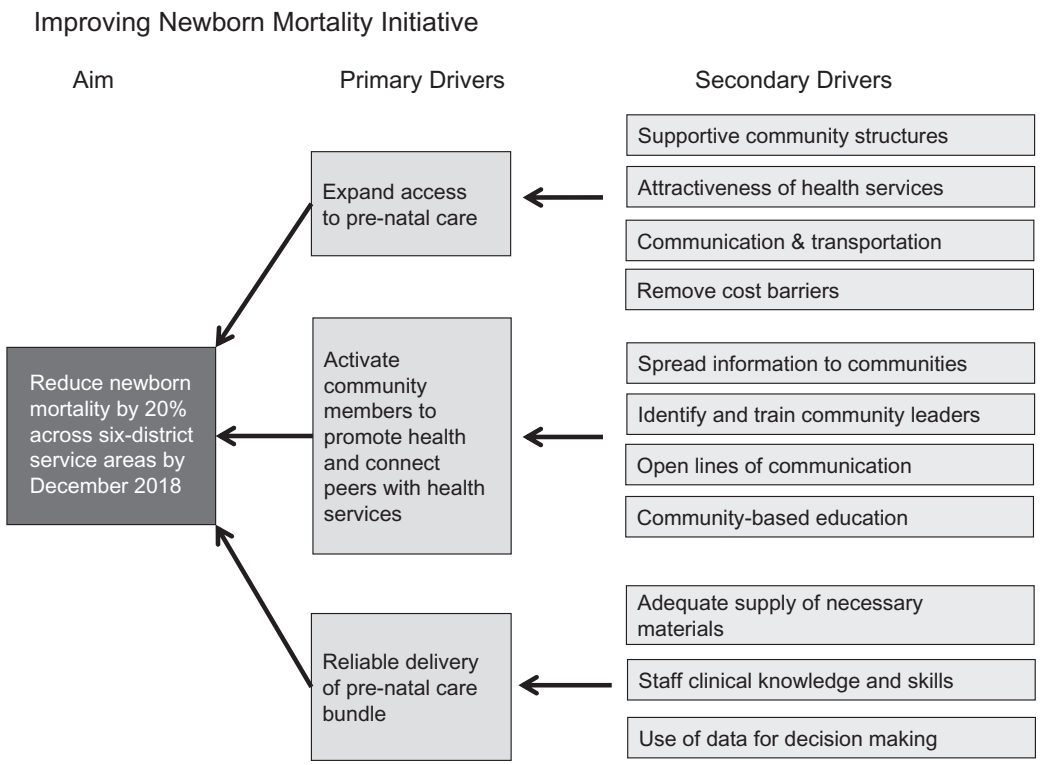

Figure 1. Example of a Driver Diagram summarizing 'What' changes the initiative predicts will lead to the improvement goal.

Box 1 Improving newborn mortality initiative

The recommendations can be illustrated by following the design of an evaluation for an initiative to improve newborn mortality in 15 maternity centers in a region of a low-income country.

Following the evaluability evaluation approach:
Agreeing on the theory of change:
The What: The improvement initiative leaders summarized the changes they recommended teams embarking on the improvement initiative follow in the form of a Driver Diagram (Fig. 1). The primary drivers of change they selected were to expand access to pre-natal care, activate community members to promote health and connect peers with health services and reliable delivery of pre-natal care bundle. Each of these primary drivers had associated secondary drivers, for example, to expand access to pre-natal care, the improvement leaders identified the need to provide supportive community structures, attractive health services, communication and transportation and remove cost barriers for pregnant women.

Box 2 Agreeing on the theory of change:

The Context: Following a series of interviews with participants, the improvement leaders understood that many of the participants were new to quality improvement, and required additional training in improvement methods. In addition, they found some variation among the participating sites in terms of leadership engagement, and the extent to which the goals of the initiative aligned with the strategic priorities of the organizations.

Logic models are used to depict the causal pathway from available resources to program activities to local short-term process changes, to long-term process changes and to improved outcomes [11]. Figure 2 shows an example of a Logic Model for the Newborn Mortality Reduction Initiative. The model aims to illustrate links between the planned improvement activities and how, over time, it may result in teams testing and showing improvement in process measures and implementing, spreading, and achieving sustained improved outcomes. A causal pathway illustrated in a logic model should specify:

- How planned activities in an improvement initiative will lead to outputs in local settings.
- How activities will lead to short-term outcomes in, for example, short-term process measures.

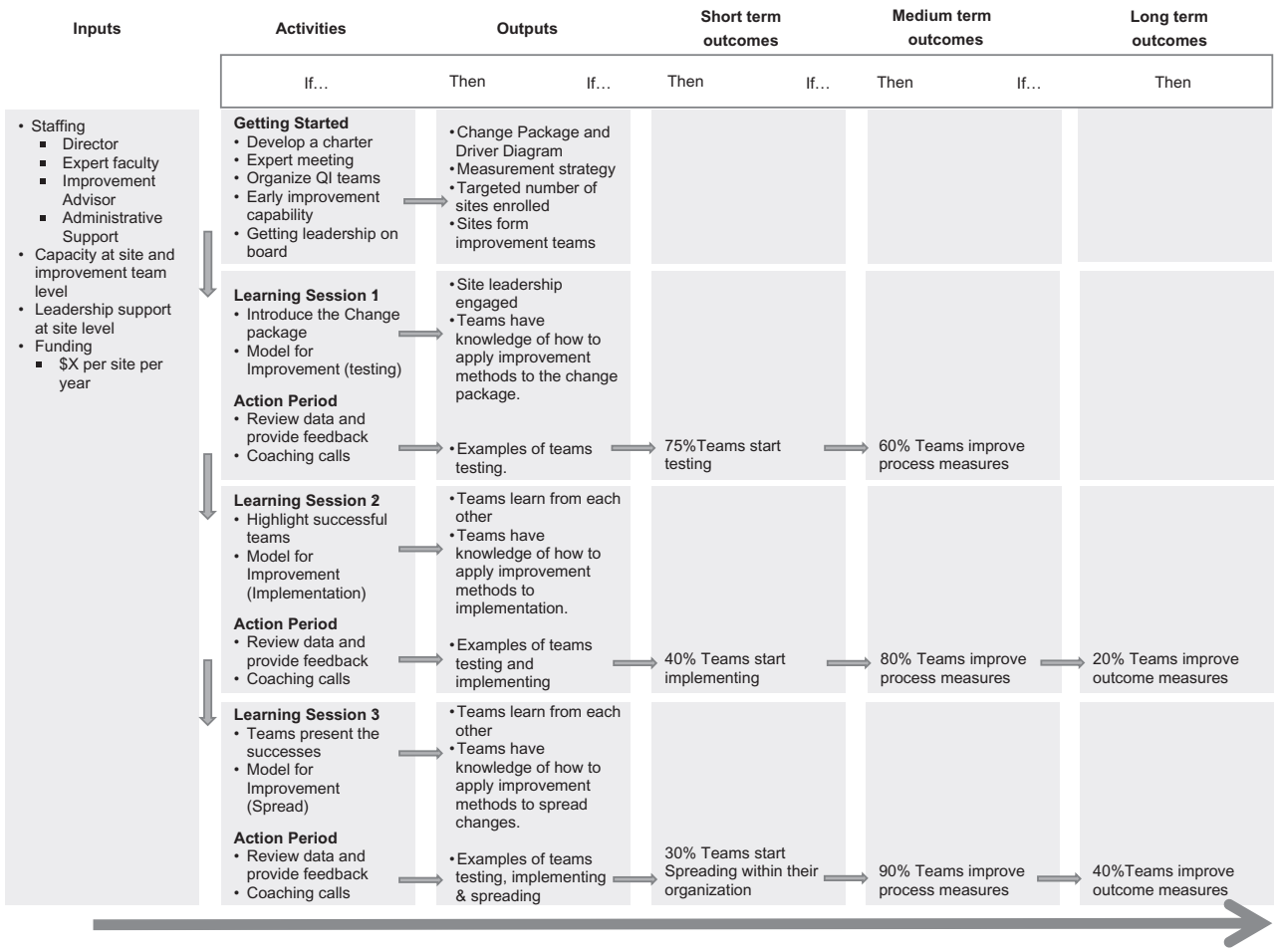

Figure 2 Example of a logic model Illustrating 'How' activities of the improvement initiative will facilitate local testing of the changes.

Box 3 Agreeing on the theory of change:

The How: The improvement leaders initially proposed to follow an IHI Breakthrough Series Collaborative design, where a series of three learning sessions, bringing together participating teams are conducted. The learning sessions aim to teach improvement methods that will encourage participants to test local changes aligned with the content of the driver diagram. Between learning sessions, participants are expected to test changes and share learning with each other. Based on feedback related to the Context, the improvement leaders decided to add additional coaching calls with the aim of strengthening improvement capability within participating teams. The improvement leaders summarized the How into a Logic Model (Fig. 2).

- How local short-term process measures will lead to the improvement depicted in the primary or secondary drivers of improvement (Box 3).

Developing the evaluation design

As the theory of change is developed, appropriate evaluation questions and designs must be developed simultaneously. For key stakeholders and evaluators to agree on an evaluation design, the evaluability assessment approach recommends they review the theory of change and decide what evaluation questions they want to address, whether to use a formative or summative approach, and what data to collect. Only then will they be able to identify options for designs that will address attribution. The final choice of design will then be made in relation to available resources.

Agreeing on the evaluation questions

A key component of improvement initiatives is having a clear goal, agreed by stakeholders for what will be achieved. An evaluation plan must demonstrate an understanding of the major stakeholders and their roles, including a description of their evaluation goals and how the evaluation will align with them. For example, a government agency may want to know the overall impact of the initiative, and the cost to achieve wider scale-up. Improvement leads may want to know where the program worked and how a program can be

Table 1 Examples of evaluation questions by improvement phase

The What	The Context	The How
Innovation phase: Model development typically takes place in a small number of settings, and evaluation questions should focus largely on the What, for example:		
• What is the overall impact of the model on health care quality and patient outcomes? • Which elements of the model had the greatest impact on patient outcomes?		
Testing phase: Testing phase, the aim is to identify where a model works or can be amended to work. Hence, although refining The What will occur, developing The How and The Context will also be important. Example evaluation questions include:		
• What is the overall impact of the overall model on health care quality and patient outcomes? • Which elements of the model had the greatest impact on patient outcomes?	• To what extent can all the changes be implemented? • What are barriers and facilitators to implementing the changes locally? • What are the barriers and facilitators to undertaking the improvement activities as planned?	• To what extent can all the changes be implemented? • What are barriers and facilitators to implementing the changes locally? • What are the barriers and facilitators to undertaking the improvement activities as planned?
Spread and scale-up phase**:** The aim is to spread or scale-up the model in contexts earlier work has indicated it is likely to work or be amended to work. Here, the What and the Context should be well developed, and the focus will be primarily on the How. Evaluation questions may include:		
• What is the overall impact of the overall model on health care quality and patient outcomes?	• To what extent did the impact of the model vary across settings? • To what extent did the implementation vary from the model vary across settings? What contextual factors are associated with the implementation of the model?	• To what extent can all the changes be implemented? • What are barriers and facilitators to implementing the changes? • What are the barriers and facilitators to undertaking the activities as planned?

Box 4 Agreeing on the evaluation design:

The evaluation questions

The improvement leaders had data demonstrating the changes described in the Driver Diagram had been implemented with a positive impact on outcomes in a number of other locations and settings. Working with the funders, and with an evaluation team the funders had identified, they jointly decided they were at the Spread and Scale-Up phase (Table 1), where they were aiming to spread the changes to contexts in which earlier work has indicated they would likely work. They decided to focus their evaluation questions primarily on how the changes would be implemented and secondarily on the overall impact of successful implementation on patient outcomes. The evaluation questions were:

1.1) To what extent can all the changes be implemented?
1.2) To what extent did the implementation vary from the model across settings?
1.3) What are barriers and facilitators to implementing the changes?
1.4) What are the barriers and facilitators to undertaking the improvement activities?
2) What is the impact of the overall model on patient outcomes?

amended to work in the future. Examples of evaluation questions by improvement phase are shown in Table 1.

Having clarified stakeholder questions, options for evaluation designs are required. The evaluation design should be clearly described. Issues to consider include the choice of summative and/or formative approaches, the desired strength of internal and external validity, availability of data, identification of counterfactuals and available resources for evaluation (Box 4).

Formative and/or summative approaches

Summative evaluation assesses program impact and determines the degree to which the program was successful. Formative evaluation aims to improve programs as they evolve. As shown in Table 1, the assignment of an initiative into one of the improvement phases of innovation, testing, and spread and scale-up can guide the choice of using summative or formative evaluation approaches. In the innovation and testing phases, the 'What' and 'How' are under development, and will adapt over time. In the spread phase, the 'what' and 'how' will be more developed, and less likely to adapt over time.

For initiatives in the innovation and testing phase, formative approaches are primarily used and feature mixed quantitative and qualitative methods, to capture how, where and with what impact models are being adapted over time. For initiatives in the spread and scale-up phase, summative approaches are primarily used and feature quantitative methods to estimate the overall impact of the model. Formative evaluation approaches should be considered in the spread and scale-up phase, particularly if adaptation and issues related to context are likely to occur during spread and scale-up (Box 5).

Box 5 Agreeing on the evaluation design:

Formative and/or summative approaches

The evaluation team recommended a formative approach to address evaluation questions 1.1–1.4, where the evaluation team would work with the improvement leads to collate and analyse process and outcome data at quarterly intervals to provide feedback on progress. For question 2, the evaluation team recommended a summative approach, where they would collate and analyse data at the end of the improvement work to assess overall impact on patient outcomes.

Box 6 Agreeing on the evaluation design:

Availability and use of data:

For evaluation questions 1.1 and 1.2, the evaluation team and improvement leads identified several key process measures aligning with the driver diagram that participants were expected to collect as part of their improvement work on a monthly basis. They agreed for the evaluation team to have access to this data every quarter, and for the evaluation team to plan a data quality exercise to assess the validity of the data.

For evaluation questions 1.3 and 1.4, the evaluation team recommended semi-structured qualitative interviews and site visits be undertaken with a sample of eight participating sites.

Moreover, the evaluation team recommended setting aside a time every quarter to understand whether the improvement leaders planned to change their approach based on the feedback, so that the evaluation plan could also be amended accordingly.

For evaluation question 2, the evaluation team planned to obtain data from an administrative system, available 6 months after the end of the improvement work. This administrative system was subjected to regular data quality audits.

Availability and use of data:

To understand where an improvement change works or can be adapted to make it work, clearly defined and prioritized data and measurement plans are required. These allow the evaluation to explore overall impact and variation across sites. To understand progress towards the goals of an improvement project and how progress can be attributed to the theory of change, several high-priority measures are required, aligning with progress on the Logic Model and Driver Diagram. In the Newborn Mortality example, the Driver Diagram in Fig. 1 and Logic Model in Fig. 2 suggest an outcome measure of newborn mortality, and process measures indicating access to pre-natal care, community champion activation, and reliability of care bundle delivery. Additional key measures on the causal pathway suggest those indicating gains in knowledge and application of improvement methods. As suggested in Fig. 2, specific dates by which measures are predicted to reach a specific target are needed. Measurement should focus on the overall outcomes, areas where the logic model suggests are high-leverage drivers or steps that are critical to have achieved, as well as areas where there exists a lack of consensus in the strength of evidence related to a step. Details should be provided regarding how frequently the data will be collected and how it is collected and validated. The data are ideally available in a format that allows for analysis of variation across sites. Measurement by participants at the point of care is an important feature of most improvement initiatives, and evaluators might consider building on existing participant data collection activities. However, if they do, it will be important to pay specific attention to data quality issues. As the limitations of data collected by participants might be difficult to overcome (e.g. data quality changing over time as the data collection skills of participants improve, and the positive bias that can be introduced when workers rate their own performance), evaluations desiring high internal validity might need outcomes to be measured by objective data collectors (e.g. high-quality surveyors). Finally, consideration should be given to measures that indicate unintended consequences, indicating whether a loss of quality has occurred elsewhere in the system (balancing measures) (Box 6).

Assessing attribution

With an understanding of the above areas, several mixed-method formative and summative evaluation designs relevant to the evaluation questions can be developed. We recommend using quantitative methods to estimate the impact or attribution of the initiative, and explore variation across settings; and qualitative methods to explore questions related to how or why an initiative did or did not progress and to understand issues related to adaptation of the models being implemented [12].

To assess attribution, the extent to which measured improvement results from the initiative, evaluation plans must describe a counterfactual—what is likely to happen if the initiative is not introduced. Table 2 summarizes approaches that can be used as the core of an evaluation design to assess attribution. Randomization should be explored where possible to assess attribution of specific aspects of the improvement initiative. We recommend evaluators base their design on one of these approaches and build the qualitative methodology, exploring questions of how and why around them. Table 3 summarizes additional issues for evaluators to consider depending on the Improvement Phase of the Initiative (Box 7).

Availability of human and financial resources

Developing, establishing and maintaining data collection systems are vital to the success of improvement initiatives, but is resource intensive. An evaluation may also want to extend data collection to comparator sites, increasing resource requirement. Moreover, surveys, site visits, and qualitative interviews, and ethnographic techniques whilst providing valuable insights, are resource intensive. Following the Evaluability Assessment approach, the evaluation designers will need to re-visit and amend the questions they identified earlier (see 3.1), and prioritize resources accordingly (Box 8).

Table 2 Core evaluation designs for assessing attribution

Basic features	When to use
Factorial design Two or more interventions and a control group are compared. Participants (patients or sites) are randomized to each intervention independently	To compare two or more models of a multifaceted model in one or two settings, where the context is well-understood (e.g. comparing three improvement initiatives: one with coaching only, one with learning sessions only, and one with both coaching and learning sessions).
Stepped-wedge design Participants are assigned to an intervention or control group for a defined period. After this initial study period, control participants transfer into the intervention group. Can be randomized or non-randomized	To take advantage of the delayed implementation of an intervention in a representative sample of settings, and explore their use as comparator sites; or when an intervention is thought to be beneficial and/or when it is impossible or impractical to deliver the intervention to everyone at the same time
Controlled before and after study (CBA) Data are collected before and after the implementation of an intervention, both in one or more groups that receive the intervention and in a control group that is similar to the intervention group but that does not receive the intervention	This non-randomized design can be used to establish a counterfactual when random assignment is not ethical, possible or practical
Interrupted time series study (ITS) Data are collected at multiple time points before and after an intervention to determine whether the intervention has had an effect significantly greater than what would have been predicted by extending the baseline trend into the follow-up period	Where comparators are not available, a counterfactual can be estimated by the baseline trend (i.e. a historical control). The validity of this design can be strengthened by having a separate control group. Random allocation may be used to allocate to the intervention).
Cluster randomized controlled trials Randomization allocation of subjects or sites to an intervention or control group can be introduced if it is possible and appropriate. The internal validity of the above evaluation designs are stronger if study clusters such as health facilities are randomly assigned to intervention or control groups. Validity can be further strengthened by matching cluster groups on attributes associated with study outcomes[a]	

[a]E.g. in a two-armed study, create pairs of health facilities with similar attributes, then randomly assign one facility per pair to the intervention group, with the remaining facility being a control.

Table 3 Design considerations by improvement phase

Innovation phase:
- When developing a multifaceted model in a small number of settings, with a well-understood context, approaches such a factorial design can be explored [16]

Testing phase:
- Explore the possibility and potential implications of delaying implementation of the initiative in a representative sample of settings, and explore their use as comparator sites [17]
- Consider whether the model can be disaggregated and assessed using a stepped-wedge design [18]
- Where comparators are not available, a counterfactual may be available through use of longitudinal baseline data from participating sites, prior to participation in the initiative
- Where comparators are not available, use longitudinal data to assess for association in the take up of processes or activities with changes in outcomes
- In the testing phase, an approach to understanding where a model can be attributed to have worked or amended to work is important, and an exploration of the activity, process and outcome data across sites, and across specific contexts should be undertaken

Spread and scale-up phase:
- Explore the possibility and potential implications of delaying implementation of the initiative in a representative sample of settings, and explore their use as comparator sites
- When aiming to develop an approach for how a model can be implemented, consider whether the implementation model can be disaggregated and assessed using a stepped-wedge design
- Where comparators are not available a counterfactual may be available through use of longitudinal baseline data from participating sites, prior to participation in the initiative
- Where comparators are not available, compare within settings, activity and process data

For all improvement phases:
- If an experimental approach is not appropriate, the evaluation plan must offer an option for describing how attribution of outcomes can be assessed through triangulation of numerous analytical and research tasks

Summary

In this paper, guided by the iterative evaluability assessment approach, we provide recommendations to inform funders and evaluators on what to look for when evaluating improvement initiatives. We focus these recommendations on designs that allow for the attribution of improvement results to improvement initiatives. These recommendations can be used to inform the development and review of Requests for Proposals, or in discussion with improvement practitioners when developing projects or programs. Additionally, our recommendations align with the SQUIRE 2 [1].

Box 7 Agreeing on the evaluation design:

Availability and use of data to assess attribution

A neighboring region also had access to the changes described in the Driver diagram (Fig. 1), but were not intending to actively support the spread of them. The evaluation team recommended that data on implementation of the changes and patient outcomes in this region be used to assess attribution. Moreover, the evaluation team recommended undertaking more in-depth qualitative approaches to explore attribution and to supplement the analysis by exploring the association of process and outcome measures over time for each participating site.

Box 8 Agreeing on the evaluation design:

Available human and financial resources

In review with the funder and the improvement leaders, funding was not available to cover the initial evaluation design. The funder, improvement leaders and evaluation teams worked together to update the evaluation questions and design so that process and outcome measures were reviewed two times per year rather than four, and to no longer prioritize evaluation question 1.4.

Recognizing the call for strengthening the evidence-base, the use of randomization in evaluation designs should also be explored [13]. Moreover, evaluation designs must consider the iterative, adaptive nature of improvement. Perhaps the most important aspect to look for in evaluation is the flexibility of evaluation design to respond to likely adaptations in how the improvement activities are applied across local contexts, recognizing that some sites will move more quickly than others. The complex adaptive nature of improvement requires additional interpretative work in evaluation design to fully understand The What and How and their interaction with The Context. As Dixon-Woods and others have argued, such work will benefit from sound theoretic underpinnings to understand underlying concepts and themes related to the 'What' and 'How' that may be applicable to other settings [14, 15].

Acknowledgments

The authors would like to acknowledge L.M., R.B., J.H. and K.L. of USAID for their review and suggestions for this paper.

References

1. Goodman D, Ogrinc G, Davies L *et al*. Explanation and elaboration of the SQUIRE (Standards for Quality Improvement Reporting Excellence) Guidelines, V. 2.0: examples of SQUIRE elements in the healthcare improvement literature. *BMJ Qual Saf* 2016;**25**:e7.
2. Leviton LC, Khan LK, Rog D *et al*. Evaluability assessment to improve public health policies, programs, and practices. *Annu Rev Public Health* 2010;**31**:213–33.
3. Parry GJ, Power M. To RCT or not to RCT?: more on the role of randomisation in quality improvement. *BMJ Qual Saf* 2015. DOI:10.1136/bmjqs-2015-004862.
4. Parry GJ, Carson-Stevens A, Luff DF *et al*. Recommendations for evaluation of health care improvement initiatives. *Acad Pediatr* 2013;**13**:S23–30.
5. Ovretveit J, Leviton L, Parry GJ. Increasing the generalisability of improvement research with an improvement replication programme. *BMJ Qual Saf* 2011;**20**:i87–91.
6. Lilford RJ. Implementation science at the crossroads. *BMJ Qual Saf* 2017. doi:10.1136/bmjqs-2017-007502.
7. Bennett B, Provost L. What's YOUR Theory? *Qual Progrss* 2015;**48**:36.
8. Kaplan HC, Provost LP, Froehle CM *et al*. The Model for Understanding Success in Quality (MUSIQ): building a theory of context in healthcare quality improvement. *BMJ Qual Saf* 2012;**21**:13–20.
9. Damschroder LJ, Aron DC, Keith RE *et al*. Fostering implementation of health services research findings into practice: a consolidated framework for advancing implementation science. *Implement Sci* 2009;**4**:50.
10. The Breakthrough Series. *IHI's Collaborative Model for Achieving Breakthrough Improvement. IHI Innovation Series White Paper*. Boston: Institute for Healthcare Improvement, 2003. www.IHI.org.
11. Goeschel CA, Weiss WM, Pronovost PJ. Using a logic model to design and evaluate quality and patient safety improvement programs. *Int J Qual Health Care* 2012;**24**:330–7. mzs029.
12. Coly A, Parry G Evaluating Complex Health Interventions: A Guide to Rigorous Research Designs. AcademyHealth. 2017 http://www.academyhealth.org/evaluationguide (10 January 2018, date last accessed).
13. National Academies of Sciences, Engineering, and Medicine. *Improving Quality of Care in Low-and Middle-income Countries: Workshop Summary*. Washington, DC: National Academies Press, 2015.
14. Davidoff F, Dixon-Woods M, Leviton L *et al*. Demystifying theory and its use in improvement. *BMJ Qual Saf* 2015;**24**:228–38. bmjqs-2014-003627.
15. Lipsey MW. Theory as method: small theories of treatments. *New Direct Prog Eval* 1993;**57**:5–38.
16. Fisher RA. *Statistical Methods, Experimental Design and Scientific Inference*. Oxford, UK: Oxford University Press, 1935.
17. Campbell DT, Stanley J. *Experimental and Quasi-Experimental Designs for Research*. Chicago: Rand-McNally, 1963.
18. Hemming K, Haines TP, Chilton PJ *et al*. The stepped wedge cluster randomised trial: rationale, design, analysis, and reporting. *BMJ* 2015;**350**:h391.