OXFORD

Methods Article

Criteria for evaluating programme theory diagrams in quality improvement initiatives: a structured method for appraisal

LAUREL ISSEN, THOMAS WOODCOCK, CHRISTOPHER MCNICHOLAS, LAURA LENNOX, and JULIE E. REED

Imperial College London, National Institute for Health Research Collaboration for Leadership in Applied Health Research and Care Northwest London, Chelsea and Westminster Hospital, 4th Floor, Lift Bank D, 369 Fulham Road, London SW10 9NH, UK

Address reprint requests to: Julie E. Reed, Imperial College London, National Institute for Health Research Collaboration for Leadership in Applied Health Research and Care Northwest London, Chelsea and Westminster Hospital, 4th Floor, Lift Bank D, 369 Fulham Road, London SW10 9NH, UK. Tel: +020 3315 8141; E-mail: Julie.Reed02@lmperial.ac.uk

Abstract

Background: Despite criticisms that many quality improvement (QI) initiatives fail due to incomplete programme theory, there is no defined way to evaluate how programme theory has been articulated. The objective of this research was to develop, and assess the usability and reliability of scoring criteria to evaluate programme theory diagrams.

Methods: Criteria development was informed by published literature and QI experts. Inter-rater reliability was tested between two evaluators. About 63 programme theory diagrams (42 driver diagrams and 21 action–effect diagrams) were reviewed to establish whether the criteria could support comparative analysis of different approaches to constructing diagrams.

Results: Components of the scoring criteria include: assessment of overall aim, logical overview, clarity of components, cause–effect relationships, evidence and measurement. Independent reviewers had 78% inter-rater reliability. Scoring enabled direct comparison of different approaches to developing programme theory; action–effect diagrams were found to have had a statistically significant but moderate improvement in programme theory quality over driver diagrams; no significant differences were observed based on the setting in which driver diagrams were developed.

Conclusions: The scoring criteria summarise the necessary components of programme theory that are thought to contribute to successful QI projects. The viability of the scoring criteria for practical application was demonstrated. Future uses include assessment of individual programme theory diagrams and comparison of different approaches (e.g. methodological, teaching or other QI support) to produce programme theory. The criteria can be used as a tool to guide the production of better programme theory diagrams, and also highlights where additional support for QI teams could be needed.

Key words: quality improvement, quality management, measurement of quality, programme theory

Introduction

Quality improvement (QI) initiatives have grown rapidly in number in response to the need to reduce unwarranted variation and improve quality and value of care. Despite this growth [1, 2], evidence of sustained benefits remains limited [3–8].

A recurring challenge to improvement practice and evaluation is that QI initiatives often lack a clear programme theory linking interventions directly to intended outcomes [9, 10]. Accurately defining hypothesised relationships (cause and effect) provides a comprehensive and prioritised list of interventions, plus support for subsequent monitoring of implementation and effectiveness [11], and consideration of how interventions may translate to other contexts [12]. Well-defined programme theory allows the social challenges of QI to be addressed by creating a shared aim among all who will be impacted by the proposed service change and increasing staff engagement to support implementation [13, 14].

Several conceptual models exist to identify and articulate programme theory including driver diagrams, action–effect diagrams and logic models [15–19]. Although differences among these approaches exist, key features of these include the ability to:

- help a group to explore the factors that they believe need to be addressed in order to achieve a specific overall goal or outcome,
- show how the factors are connected,
- act as a communication tool for explaining a change strategy and
- provide the basis for a measurement framework.

Evidence suggests that programme theory remains underdeveloped and/or poorly articulated [19]. To date the practical application of such models in frontline healthcare settings has been poorly studied; consequently, there is a lack of information about how to evaluate the quality of programme theory diagrams. A systematic method of assessing programme theory quality would guide better use from initial setup through implementation and potentially maximise benefits in routine practice. Clear evaluation criteria would also provide a future research method to determine factors that best facilitate programme theory articulation, and assess the overall impact of programme theory on QI conduct.

The objective of this study was to develop scoring criteria to assess the quality of programme theory diagrams; to test the usability and inter-rater reliability of scoring programme theory diagrams and to assess whether the criteria could be used to compare different approaches to constructing programme theory diagrams.

Methods

Overview

Theoretical and practical benefits of programme theory were codified into scoring criteria. The usability of these criteria was tested by assessing programme theory diagrams of different types and from different organisational sources and calculating inter-rater reliability between two scorers.

To compare and identify strengths and weaknesses of different approaches to constructing programme theory diagrams, this study compared driver diagrams to action–effect method diagrams, and compared diagrams generated within a single organisational context (NIHR CLAHRC NWL) to diagrams generated in other settings.

Criteria development

Theoretical and practical benefits of programme theory were codified into scoring criteria. The criteria were based on established literature describing the theorised benefits of driver diagrams [15, 17] and more general theory about the aspects of pre-planning deemed important for the long-term success of QI [1, 5–8, 10–14, 20–23]. This theoretical knowledge was combined with the practical knowledge gained from experience of Collaboration for Leadership in Applied Health Research and Care Northwest London (CLAHRC NWL) staff involved in supporting planning, conduct and evaluation of improvement initiatives.

Iterative development of the criteria was led by one author (L.I.) who undertook informal interviews with CLAHRC NWL staff and appraisal of the proposed criteria. L.I., a postdoctoral mixed methods researcher, was not previously involved in development or teaching of the production of programme theory diagrams using either driver diagrams or the action–effect method. Interviews were conducted after the production of all diagrams. Two scorers (L.I. and L.L.) then tested the criteria with 10 sample diagrams. L.L., a PhD student in healthcare QI and registered nurse, had not previously been involved in development of the action–effect method or scoring criteria. After a further four cycles of criteria testing and clarity-based modification between L.I. and L.L., inter-rater reliability was 92% on an expanded set of 10 sample diagrams.

Setting

The NIHR commissioned regional CLAHRCs to support the systematic and effective translation of research into practice, and to improve the quality of care for patients (NIHR, 2011). In CLAHRC NWL, a suite of QI methods supported initiatives to deliver care improvements. This approach was driven by an overarching research agenda to investigate the application and impact of QI methods in healthcare.

Four rounds of QI projects (March 2009–September 2013) were selected by a competitive process open to healthcare organisations in NWL. About 55 initiatives were selected from primary, secondary, mental health and public healthcare settings covering diverse clinical topics. The initiatives established frontline QI teams which engaged multidisciplinary staff and patients, and were supported by CLAHRC NWL in training, facilitation and expert support to use QI methods.

Diagram inclusion

About 63 programme theory diagrams were selected for assessment with the scoring criteria. About 22 driver diagrams (produced between 2009 and 2011), 21 action–effect diagrams (produced between 2011 and 2014) in QI initiatives affiliated with NIHR CLAHRC NWL and 20 driver diagrams from a systematic search of driver diagrams published externally between 2009 and 2011.

CLAHRC NWL diagrams

In two rounds of CLAHRC NWL QI projects (March 2009–April 2011), teams were encouraged to use a suite of QI methods including driver diagrams [17] (Section 1, Appendix A). We refer to diagrams produced in this phase as CLAHRC NWL driver diagrams.

In the next two project rounds (April 2011–September 2013), the action–effect method [16] was iteratively co-developed with frontline QI teams to include greater clarity around diagram components, and how these components are distinguished and interrelate with each other. (Section 2, Appendix A). This provides the distinction between the 22 driver diagrams produced in Rounds 1 and 2, and 21 action–effect diagrams produced in Rounds 3 and 4. Sections 1 and 2, Appendix A represent formal training received by teams using driver diagrams and action–effect methodology, respectively.

External driver diagrams

A systematic search was conducted in January 2012 (concurrent with the end of CLAHRC NWL Round 2) for peer-reviewed journal articles containing the terms 'driver diagram' and 'health'. This search primarily produced articles advocating the use of driver diagrams with few examples of published driver diagrams produced to aid improvement in existing healthcare practice. To find a sample of diagrams produced in a similar context to CLAHRC NWL driver diagrams, we conducted a Google Image search in January 2012 for 'driver diagram' and 'health' and selected the first 20 published diagrams that indicated they had been produced as part of service improvement and redesign.

Application of scoring criteria to programme theory diagrams

The scoring criteria were used to assess the 63 diagrams by two authors (L.I. and L.L.). The scorers used the final criteria outlined in Table 1, with possible scores of zero (does not meet the requirement) to three (excellent example of requirement) for each question.

Total scores for each diagram were the composite sum of scores for each of the eight criteria questions giving a maximum score of 24. In attempt to blind scorers to diagram source, diagrams were unlabelled as to diagram type, and were grouped by clinical subject matter. This resulted in a shuffled but not strictly random ordering of diagram type. Due to stylistic and formatting similarities among diagrams of the same type, scorers may have inferred diagram type.

Our primary consideration in this paper was to determine whether the scoring criteria could be applied in a standardised way. Thus, inter-rater reliability was calculated between final scores given to each of 63 diagrams by the two scorers, using the ordinal score of Krippendorff's alpha [24].

Assessing and comparing approaches to developing programme theory diagrams

Comparative analysis was performed between programme theory types and settings (external driver diagram, CLAHRC NWL driver diagram and action–effect diagram). Each diagram's total score and score for each individual criterion were calculated as the average between two scorers, comparing median scores for each set of diagrams. Significance of this comparison was tested using the non-parametric Kruskal–Wallis rank sum test with significance level of 0.05. Thematic assessment of results was used to consider the strengths and weaknesses of diagram types.

Results

Firstly, we present the scoring criteria with rationale as to their development. This is followed by an assessment of their reliability when applied to scoring a variety of programme theory diagrams. Finally, we consider the application of scoring criteria to compare different approaches to developing programme theory.

Programme theory scoring criteria

The programme theory scoring criteria are summarised in Table 1 and presented in full in Section 3, Appendix A.

Rationale for scoring criteria—compiling evidence and experience

Scoring category: overall aim (Question 1). The need for healthcare improvement efforts to articulate an overall aim is well-documented [1, 5, 8, 15–17, 21, 25, 26]. The aim should be focused on service user needs [15–17, 21] and agreed by all major stakeholders [1, 8, 15–17, 25, 26]. While some sources recommend including measures, interventions and timelines in the aim statement [15, 17], this often conflicts with the objective of widespread engagement and agreement on the aim. Often, not all stakeholders agree that a particular intervention would be effective, or that a particular outcome measure is the most critical [16]. Furthermore, including these aspects in the aim statement serves to obscure cause–effect relationships and measurement concepts which should be made explicit through fully articulated programme theory. Thus, the quality criteria maintain that the overall aim should be stated separately from other components.

Scoring category: logical overview (Question 2). Programme theory dictates that broad categories of factors should be considered to

Table 1 Abbreviated scoring criteria. Full scoring guidance is presented in Section 3, Appendix A

Criteria category	Criteria questions
Overall aim (Q1)	1. Is the overall aim: – High-level? – Focused on the service user? – Indicating direction and aspiration? – Free from interventions, cause–effect relationships, hypotheses, assumptions?
Logical overview (Q2)	2. Is the first column (major contributing factors): – A comprehensive breakdown of the aim? – A systematic breakdown of the aim?
Clarity of components and Cause–effect chains (Q3, Q4, Q5)	3. Do all factors have a clear meaning to the potential audience? 4. Cause–effect relationships: – Is it clear which factors are proposed to be causally linked? – Is it clear which linked factors are proposed to be the cause and the effect? – Are the proposed cause–effect relationships between linked factors plausible and free from leaps in logical reasoning? 5. Is it clear the extent to which cause–effect relationships are evidenced?
Measurement and evaluation (Q6, Q7, Q8)	6. Do the measure concepts have a clear meaning to the potential audience? 7. Is it clear to the potential audience why and how each measure represents the factors with which it is associated? 8. Is there an even distribution of measures at different levels of control and influence across the diagram?

For each of these eight questions score from 0–3 as follows: 0 = does not meet requirements; 1 = meets some of the criteria but has major issues, or a few instances meet requirements; 2 = largely meets requirements or most instances meet requirements and 3 = excellent example of requirement.

ensure that there are no gaps in intervention planning [20, 21]. Stakeholders should be confident that if each of these factors performs well in the system, the overall aim will be achieved. If these high-level factors are not considered systematically, it is difficult to determine whether they portray a comprehensive picture of all factors that could contribute to the aim [16].

Scoring category: clarity of concepts and cause–effect chains (Questions 3–5). Programme theory involves a clear articulation of how activities and interventions are logically proposed to achieve the overall aim [10, 12, 14, 21, 23, 27, 28]. This is a complex undertaking, often underspecified in improvement efforts [10, 20, 23]. Furthermore, tacit knowledge is a precarious method of storing information [22]. Thus, cause–effect chains to be fully and explicitly articulated [28] without relying on tacit information to follow the chain of cause and effect. Often proposed factors and interventions are themselves unclear, a related but distinct issue (Section 3, Appendix A). The evidence base is an important facet of programme theory as well [7, 8, 14, 28, 29] both in terms of articulating how components of the logic model are evidenced [14, 28] as well as identifying the source and strength of the evidence [28, 29] (recognising that even where logic models are evidence-based, this understanding can guide evaluation [30]).

Scoring category: measurement and evaluation (Questions 6–8). Clarity of plans for evaluation and measurement is an important attribute of programme theory [1, 5, 7, 11–13, 21–23, 27, 29, 30]. In addition to these plans being clear to all stakeholders, it is important for the measure concepts to be aligned with proposed cause–effect relationships [5, 12, 21, 27, 30] and involve both process and outcome measures [7, 28]. Process measures help guide the implementation process and can provide rapid feedback to how well the intervention and implementation activities are working and guide adjustments. Outcome measures provide useful information about the impact of an intervention, which usually takes longer to assess.

Reliability of assessments using the programme theory scoring criteria

Inter-rater reliability between scores given to each diagram was 78% (Krippendorff's alpha, ordinal scale) which is within the acceptable range for criteria usability [24].

Use of scoring criteria to compare different approaches to developing programme theory

Of a maximum overall score of 24, the median composite score for external driver diagrams was 6.25, for CLAHRC NWL driver diagrams 5.75 and for CLAHRC NWL action–effect diagrams 11.5 (Figure 1). The setting in which programme theory diagrams were conducted (CLAHRC NWL driver diagrams compared to external driver diagrams) had no significant effect with similar scores observed for each set. The type of diagram used had a significant but moderate difference with action–effect diagrams scored higher over both types of driver diagrams (chi-squared = 19.6941, df = 2, P < 0.0001). The average score for action–effect diagrams was under half the total possible score.

Comparing individual scoring criteria reveals that action–effect diagrams tended to perform better than driver diagrams, independent of setting, for the quality of overall aim, clarify of cause and effect relationships and distribution and clarify of measures. Both diagram types did not include explicit reference to existing evidence base supporting cause and effect chains. Boxplots for comparison of the individual
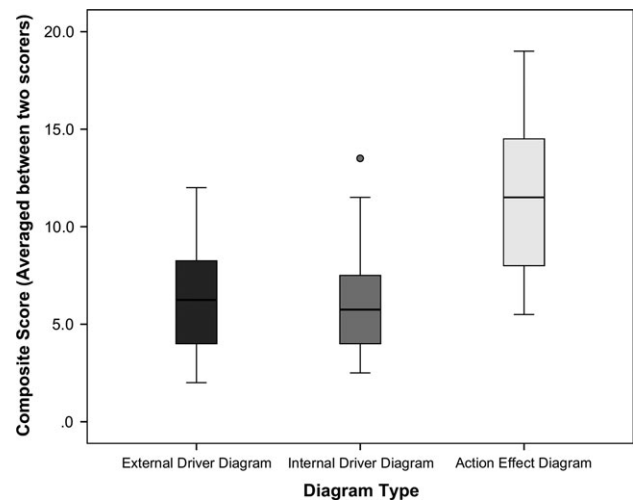

Figure 1 Boxplot diagram of total composite scores, averaged (mean) between scorers, for each diagram set (external driver diagram, CLAHRC NWL driver diagram and CLAHRC NWL action–effect diagram). For all boxplot diagrams, the median is marked by a thick horizontal line, the upper quartile by the box above the line and the lower quartile by a box below the line. Whiskers indicate values 1.5 times above and below the interquartile range. Circles indicate outlier cases between 1.5 and 3 times the interquartile range, and asterisks indicate outliers greater than 3 times the interquartile range.

scoring criteria assessments can be found in Section 4, Appendix A along with description of how rationale for assessments and examples of good practice.

Discussion

Failure to provide clear programme theory is linked with failure to deliver or sustain improvement. The criteria developed here are the first of their kind and can enable practitioners and researchers to assess the quality of programme theory output, regardless of which approach was used to construct the diagram. QI tools will be of maximum benefit to teams when they are used as intended, but evidence suggests that programme theory is often underdeveloped and poorly articulated [30]. These criteria will help guide high-quality production of programme theory diagrams, and provide a structured method for researchers to evaluate their use.

The criteria, including categories of overall aim, clarity of components and cause–effect chains, and measurement, were developed by building on the existing literature of programme theory. New concepts were introduced only when they were evidenced from practical experience and provided generalisable lessons. For example, the action–effect method guidance [16] expanded the definition of the overall aim, stating it should be based on the concept of 'To improve health for service-users'. This was informed by experience recognising the potential for a patient-centred aim to facilitate engagement with diverse stakeholder groups, and the need for this aim to be free from measures and interventions which may be inconsistent or controversial. The quality criteria proposed in this paper provide a foundation for future work to develop and achieve professional consensus on wider applicability and generalisability.

The criteria demonstrated good reliability for assessing individual programme theory diagrams, with inter-rater reliability of 78%. The criteria also show potential for comparing and contrasting different

approaches to constructing programme theory diagrams. Driver diagrams produced in different organisational settings received similar average scores suggesting that the scores reflect underlying attributes of the programme theory approach rather than specific variations by which the approach was applied in different settings. While the action–effect diagrams in this sample scored significantly higher than the sampled driver diagrams, their median score was 11.5 out of 24 points, indicating further room for improvement. The criteria highlight those aspects of programme theory that require greatest improvement and these could be improved with targeted guidance, expert facilitation or support (e.g. clarity of evidence base, logical overview quality). A limitation of this study is that only two types of approaches to constructing programme theory were studied. Further research is required to explore the application of the scoring criteria in other approaches including logic models [18, 19].

The data suggest that the construction of a quality programme theory diagram is conceptually difficult. Based on the experience of the authors, we suggest it requires significant expertise in QI methodology that cannot be provided through written instruction and light-touch facilitation alone. Further research is needed to review facilitation approaches or additional technical assistance necessary to improve the quality of programme theory diagrams. This research opportunity is provided by the existence of scoring criteria as a systematic method for determining the success of facilitation and technical assistance improving the quality of resulting programme theory diagrams.

Further research is needed to investigate whether expert facilitation or technical assistance can encourage engagement in programme theory and iterative development and use of the programme theory diagrams to support constructive dialogue and exchange of tacit knowledge between stakeholders, and to reduce the cognitive burden associated with diagram construction. Facilitation or technical assistance could lead to more substantial improvements in developing factors and cause–effect chains that are less reliant on tacit knowledge, link to the evidence base and build a robust evaluation framework [29]. This is theorised in the literature to lead to improvements in team functioning and buy-in to the improvement project as well as aiding the spread of success to other environments and initiatives [22, 23, 28]. One important consideration is the challenge of exposing tacit knowledge which, due to its 'sticky' nature, is often difficult for frontline staff to perceive and share with outsiders, and therefore presents a barrier to communicate their reasoning to a lay audience [31].

Conclusion

This is the first structured approach to assess the quality of programme theory developed by QI teams in practice. The scoring criteria incorporate a summary of published literature and practical experience regarding the benefits of programme theory.

The robustness and viability of the scoring criteria for practical application was demonstrated by 78% inter-rater reliability between two independent scorers. The scoring criteria were able to detect differences in diagram type (action–effect diagram versus driver diagram) independently of the setting in which diagrams were constructed.

Future uses include assessment of individual programme theory diagrams and comparison of different approaches (e.g. methodological and teaching) to produce programme theory. The criteria can be used as a tool to guide the production of better programme theory diagrams and also highlights where additional support for QI teams could be needed.

Supplementary material

Supplementary material is available at *International Journal for Quality in Health Care* online.

Ethics approval and consent

Only documentary analysis was used; thus, individual consent was not applicable.

Consent for publication

Not applicable: This manuscript does not contain data from any individual person.

Availability of data and materials

An anonymised sample of the dataset has been made available as Supplementary material accompanying this article.

List of abbreviations

QI	quality improvement
NIHR	National Institute for Health Research
CLAHRC NWL	Collaboration for Leadership in Applied Health Research and Care for Northwest London

References

1. Øvretveit J. Quality collaboratives: lessons from research. *Qual Saf Health Care* 2002;**11**:345–51.
2. Boaden R, Harvey G, Moxham C *et al. Quality Improvement: Theory and Practice in Healthcare*. Coventry: NHS Institute for Innovation and Improvement, 2008.
3. Øvretveit J, Gustafson D. Evaluation of quality improvement programmes. *Qual Saf Health Care* 2002;**11**:270–5.
4. Ferlie E, Shortell S. Improving the quality of health care in the United Kingdom and the United States: a framework for change. *Milbank Q* 2001;**79**:281–315.

5. Rogers P. Using programme theory to evaluate complicated and complex aspects of interventions. *Evaluation* 2008;**14**:29–48.

6. Michie S, Johnston M, Abraham C *et al*. Making psychological theory useful for implementing evidence based practice: a consensus approach. *Qual Saf Health Care* 2005;**14**:26–33.

7. Stetler C, Mittman B, Francis J. Overview of the VA Quality Enhancement Research Initiative (QUERI) and QUERI theme articles: QUERI Series. *Implement Sci* 2008;**3**:8.

8. Grol R. Beliefs and evidence in changing clinical practice. *BMJ* 1997;**315**: 418–21.

9. Davies P, Walker A, Grimshaw J. A systematic review of the use of theory in the design of guideline dissemination and implementation strategies and interpretation of the results of rigorous evaluations. *Implement Sci* 2010;**5**:14.

10. Shojania K, Grimshaw J. Evidence-based quality improvement: the state of the science. *Health Aff* 2005;**24**:138–50.

11. Walshe K. Understanding what works—and why—in quality improvement: the need for theory-driven evaluation. *Int J Qual Health Care* 2007;**19**:57–9.

12. Foy R, Øvretveit J, Shekelle P *et al*. The role of theory in research to develop and evaluate the implementation of patient safety practices. *BMJ Qual Saf* 2011;**20**:453–9.

13. Dixon-Woods M, Tarrant C, Willars J *et al*. How will it work? A qualitative study of strategic stakeholders' accounts of a patient safety initiative. *Qual Saf Health Care* 2010;**19**:74–8.

14. Bartholomew L, Mullen P. Five roles for using theory and evidence in the design and testing of behavior change interventions. *J Public Health Dent* 2011;**71**:S20–33.

15. Langley G, Moen R, Nolan K *et al. The Improvement Guide: A Practical Approach to Enhancing Organizational Performance*. San Francisco, CA: John Wiley & Sons, 2009, 512.

16. Reed J, McNicholas C, Woodcock T *et al*. Designing quality improvement initiatives: the action effect method, a structured approach to identifying and articulating programme theory. *BMJ Qual Saf* 2014;**23**:1040–8.

17. Bennett B, Provost L. What's your theory? *Qual Progress* 2015;**48**:36–43.

18. Prasanta K. Managing healthcare quality using logical framework analysis. *Manag Serv Qual* 2006;**16**:203–22.

19. Funnell S. *Purposeful program theory: Effective use of theories of change and logic models*. San Francisco, CA: John Wiley & Sons, 2011.

20. Plsek P. Tutorial: management and planning tools of TQM. *Qual Manag Health Care* 1993;**1**:59–72.

21. Plsek P. Quality improvement methods in clinical medicine. *Pediatrics* 1999;**103**:203–14.

22. Bate P, Robert G. Knowledge management and communities of practice in the private sector: lessons for modernizing the national health service in England and Wales. *Public Adm* 2002;**80**:643–63.

23. Benn J, Burnett S, Parand A *et al*. Studying large-scale programmes to improve patient safety in whole care systems: challenges for research. *Soc Sci Med* 2009;**69**:1767–76.

24. Hayes A, Krippendorff K. Answering the call for a standard reliability measure for coding data. *Commun Methods Meas* 2007;**1**:77–89.

25. Senge P. *The Fifth Discipline: The Art and Practice of the Learning Organization*. New York, NY: Doubleday, 1990: 142–52.

26. Parboosingh I, Reed V, Caldwell Palmer J *et al*. Enhancing practice improvement by facilitating practitioner interactivity: new roles for providers of continuing medical education. *J Contin Educ Health Prof* 2011;**31**:122–7.

27. Rogers P. Causal models in program theory evaluation. *New Dir Evall* 2000;**87**:47–55.

28. Grant A, Treweek S, Dreischulte T *et al*. Process evaluations for cluster-randomised trials of complex interventions: a proposed framework for design and reporting. *Trials* 2013;**14**:15.

29. Hunter S, Chinman M, Ebener P *et al*. Technical assistance as a prevention capacity-building tool: a demonstration using the getting to outcomes framework. *Health Educ Behav* 2009;**36**:810–28.

30. Davidoff F, Dixon-Woods M, Leviton L *et al*. Demystifying theory and its use in improvement. *BMJ Qual Saf* 2014;**24**:228–38.

31. Brown J, Duguid P. Knowledge and organization: a social-practice perspective. *Organ Sci* 2007;**12**:198–213.