Reviewer's report

Title: Designing evaluation studies to optimally inform policy: what factors do policymakers in China consider when making resource allocation decisions on healthcare worker training programs?

Version: 0 Date: 05 Nov 2017

Reviewer: Lewis Husain

Reviewer's report:
The article assesses the perceived usefulness of the Kirkpatrick framework for assessing capacity development programmes and whether these should be scaled up, using interviews and FGDs with a number of informants in the Chinese health system.

The article looks at an under-researched topic, at least in the Chinese case - what do decision-makers want from evaluations? What information do they think is important in deciding on whether a given programme should be scaled up or continued? How can evaluations help support decision-making through providing useful evidence?

Methodology and data

The methodology is well explained, and the use of a small sample of provincial and national officials plus staff from local-level health facilities in a mixed interview/FGD strategy is fine. However, the methodology and/or presentation of data could be stronger in the following ways:

First, the priorities and views of different interviewees are likely to be quite different. For example, a national or provincial CDC official can be expected to have a different view on the potential sustainability or scale up of a training programme to that of a clinician or manager in a local health facility. It would be helpful if the quotes and viewpoints given in the text could be better linked to the category/position of the speaker, even if this is approximate and did not identify the speaker (e.g. "senior CDC official" or "CDC official", "facility manager", etc).
Second, the sample is likely to include an inherent bias which merits discussion. Discussion in the Evaluation methodology section of the manuscript (lines 377-384) indicates a very high level of awareness of analytical and evaluation methods among the interviewees - the paper reflects an informed critical discussion of methodology with the interviewees. It is not discussed to what extent we should expect this kind of critical reflection to be present in other groups we might be interested in as users of evaluation findings. This study has worked with CDC officials and hospital managers, who are likely to have a background in clinical medicine, public health, epidemiology or statistics. In the reviewer's experience, this level of methodological sophistication may not be present across the provincial health system more broadly (or other government departments) in China. The extent to which it will exist in other LMICs with possibly substantially lower levels of general capacity is a question that should be discussed in relation to the wider relevance of this study.

Scope and conceptualisation of the study

Overall, my view is that the conceptualisation is reasonable, despite some limitations. The reason for this is assessment is, fundamentally, two-fold:

First, the study presents the Kirkpatrick framework as an evaluation approach that would be used as a standalone tool for evaluation of a programme. While it can be, many evaluations use multiple methods and approaches in combination including, for example, capacity development tools + value for money analysis, etc. in other words, while the authors' criticisms of the Kirkpatrick framework are broadly valid, the study doesn't fully reflect the real-world context of the conduct of evaluations.

In my view, this is a question of the framing of the article - it would be better if the article presented its framework not as a modification to an invariant analytical model, but rather as a finding that Kirkpatrick on its own may not be sufficient, and that it needs to be combined with other methods / approaches to data collection and analysis - rather than the current solution, which is an expanded model. This would better reflect how methods are used in practice.
Second, the study presents a composite evaluation approach of Kirkpatrick + other components. Some of these are reasonable assessments to include in an evaluation, while others are less persuasive:

1. The study calls for "Broader or indirect programmatic results of the training program". This makes sense as an add-on, and is simply a request that assessment of a training programme also explore its wider impacts.

2. Resources required. An evaluation of a training programme could relatively simply collect evidence on the direct and indirect resources required for continued implementation. Many do. The text, however, does not clearly distinguish between two related ideas: the costs of a programme and the long-term availability of resources to support a programme. Assessment of the first is a reasonable thing to request of an evaluation. The second is a political/policy question that will require negotiation between multiple actors, and is not something that an evaluation can meaningfully inform on. Discussion of this (lines 288-290 of the manuscript) should be reformulated to make clear what is within the scope of an evaluation and what is outside it.

A similar point is raised on lines 369-371: "Policymakers also emphasized that they consider the availability of sufficient resources financial and human - within different regions to cover a larger population and if a feasible scale-up plan was in place." They are certain to, but this is not something that an evaluation can meaningfully inform on, though it can inform on the resources needed in the initial intervention. These arguments should be clarified.

3. The discussion of sustainability is mixed.

Two points clearly make sense: (a) "selection of appropriate indicators for a cost-effectiveness analysis" and (b) "contextual factors that are important in determining long-term continuation. This included whether there is local support and demand from communities and commitment from partner organizations (regional health facilities) involved in implementation to continue the training program". These are points that an evaluation could meaningfully inform on: cost-
effectiveness is an analytical question that can be addressed through an evaluation, and information on buy-in of local stakeholders such as health facilities is an area in which a third party evaluation could provide an external viewpoint, outside the normal interactions of the stakeholders already engaged in the intervention. (Though note that a judgement on 'continuation' would benefit from being linked to an assessment of whether the programme has achieved its goals - something that is not discussed here, it seemingly being assumed that continuation is desirable.)

Two of the points included under sustainability are less clearly a matter for an evaluation: (c) "whether costs involved in running the training program would be met by funders willing to continue investment" and (d) "level of political support to make sufficient resources available to continue the program". In my view, there are questions regarding these components: (c) Whether costs would be met by funders - this sounds very much like a concern from an interested party. Fundamentally, this is a political/policy decision and is outside the scope of the evaluation - while the evaluation can provide recommendations on the desirability of continuing an intervention, it can't ensure that funds are in place. (d) Given the stakeholders involved in this study, it seems unlikely that the evaluation team are going to be able to substantially inform them on the policy/political priorities for continuing training - the issue at stake in the example used in this study is continuation using domestic funding of a programme that was previously supported by foreign funding. Under normal circumstances, there will be little that outside evaluators can tell the national or provincial CDC about the policy/political priority attached to an issue they work on. (Though in the case of evaluations for an external donor, this point might well be entirely valid.)

'Scalability' deserves more discussion in the model. This is a departure from the other components of the model, which are primarily focused questions of internal validity, whereas scalability involves a discussion of external validity. This is both a trope in the Chinese policy context, and a genuine methodological question.

Terminology
'Internal' and 'external' evaluators (e.g. line 404). These terms are used in the text to refer to Chinese and non-Chinese evaluators. More normal use of the term 'external evaluator' is to describe someone external to a given programme (i.e. a third party). It is recommended that terms be changed to bring them in line with mainstream usage.

'Utilisation' / 'use', e.g. "Our qualitative analysis focused on factors that influence utilization of evidence in evaluation studies by policymakers' making resource allocation decisions about HCP training or capacity building interventions" (lines 455-457), "we found that perception of the quality of the research and research team is a major factor influencing the use of research results" (500-501). There has been no discussion of actual use or utilisation of evidence from an evaluation in the article. The questions asked of policy makers are hypothetical ones, and any judgement of usefulness is necessarily prior to actual utilisation and resource allocation decisions, which are policy/political decisions, not just technical ones. It would be better to rephrase as "judged to be important for informing decision making" or similar.

The term 'policymaker' is used throughout the article. Given that we are not told about the positions of individuals within the CDC or the CMA, the departments they are in, and their actual role in policy making, it might be better to either supplement the discussion with such information or to opt for a more neutral term, such as 'officials'.

Miscellaneous

"extreme shortage of skilled healthcare providers (HCP) in LMICs" (line 93). This does not apply to China.

"The guide clearly recognizes limitations in evaluations based on the Kirkpatrick model -which has not been updated since it was developed over three decades ago" (461-462). This is not correct - there have been many adaptations to the Kirkpatrick model, due to the fact that it is widely used.
Level of interest
Please indicate how interesting you found the manuscript:
An article whose findings are important to those with closely related research interests

Quality of written English
Please indicate the quality of language in the manuscript:
Acceptable

Declaration of competing interests
Please complete a declaration of competing interests, considering the following questions:

1. Have you in the past five years received reimbursements, fees, funding, or salary from an organization that may in any way gain or lose financially from the publication of this manuscript, either now or in the future?
2. Do you hold any stocks or shares in an organization that may in any way gain or lose financially from the publication of this manuscript, either now or in the future?
3. Do you hold or are you currently applying for any patents relating to the content of the manuscript?
4. Have you received reimbursements, fees, funding, or salary from an organization that holds or has applied for patents relating to the content of the manuscript?
5. Do you have any other financial competing interests?
6. Do you have any non-financial competing interests in relation to this paper?

If you can answer no to all of the above, write 'I declare that I have no competing interests' below. If your reply is yes to any, please give details below.

I declare that I have no competing interests

I agree to the open peer review policy of the journal. I understand that my name will be included on my report to the authors and, if the manuscript is accepted for publication, my named report including any attachments I upload will be posted on the website along with the authors' responses. I agree for my report to be made available under an Open Access Creative Commons CC-BY license (http://creativecommons.org/licenses/by/4.0/). I understand that any comments
which I do not wish to be included in my named report can be included as confidential comments to the editors, which will not be published.

I agree to the open peer review policy of the journal