Author’s response to reviews

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?

Authors: 
Shishi Wu (shishi_wu@u.nus.edu) 
Helena Legido-Quigley (Helena.legido-quinley@lshtm.ac.uk) 
Julia Spencer (spencer-julia@outlook.com) 
Richard Coker (Richard.Coker@lshtm.ac.uk) 
Mishal Khan (Mishal.Khan@lshtm.ac.uk)

Version: 1 Date: 03 Jan 2018

Author’s response to reviews:

03 January 2018

Dear Dr. Gonzalez-Mcquire,

Revision Note: HRPS-D-17-00053 Designing evaluation studies to optimally inform policy: what factors do policymakers in China consider when making resource allocation decisions on healthcare worker training programs? Shishi Wu; Helena Legido-Quigley; Julia Spencer; Richard James Coker; Mishal Khan Health Research Policy and Systems

Thank you for considering our above-named manuscript for publication by Health Research Policy and Systems. We are grateful for the reviewers’ comments which we received on 06 November 2017. The text below details our responses:
[Please include a point-by-point response within the 'Response to Reviewers' box in the submission system and highlight (with 'tracked changes'/coloured/underlines/highlighted text) all changes made when revising the manuscript. Please ensure you describe additional experiments that were carried out and include a detailed rebuttal of any criticisms or requested revisions that you disagreed with. Please also ensure that your revised manuscript conforms to the journal style, which can be found in the Submission Guidelines on the journal homepage.]

We are pleased to say that we have been able to address all of the reviewers’ points and this has greatly strengthened the manuscript.

[The due date for submitting the revised version of your article is 04 Feb 2018. This deadline may be extended upon specific request.]

We have submitted within the deadlines on 03 January 2018.

[Please note, if your manuscript is accepted you will not be able to make any changes to the authors, or order of authors, of your manuscript once the editor has accepted your manuscript for publication. If you wish to make any changes to authorship before you resubmit your revisions, please reply to this email and ask for a 'Request for change in authorship' form which should be completed by all authors (including those to be removed) and returned to this email address. Please ensure that any changes in authorship fulfil the criteria for authorship as outlined in BioMed Central's editorial policies (http://www.biomedcentral.com/about/editorialpolicies#authorship).]
We have requested a change of authorship.

Reviewer 1

I

[(1) Having read this paper, I feel that the authors put the Kirkpatrick Model at the center and much emphasize its weaknesses. In my opinion, this Model was developed for one particular purpose: determining whether training is effective, by focusing on the four outcome components. If an evaluation aims to inform future resource allocation to the training program, additional elements should definitely be assessed. Given the complexity of policymaking, evaluators of any programs should not expect that studies based on one single theory or model can offer adequate information for policy decisions. Focusing on the Kirkpatrick Model, the authors may review key features of this Model and also find out how the Model is usually applied in training evaluations and why it remains popular, even though its limitations have been identified. A manual for training evaluation prepared by JICA under a project to improve local administration in Cambodia is a good example, as it suggests evaluators identify the purposes of evaluations and then identify relevant methods and tools accordingly. It is unfair to argue that the Kirkpatrick Model does not capture everything policymakers wanted to consider. In particular, the features of evaluation studies, i.e. the robustness and composition of evaluation team should not be part of the extended Kirkpatrick Model. Is the development of the extended Kirkpatrick Model an objective of this study?]

We thank the reviewer for raising this important point. In line with your feedback, and that of the other reviewers, we have reframed the paper (please see changes to the introduction and results section).

Specifically, we are no longer presenting a modified Kirkpatrick model but a framework for the translation of program evaluation evidence into policy (that incorporates and acknowledges the role of Kirkpatrick in direct programmatic elements but moves beyond this model to consider other relevant factors).
[(2) I concur with the authors that policymakers in other LMICs may have different issues of concerns when they make decisions on HCP training programs (line 506-507). However, discussion is needed on whether the study findings can be generalized among different groups of Chinese policymakers, for example, those who make decisions in training programs for non-infectious control and other public health problems. Furthermore, as China is a large country, consisting of many administrative territories, I wondered if the 30 participants (13 from provincial level) in the interviews and FGD sessions could really represent Chinese policymakers, despite consistent viewpoints found in this study.]

We agree with your useful suggestion and have added discussion on the generalisability of our findings in the subsection on ‘Strengths and limitations of proposed framework and the study methodology’.

3

[(3) The Data Collection section (page 7) should be expanded. It is unclear what interview questions were employed and how they were developed (and tested?). Some questions on the limitations of current training evaluation approaches might lead to certain answers. The development of the FGD exercise should also be described. Although this study involves 2 major issues, namely training evaluation (research) and research-policy nexus, its methodology can be strengthened by reviewing public policy models and literature to get understanding on how a policy is developed, especially the concerns of those who made decisions. The review results may be helpful in devising the interview guide, FGD exercise and data analysis framework. From this, the authors may find that there are different consequences and impacts of the training programs, considered by policymakers. Only some of these factors can be assessed as part of training evaluations, while others require different assessments.]

We have addressed both of the reviewers’ suggestions as follows:
1. Additional details have been added to the data collection/methodology section about IDIs and FGDs (including a reference to theory that guided FGD development)

2. Text has been added in the background section engaging with the interactive model of policy process

4

[(4) Please the authors elaborate the statement: "the Kirkpatrick Model has not been updated since it was developed over three decades ago …" (line 462-463). Given that this paper refers to the 3rd edition of the book on the Model (reference #21), some sections might have been rewritten, based on feedback from the readers and current contextual situation.]

Thank you for your comment, with which we fully agree. Owing to further investigation and a reframing of the paper, this statement has now been removed.

Reviewer 2

5

[Broadly, the paper addresses the issue of translating program evaluation to policy by the policymakers. It is focused narrowly on (a) healthcare worker training programs, (b) resource allocation decisions, and (c) informing the policymakers for the decisions, and not the ultimate policymaking or policy. The study is conducted in China with a small group of local policymakers. Its results extend and expand the Kirkpatrick model.

The paper is insightful. The method is thorough. To be able to generalize their findings, and fulfil the expectations raised by the first part of the paper's title, I would suggest the following: (a) frame the paper within a broader model of translation of program evaluation to policy, and (b) reorganize the final model slightly. I will explain the two below.
The current paper is more an extension of the Kirkpatrick model and less of "[d]esigning evaluation studies to optimally inform [and formulate] policy." The latter is the desired outcome. By framing the study within a general model for translating program evaluation to policy, it would be easier to generalize the findings of the study beyond (a) healthcare worker training programs, (b) China, and (c) resource allocation decisions.

The authors present their final result as an extension of the Kirkpatrick model (with the added broader programmatic results), and expansion with two other components -- broader implementation considerations, and perceived reliability of the results. To this reviewer, the final model may be summarized as Kirkpatrick model focused on the specific program elements + broader programmatic elements + evaluation credibility. The broader programmatic results would be part of the broader programmatic component. The broader programmatic component would contextualize the results of the specific program component. The evaluation credibility could amplify or attenuate the willingness to formulate policy based on the program evaluation information.

In short, the message of the paper and the general underlying model appears to be:

Translation of program evaluation to policy = f(program evaluation, program contextualization, evaluation credibility). The Kirkpatrick model focuses primarily in the first component, the authors have added two other components to it.]

We are grateful for the positive comments on the insights and thorough methodology of the paper. We found the reviewer’s suggestions for improvement very useful and have followed them by a) reframing the paper and b) modifying the discussion about the final model to focus on the translation of program evaluation to policy. As you suggested, the Kirkpatrick model is incorporated into our framework to represent specific program elements, but we add two new components, broader programmatic elements and evaluation credibility. Thus, rather than an
extension of the Kirkpatrick model, we create a framework which can aid in the design of policy-
relevant evaluation studies.

Reviewer 3

[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).]
Thank you for noting that we address an important, under-researched topic and for your positive comments on the methodology section.

Based on your suggestion we have added information about each respondent (national policymaker or hospital director) without giving away his/her identity.

[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.]

Thank you for raising this important point about the high level of technical skills/analytical awareness of the interviewees. We have emphasised the technical and analytical sophistication of the interviewees in the Evaluation methodology section of the manuscript and also reflected upon this in our discussion section as suggested.

[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.

Thank you for appreciating our conceptualisation and suggesting specific improvements which we have made. Based on your suggestion, we have reframed the article so as to present a broader framework for the translation of program evaluation to policy. Rather than expanding on the Kirkpatrick model, we suggest that this model on its own may not be sufficient, and highlight other factors that should be considered in an evaluation study to better meet the information needs of policymakers. These other factors could be captured through a combination of methodologies, as you suggested.
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. ]

Many thanks for the comment, we agree with the reviewer that we are requesting that assessment of a training programme also explore its wider impacts. Unfortunately, in our experience this is not often taken into account or being considered.

[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.]

We agree with the reviewer’s comment and we thank her/him for highlighting this point. We have clarified according to the reviewer’s suggestion by adding the following text: Here a useful evaluation could present information about costs and resources required to continue or scale up the training program, but evaluators may not be in a position to assess future funding commitments.

We have also addressed the related comment about the scalability section.

11

[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.)]
Many thanks for your positive feedback. We have amended the text to highlight your point that a judgement of continuation would be linked to an assessment of whether program goals have been met.

12

[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.)]

Many thanks for these points. We have edited the text based on your feedback to make it clear that some components of sustainability that are considered important by domestic policymakers may be outside the scope of typical evaluations.

13
‘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.

We agree the findings around scalability (and policymaker attention to this) are important and we value the feedback about discussing these in more detail. Accordingly, we have done the following:

1. In the results section discussing scalability we have included a point about some assessments of scalability being outside the scope of evaluations

2. Added the following sentence to the discussion: Furthermore, our findings indicate that policymakers are not only concerned with the internal validity of the evaluation results, but also external validity in terms of whether the evaluation results demonstrate scalability.

Many thanks for your comment. As per your suggestion, we have amended the term “internal” to “local (Chinese)” and “external” to “foreign (non-Chinese)” evaluators throughout the text to better reflect 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.]

Thank you for raising this important point. As per your helpful suggestion, we have rephrased discussion of ‘use’ or ‘utilisation’ of evidence by policymakers throughout the paper to ‘factors that are judged to be important for informing decision-making’.

16

[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'. ]

We agree with the reviewer’s comment and we thank her/him for highlighting this point. We have amended the text. When referring specifically to interview respondents we use the term ‘officials’ instead of ‘policymakers’. We have also added information about the positions of respondents when quoting.
"extreme shortage of skilled healthcare providers (HCP) in LMICs" (line 93). This does not apply to China.

We have amended language from ‘extreme’ to ‘acute’ and added the following reference, which evidences the shortage of HCPs in China. http://www.bmj.com/content/354/bmj.i4860.full

["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.]

Thank you for your important comment. Upon further review, we have removed this sentence.

[If improvements to the English language within your manuscript have been requested, you should have your manuscript reviewed by someone who is fluent in English. If you would like professional help in revising this manuscript, you can use any reputable English language editing service. We can recommend our affiliates Nature Research Editing Service (http://bit.ly/NRES-HS) and American Journal Experts (http://bit.ly/AJE-HS) for help with English usage. Please note that use of an editing service is neither a requirement]
Improvements have not been suggested by the reviewers. However, we have proof read the paper one more time and made minor edits.

We attach a revised copy of the manuscript, dated 03 January 2018, for your consideration.

We hope that these responses are to your satisfaction, and we hope to hear from you soon.

Yours sincerely,

Mishal Khan