Reviewer's report

Title: Instrumentation Issues in Implementation Science

Version: 1 Date: 2 December 2013

Reviewer: Julie Lowery

Reviewer's report:

This is a nice description of some of the key issues in implementation science. However, these issues are not unknown to implementation science researchers. Therefore, the meat of the manuscript is in the recommendations. With some enhancements (noted below), I think the manuscript could be highly beneficial to researchers.

Major compulsory revisions.

1. Implementation Issue #1. Under recommendations is the statement, “In sum, clear and careful definitions…will allow stakeholders to select appropriate instruments for the constructs under investigation.” Can the authors offer additional advice for how to come up with these clear and careful definitions for those constructs relevant to a given study? And given the many constructs that are similar or that overlap across frameworks, how does one select the construct name that best meets the definition or, more importantly, how does one find the right instrument? For example, I was intrigued with the findings from the SIRC IRP regarding the similar constructs of appropriateness, perceived fit, relevance, compatibility, suitability, usefulness, and practicability. The authors found that “unique instruments (i.e., with different item composition and content) are used to measure these different terms/constructs. Therefore, our preliminary results suggest that these constructs may actually represent nuanced, but unique factors potentially worthy of our consideration.” Would it be possible to use this finding to provide an example of how one would devise a clear definition around one of these constructs and then use it to find the best instrument? I’m very curious about the differences in item content across instruments measuring these very similar constructs, and how one would decide which one to use. This is a very practical problem on which the authors’ insights could be especially helpful.

2. Instrumentation issue #2. The recommendations for this issue are titled, “Recommendations for reliability reporting” and “Recommendations for validity reporting.” I would suggest that the issue is less about reporting (although of course this is very important) and more about conducting reliability and validity assessments. In particular, the authors note “the field’s nascent state and the complexity of the projects placing demands on researchers.” Given these challenges, can the authors offer some specific suggestions to implementation science researchers on how best to incorporate these assessments in their research? Should researchers report certain data/statistics, even if they clearly do not have sufficient sample size to fully investigate reliability or validity (which
is likely to be the case in many implementation studies), that can be used in subsequent meta-analyses for establishing validity and reliability of various instruments? For example, the authors write, “At a minimum, researchers should assess and report an item’s internal consistency.” Since psychometrics is not my area of expertise (and I imagine this is true for many of my colleagues), it would be worth adding one or two more statements about how to do this, if these are critical data to collect and report.

3. Also, which assessments of reliability or validity could be relatively easily incorporated into studies, without significantly increasing the required resources and timeframe? For example the authors write, “It is important to note that there are additional forms of reliability, including test-retest and inter-rater reliability, which vary with respect to their applicability to the study and should be reported wherever possible.” The qualifier, “which vary with respect to their applicability to the study,” is key. Can the authors provide examples of when these might be applicable, thus allowing their incorporation into studies?

4. Similarly with recommendations for validity reporting—the authors mention the different kinds of validity: construct, content, concurrent, divergent, criterion-referenced. Which ones of these could most easily be measured in implementation studies, and how would this be done? I think the issue that implementation researchers have is that their primary goal is to investigate the implementation of an intervention, not to establish the psychometric properties of an instrument. How can both goals be met at the same time?

5. Instrumentation issue #3. This is a very interesting and complex issue. In fact, it is so complex, I am not sure that it can be adequately addressed in this paper along with the other issues. If it is going to be included, then I think more specific suggestions need to be provided than are currently included in the recommendations, as explained below:

Who? What are the recommendations for this section?

When? “Careful attention must be paid to which constructs are assessed at each stage?” What does this mean exactly? How would this be done? “The typical pre-implementation/post-implementation evaluation of constructs has likely impeded the field of implementation science from identifying predictors, moderators, and mediators of the implementation process.” Why is this true? What is the problem with this type of evaluation? What are alternative designs, and what are the implications of these on study resources and timelines?

What? “Researchers are encouraged to consider these constructs to determine what to measure, when in the implementation process, and at what level...” Are there any specific recommendations to guide researchers in figuring out the answers to these questions?

Also, I don’t see this particular issue in Table 1. Where did it come from?

6. Instrumentation issue #5. Do the authors have any suggestions for addressing the self-report issue?
7. Implementation issue #6. No specific recommendations are provided under “Length.” Are there any recommendations or rules of thumb that can be provided from a statistical perspective? E.g., “To establish internal consistency of a particular scale/construct, a minimum of X items is generally needed” or “Y items are recommended.” No recommendations are provided under “Language.” Perhaps suggest piloting instruments with a few subjects using a “think aloud” technique?

Minor compulsory revisions

Please provide additional information on the design:

8. “The authors identified numerous instrumentation issues upon review of the preliminary results from the SIRC IRP. The authors then created a survey designed to engage implementation stakeholders…” Please provide a little more information on how the preliminary results informed development of the survey.

9. “Participants were identified via implementation science related listservs.” Examples of three different groups are provided. It would be interesting to see all groups that were used, to get a better idea of the diversity of respondents.

10. How many people received a request to complete the survey? What was the participation rate?

11. Table 1: Please explain how the average rank can exceed 10 (e.g., issues 12 and 13, respectively), if all items were ranked on a scale of 1 to 10.

Discretionary revisions

12. Consider eliminating Instrumentation issue #4. It overlaps quite a bit with issues #2 and #7, and could possibly be incorporated into these. As it is, it doesn’t seem to be a sufficiently distinct issue from these other two.

Level of interest: An article of importance in its field

Quality of written English: Acceptable

Statistical review: No, the manuscript does not need to be seen by a statistician.

Declaration of competing interests:

I declare that I have no competing interests.