Reviewer's report

Title: Advancing the Application, Quality and Harmonization of Implementation Science Measures

Version: 2 Date: 18 July 2012

Reviewer: Cara Lewis

Reviewer's report:

Re: ‘Advancing the Application, Quality and Harmonization of Implementation Science Measures'

The premise of the submitted manuscript, to identify, categorize, and make available an interactive repository of implementation science measures that could possibly lead to harmonization of implementation research efforts, is ambitious. Use of the wiki platform allows this team to access the collective wisdom developing in the field of D&I around the globe. The authors have proposed an evaluation system for rating the measures along two dimensions: psychometric quality and practicality. This project therefore has the potential to facilitate systematic and rigorous evaluation of implementation efforts. However, several weaknesses have emerged that, from my perspective as a reviewer working on a similar project, limit the potential for this manuscript to make a substantive contribution to the literature in its current form. I have attempted to provide constructive comments and suggestions that will be helpful in disseminating this information, even if not in Implementation Science.

Major Compulsory Revisions

1. The authors do not describe the focus of this project other than to state that it is focused on measures relevant to “implementation science” nor do they define the scope. Is the scope all fields or disciplines (e.g., education, mental health, business) or is it focused on health only, for instance? Are qualitative interviews considered measures or is this project restricted to quantitative instruments? This has implications for stakeholder involvement, organization of the wiki, and identification and inclusion of measures.

2. The authors do not indicate where the definitions for the constructs included on the wiki come from--are they provided by champions, the core research team, or stakeholders? Care in defining these constructs and categorizing the measures seems particularly important for their project especially because the authors themselves have highlighted issues of conflicting definitions as a main factor guiding their work (see paragraph 3 of the background). Given the lack of cohesion and consistency of taxonomies and construct definitions for D&I research, it is a concern that relying on stakeholders to (1) identify and define constructs and (2) categorize measures without a systematic or standardized way of reviewing this work could perpetuate this problem.
3. The authors indicate on page 2 that “...there is little or no information about practical, actionable measures that are appropriate for use in real world settings that have competing demands...” Do the authors mean that no one has before summarized this information across measures? If so, a simple clarification would suffice. This sentence in its current form is slightly confusing because data regarding several of the dimensions in their “Practical Measure Rating Criteria” are available along with having access to the measure itself (i.e., # items, cost). Or, do the authors mean that valid and reliable assessment of these dimensions (e.g., “Actionable”, “User Friendly”) has not yet been conducted? I view this as a separate and important issue. I would caution the authors to explicitly state the process by which they intend to rate these dimensions as some methods would be more rigorous than others and it is not yet clear from this manuscript how they intend to proceed. Specifically, experts in the field of D&I have indicated that one of the only valid and reliable ways of rating these dimensions would be to systematically code for all relevant information available for each measure--albeit an incredibly time-consuming process. However, this method would avoid reification of opinions that may not be terribly representative (especially given the user breakdown provided in the manuscript). Publicizing ratings on these dimensions without knowing the reliability of the rating system (or the source/rater’s area of expertise) could be seen as problematic. This is not to say that having opinions of stakeholders made available to others is a useless endeavor; on the contrary, with transparency this could be very helpful and important to a number of stakeholders. Rather, I would simply caution the authors to be very transparent about this method (and all steps and stakeholders involved in this process) and appropriately note the limitations of this approach in the manuscript.

4. There is a particularly confusing section of the manuscript that would benefit from additional attention and clarification. The first sentence of the second paragraph on page 7 is confusing. What do the authors mean that “…this method has been shown to create more accurate outcomes such as those created by GEM” as it relates to their current project? In this same paragraph the authors use the term “standardization” (as they do in several places throughout the manuscript) referring to D&I measurement and it is not entirely clear what they mean by this term in this particular context. Finally, in this section the authors also discuss the merging of independent data as an outcome of GEM yet this is not included in their purposes for the project nor do they talk about how this would take place; this might better be referred to clearly as a future direction rather than a natural outcome of the wiki platform as this would be a large additional effort to orchestrate beyond this initial product.

5. Can the authors say more about their “focused, non-systematic search of the literature for additional measures”? What was the focus? Why was this the approach taken? What benefits did this non-systematic review have over a systematic search? I am concerned with the lack of methodological rigor that seems to guide this project and I am even more concerned about the authors sharing results stated conclusively such as: Acceptability has only 7 associated
measures. Because of my involvement leading a similar project, I can say that there are at least 43 measures resulting from a systematic review of Acceptability that are best categorized as Acceptability of the (a) intervention or innovation and (b) the training. The authors state the “website has been sufficiently populated”, which I would argue with given the example provided. Because of the stage of the current project and the approach, I am concerned that publishing this manuscript in its current form would paint an inaccurate and confusing picture for stakeholders. Careful attention to revising the way in which the approach and results are presented would be important.

6. The authors indicate that the measures will be rated in terms of reliability but they don’t indicate how the data required to complete these ratings will be collected and provided to raters. That is, are they systematically reviewing the literature to collect the associated psychometric papers or are they relying on stakeholders to post to the web and then assuming that all the necessary information is there? And, if they are intent on rating the reliability of measures (thereby implicitly indicating they value this) then I would like to see how they plan to achieve reliability in their own ratings.

7. The authors indicate that “Users have been asked to enter both quantitative ratings and qualitative comments at anytime during the campaign if they are familiar with the measure.” What or who determines familiarity with the measure? Who ensures accuracy of the data? What determines the “end of the Populate phase”? What does the phrase “sufficiently complete” mean on page 11? What defines a “champion”? If champions are relied upon to determine whether “any key IS measures or metadata were missing” then the identified champions may not be sufficiently trained or well-versed in these areas, as I previously noted, the project website to date is nowhere near populated to reflect a comprehensive review. Defining these terms and phrases in the manuscript will be important for stakeholders to better understand the context for this project.

8. The authors indicate that they will use a 5-point rating scale to rate each measure along several dimensions across two scales. They indicate that the anchors for the Likert-scale are qualitative phrases such as “least favorable”. Given the degree of subjectivity implicit in the dimensions of their rating scales, it is unlikely that these qualitative anchors will be reliably applied. Do the authors have plans to have each measure rated at least twice to evaluate the reliability? From their brief description, it is my understanding that they will simply “take an average”. The way the project is described suggests that minimal oversight and no training in conducting the ratings is provided, is that assumption correct? If not, please clarify in the manuscript.

9. With respect to the specific rating criteria, I have the following questions and concerns:

a. Why did the authors create their own “Gold Standard” quality criteria as opposed to accessing or modifying one of the many relevant, published rating scales that exist? See Hunsley & Mash (2010) or Terwee et al (2007) for just two examples.
b. Why is test-retest reliability “emphasized” over internal consistency? There are actually fewer cases in D&I where test-retest reliability is relevant whereas internal consistency could be argued is always relevant. Please explain.

c. How will raters know that construct or criterion validity is “most favorable” for a given measure when no additional anchors are provided?

d. How do the authors plan to deal with double-barreled anchors such as those asking raters to rate construct and criterion validity simultaneously?

e. How will raters know if “sensitivity to change” is “applicable”?

f. How will raters be instructed to deal with the subjectivity inherent in the practical rating criteria (e.g., feasible, importance, actionable, user friendly, enhances engagement, no harm)?

i. The authors indicated that they could not always access the measure itself making it difficult for stakeholders to rate this information. Will it be the case that only those who have actually used each particular measure will be invited to provide ratings on these dimensions?

g. Some people have a strong investment in having their measures used. How will you ensure that ratings are coming from a person free of conflicting interests?

h. In Phase 4, how will you weight the two scales and all the dimensions to determine the “top” measures?

Minor Essential Revisions

1. The authors need to add page numbers.

2. The authors are very clear and explicit with their timeline, yet they state that personal communication with Cara Lewis (this reviewer) occurred June 12, 2012. To be clear, I would ask the authors state the correct date of our communication which was March 13, 2012.

3. Figures 3 and 4 are missing a title.

4. It would be optimal if the authors were consistent with their use of terms by either selecting “measures” or “instruments” to make clear that they are referring to the same thing each time. On page 10, the authors state, “upload or add instruments to key measures” did they mean to say, “upload or add measures to key constructs”? On a related note, the example they provide on that page in parentheses is unclear; please clarify.

5. The closing parenthesis is missing on the bottom of page 16.

Discretionary Revisions

1. The abstract is structured according to the categories indicated in the instructions to authors, but the content could be reorganized to better fit the recommended topics per category. For instance, it might be optimal to include the purpose in the background section and more discussion of method in the method section.
2. The authors indicate in the 3rd paragraph of the Background that “other constructs do not have measures” which may indeed be true, but unless a systematic review has been conducted I might encourage tempering this language.

3. The authors’ main outcomes reported in this paper have to do with usage and views on the website. They interpret greater number of views on pages to be a “good” sign which makes sense. However, from a user’s perspective, I visited the website numerous times during this review to try to orient myself with the initiative and I found it difficult to navigate to the particular D&I measures and therefore visited many pages that I did not mean to in search of D&I constructs. Is it the case that the wiki does not separate the D&I constructs out from the other constructs they have posted? Because the authors do not use a framework or theory to organize their D&I constructs, searching their website is challenging, particularly for those who are less familiar with D&I.

4. The authors present data on the geographic regions from which GEM D&I followers/contributors came as well as some data regarding their affiliation. It would be helpful and interesting to see the full breakdown of both categories in a pie chart format, for instance.

5. The authors refer to a “factsheet” in the second paragraph on page 9 that might be helpful to include as an additional file for readers.

6. It might be helpful to address the minimal number of comments (18 total) posted on the wiki to date (and shared in the additional file) given this has been a focus of the ratings you intend to serve as a decision tool.

7. The last sentence on page 16 is confusing because the authors state that 92% of the added measures are ready to be rated but only 48% of measures contain psychometric data. Can you please clarify what seems to be a discrepancy?

8. The authors indicate at the top of page 17 that “Some IS constructs (e.g., fidelity, dissemination) would also benefit from additional measures, and will be the focus of future efforts.” The way this sentence reads suggests that measures for these constructs do not exist when in fact several are available. I would recommend rewording this sentence to avoid sending the wrong message.

Minor issues not for publication

1. The authors use “etc.” at least three times; this should be avoided.

Level of interest: An article whose findings are important to those with closely related research interests

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.