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

Authors:

Adrienn B Rabin (borsika.a.rabin@gmail.com)
Peyton Purcell (pcurcellp@mail.nih.gov)
Sana Naveed (sana.naveed@nih.gov)
Richard P Moser (moserr@mail.nih.gov)
Michelle D Henton (michelle.d.henton@kp.org)
Enola K Proctor (ekp@wustl.edu)
Ross C Brownson (rbrownson@wustl.edu)
Russell E Glasgow (glasgowre@mail.nih.gov)

Version: 3 Date: 3 October 2012

Author’s response to reviews: see over
October 3, 2012

Dr. Gregory Aarons

Editor

Implementation Science

RE: Revision of manuscript 1359409465748507 “Advancing the Application, Quality and Harmonization of Implementation Science Measures”

Enclosed is a revised manuscript for your consideration in Implementation Science. In the attachment, we have provided a detailed response to each review comment. We have revised the manuscript based on the helpful review comments and indicated changes using track changes.

All authors of this research paper have directly participated in the planning, execution, or analysis of the study. All authors of this paper have read and approved the final version submitted. The contents of this manuscript have not been copyrighted or published previously. The contents of this manuscript are not now under consideration for publication elsewhere. The contents of this manuscript will not be copyrighted, submitted, or published elsewhere while acceptance by the Journal is under consideration. There are no directly related manuscripts or abstracts, published or unpublished, by any author(s) of this paper.

We appreciate your offer to help the manuscript move through the review process in a timely manner. We look forward to hearing from you. Thank you for your assistance.

Sincerely,

Borsika A. Rabin, M.P.H., Ph.D.
Staff Researcher
CRN Cancer Communication Research Center
Institute for Health Research
Kaiser Permanente Colorado
Phone: (303) 614-1295
Regular mail: P.O. Box 378066, Denver CO 80237 USA
E-mail: borsika.a.rabin@gmail.com

Russ Glasgow, PhD
Deputy Director, Implementation Science
Division of Cancer Control and Population Sciences, National Cancer Institute
Responses to Reviewers’ comments:

Comments from Editor:

I thought that you over-used inverted commas unnecessarily (e.g. “spread the word” could be replaced by shared or disseminated). Please review and reconsider frequency and impact of their use.

We have reviewed the manuscript and decreased the use of inverted commas.

You also refer to a ‘seminal paper by Proctor;' this could come across as rather self-serving given Enola Proctor’s co-author status and it might be premature to describe a paper published in 2011 in such terms.

We revised the language and replaced the term ‘seminal’ with the term ‘recently published’ (p. 8).

The current manuscript length is just over 5000 words. I thought that a lot of the writing could have been much tighter and less descriptive. I’d particularly suggest that you tighten the Discussion and revise the Conclusion (which currently introduces new material instead of summarizing your key points). If you respond to Reviewer 1 and make any changes, I’d still expect any revised manuscript to come in at less than 5000 words.

We have condensed the manuscript as suggested while addressing the majority of reviewer recommendations below. The current length is 5197 words. We felt that further omissions from the manuscript would take away from the content of the manuscript and we are asking the Editor to consider accepting this length.

Reviewer: Cara Lewis

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.

The primary focus of our project is on health (including public health, health care, and mental health) relevant measures and constructs as reflected by most of the current constructs and measures in the GEM database. However, since GEM has a participatory approach, we are relying on stakeholders to identify areas that seem relevant to D&I. Hence, inclusion of measures and constructs from other areas and disciplines might occur in the future and will not be censored from the website.

The GEM database is mainly set up for quantitative measures, at this time all of our measures are from the quantitative domain. However, as mentioned above, our participatory approach heavily relies on stakeholder input and evolution overtime, so we do not rule out the possibility that qualitative instruments will also be added in the future.
To clarify the primary purpose of the paper we added a sentence to the purpose on p. 5.

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.

*Thank you for raising this issue. The definitions for constructs provided in GEM are entered and edited by our core research team (in the case of the pre-populated constructs) and by champions (for later additions) using definitions from referenced publications. Classification of the measures under constructs are decided based on literature and the IS expertise of the research team and champions/users. We agree with Dr. Lewis that the wiki approach described in this paper is certainly new and different from traditional, small group expert approach of developing shared definitions and making recommendations on measurement and other topics. We believe that this approach comes with its own advantages and disadvantages. To clarify this, we included a note about how definitions for constructs were created and how measures were classified (pp. 9,10 and 11) and also expanded our limitations sections to acknowledge the possible shortcomings of this approach (p. 19). The advantages of using a wiki approach are already discussed in our paper on pages 7 and 8 and the references are provided.

We do think that many of the remaining comments are related to this basic issue however- and feel it important to re-iterate that wiki and crowd sourcing approaches are different than traditional expert, small group based approaches and must be judged accordingly- not solely on the basis of traditional criteria.

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.

*We referred to the former explanation, where as there is not enough information summarized on these measures to date. We clarified this in the manuscript (p. 5.)*
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.

Thank you for this comment. We believe that the reviewer’s comment would make a lot of sense in the context of a systematic, expert-based approach for measure harmonization. However, our proposed wiki-based collaborative approach does not seek to meet the same a priori criteria as the traditional, expert-based approach does. Our intention is to invite and engage stakeholders to guide this process by contributing with the information and expertise they have. We have discussed this in an earlier response in more detail and also clarified the difference between these two methods in the manuscript. (see especially p. 7). This said, we have made every effort to be transparent about the process involved and what is being rated, by whom and how both here and throughout the issues below and revised the methods section accordingly.

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?

Some text was added to this paragraph to explain how the environment created by GEM and the use of diverse stakeholders should create the most accurate outcomes, in this case, vetting and choosing the best measures to use (p.6).

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.
A definition of ‘standardization’ was added to information on pages 6 and 7. The data sharing feature of GEM has not been used in this campaign and is only described in the manuscript as an existing feature of GEM that could facilitate future collaborative science efforts. Additional text was added at the end of the paragraph to make the point that this ability to merge independent data is being developed for GEM.

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.

As mentioned above, our goals and methods were admittedly different from the goals of a systematic review. The purpose of the pre-population of the GEM website was to create a starting platform for discussion not to provide a comprehensive list of measures. The wiki approach then builds on a collaborative, participatory process to enrich the existing repository. Due to this methodological difference, systematic approach for the identification for initial constructs and measures was not necessary/feasible and comprehensiveness was not the goal.

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.

Again, this reflects the meta-issue that wiki approaches are expert review- they are continually evolving and never ‘completed or done’. It sounds like the reviewer is saying only approaches that reflect her perspective on what the results should look like should be published.

That said, we have made repeated efforts to clarify all of the issues she raises (given word limits stated by the editor).

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.
As mentioned earlier, the wiki process builds on input from stakeholders rather than information provided by a core group of experts. Ratings are asked to be completed for measures where stakeholders feel enough information is available and meet a set of criteria specified by the developers of the GEM website—not our author group. They are also encouraged to edit measure profiles and add additional information on any characteristics (including reliability) when appropriate. Also, our team does not attempt to achieve reliability or to have users rate ‘reliability’ separately. We have reworded this section to clarify this issue: Reliability—like sensitivity to change, breadth of applicability is one of the criteria to consider, but we do NOT ask users to explicitly provide a rating of reliability. Reliability data, when available, are taken directly from the literature in ratings. Each rating will be posted with the raters name to increase transparency.

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 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. Familiarity with measures is determined by the user who can elect to comment on and rate measures or not.

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.
We have corrected the anchors listed as reflective of those on the site (p.12). We recognize that these are subjective, qualitative anchors and in a wiki tool there is no guarantee that they will be reliably applied by all users and that may be a limitation of this approach. However, based on the experience of other fields that use a crowd-sourcing approach, there can be important knowledge and information gleaned from this type of feedback which can be used as an important point for discussion in collaboration with more rigorous evaluation methods.

From their brief description, it is my understanding that they will simply “take an average”. Information on the “Gold standard” criteria and practicality will be reported separately for the measures. Measures with high ratings for either of these scales will be separately listed. This has been clarified in the text (p. 13).

Do the authors have plans to have each measure rated at least twice to evaluate the reliability? 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.

Yes, our approach will have little to no oversight and will follow a less controlled approach. The hope and intent through publication of this article and other announcements is that each measure will be rated many times - by many different raters - not just once or twice.

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.

In our use of “Gold Standard” quality criteria, instead of thinking of separate dimensions listed, we were intending direct the raters focus to the multiple facades of this scale and differentiate it from the less commonly used practicality scale. As such, “Gold Standard” is used as one concept rather than a set of separate items.

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.

For pragmatic ratings, test-retest and sensitivity to change are emphasized over internal consistency because use for tracking progress and longitudinal, repeat measurement are critical for pragmatic measures. Internal consistency is achieved often though both separate unidimensional constructs and especially by inclusion of a large number of survey items - which are neither necessary nor appropriate for pragmatic measures.

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)?

*The remaining questions in this section (c through f) are essentially re-iterations of the basic and meta-issued discussed above and the answers are essentially the same as provided above.*

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?

*No, any user can decide to rate measures when they use the GEM website. As noted earlier, users will determine if they feel sufficiently familiar with a measure to rate it.*

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?

*The proposed wiki approach favors stakeholder control over central oversight. As such, we allow stakeholders to decide whether they feel comfortable rating a certain measure. The ratings and comments are posted with the raters’ name to increase transparency.*

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

*As discussed above, information on the “Gold standard” criteria and practicality will be reported separately for the measures. Measures with high ratings for either of these scales will be separately listed.*

**Minor Essential Revisions**

1. The authors need to add page numbers.

*Thank you for catching this. Page numbers were added to the manuscript.*

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.

*We were referring to a follow-up email exchange with a June 12, 2012 date but per Dr. Lewis’s request we modified the date of the personal communication to March 13, 2012 in the manuscript (p. 19).*

3. Figures 3 and 4 are missing a title.
As indicated in the submission process for the Implementation Science, figure titles were left off of the figure files and instead we have included a summary of all figures and tables, with title and legend.

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.

We have tried to make the language consistent in our revised manuscript and have used the term measures. Furthermore, we clarified that, in the context of GEM, the term instrument refers to the actual set of items/scales that can be used to assess a given measure. Instruments can be uploaded to each measure. (p. 19).

5. The closing parenthesis is missing on the bottom of page 16. This has been fixed. Thank you.

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.

We believe the current structure helps readers better understand the flow of the process and we decided to keep the structure as is. We did change/remove some of the headings to improve the flow of the manuscript.

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.

We modified this to “easily accessible measures (p.4).”

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.
We agree that in some cases the navigation of GEM can be challenging for the first time user. As indicated in the manuscript, developers of GEM have been conducting ongoing usability testing of the webportal and made multiple changes to facilitate navigation. Furthermore, our team has developed and are distributing videos and specific “how to” documents to possible stakeholders via email and also on the website.

In terms of the use of a framework, as indicated in our manuscript, we began with a framework proposed by Proctor and colleagues, but unlike a single author or small group approach using a single theory (which admittedly has certain advantages of simplicity), we purposively did not restrict GEM D&I to a single theory approach. This was especially important since we believe that a federal government resource should not be limited to a single theory or framework.

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.

At this time, we feel that a figure would not add to the text information already provided which provides the breakdown of organizational affiliation. In regard to the international contribution, because the overall sample size is not that large, the representation from other countries (a total of 10% across 7 countries) would be only 1 or 2 from each country and therefore would not seem to be a useful chart for the reader.

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. Thank you for the suggestion. We have now included the factsheet as well as an instruction guide for adding/rating measures which we hope will be useful to readers interested in learning more about the site and contributing to the database (see Additional files 1 and 2).

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. Since the website is publically available (and one goal of this article is to encourage interested parties to access the website) we did not summarize these comments in the article. We have added language to state that since the rating phase had not yet officially begun, we anticipate that additional comments will be added as further promotion is conducted (p. 16).

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?
We have revised the language to explain this discrepancy (p. 18). Measures are automatically categorized as ready for rating if they have a certain set of measure characteristics available. This does not include psychometric properties. The criteria was decided for the larger GEM initiative and applies automatically to all workspaces such as GEM D&I.

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. 
*We modified this as “will also require the addition of measures into GEM (p.18).”*

**Minor issues not for publication**
1. The authors use “etc.” at least three times; this should be avoided.
*We removed these from the text.*

**Reviewer:** Jonathan Tobin

**Major Compulsory Revisions**
NONE

**Minor Essential Revisions**
1. GEM’s reach to date, not surprisingly, has been primarily academic researchers. Please comment on strategies to increase reach among private and non-profit organizations which represented only 14% of the early adopters. This heterogeneous target audience will be critical to enhancing both research participation and impact. 
*Since the primary user group for GEM is researchers, our efforts to date primarily focused on researchers with D&I and practical interest. Outreach to the stakeholder community will be our next step and will be most likely achieved through contact via CTSA partners. We added this excellent point to the Next steps section of the revised manuscript (p. 21).*

**Discretionary Revisions**
1. Are there examples of well-described measures in both the gold standard and practical domains? If so, it would be helpful to establish the level of concordance between the gold standard and practical measures, which would strengthen the selection of practical measures without sacrificing scientific rigor (consider adding to Table 1 and might enhance acceptability in study sections). Are these “separate but equal” or inferior but implementable? 
*This is an excellent point. We know of sparse examples of this but decided not to include in the manuscript to meet length limitations. We will certainly consider discussing it using other forums.*
2. It would be helpful to know what specific interventions will be implemented to help populate sparse measures, and are there innovative approaches that utilize Web 2.0 methods?

*During our work we used NCI contractors to populate some of the measures using a targeted search effort with special focus on psychometric properties and clarified this in the manuscript (p. 12). The wiki approach itself should serve as an innovative Web 2.0 approach for further population of the metadata for the measures as well as the increase of number of measures and constructs.*

3. Are there metrics being considered (e.g., bibliographic/bibliometrics using PubMed; Web of Science;) to examine over time whether D&I research is converging on a more narrow set of measures, and if so, does this convergence correspond to better described GEM concepts/measures as compared to less well described concepts/measures. This might be a more robust measure of GEM’s impact on science.

*This is an excellent idea. In order to meet length limitations we did not address this in the manuscript but we initiated discussions with our collaborators to reflect on how impact of GEM could be measured in the longer run.*

4. A framework to the planned qualitative analysis of feedback would be helpful, and might stimulate discussion of how to implement this feedback.

*Again, while this is an excellent idea, length limitations prevented us from address this issue in the manuscript. However, we will consider addressing this using other, GEM related platforms such as the Discussion Forum and the R2R platforms.*

5. It would be helpful to provide further details to understand how the metadata are collected and coded. Metadata refers to the fields for each measure and is added by both users as well as the authors, as part of the pre-population, and there is no standard coding of the open text fields (pp. 6 and 20). We have tried to add some language to describe the process by which measures are assigned to a construct (one of the metadata fields), but, again, as a wiki approach, there is not always a systematic approach, but rather we rely on the collective group in making decisions and assignments (p. 9). For those fields that have defined categories, these were existing categories developed by the overall GEM database and whose origins are described in previous papers discussing GEM (referred in manuscript).

6. The practicality dimension opens up a number of D&I science possibilities, and the authors should comment on the plans to summarize the user feedback around collection, interpretation, feedback, and impact (e.g., are there any opportunities for users/viewers to share time-series data from practical measures adopted in practice-based settings?).

*Yet another great idea. We added a brief sentence into the discussion for the manuscript and will plan to further explore this issue using alternative platforms (p. 16).*

7. The authors present data from Google analytics and GEM user reports (Figure
4), which may represent measures of adoption and/or reach. Please comment on the suitability of using these measures (also, the details in the text about Twitter, etc. should be added to the x-axis). Are the y-axis measures simple or cumulative counts (cumulative counts indicate growth in science and may be more indicative of the developmental stage).

We have added language to indicate that the metrics and tools used to assess adoption and reach are commonly used for evaluating web-based resources (p.14). Our current figure uses cumulative counts - we clarified this on the figure and title. Since we had only one Twitter event to date, we decided to keep this component off the figure for the sake of simplicity.