Author’s response to reviews

Title: Instrumentation Issues in Implementation Science

Authors:

Ruben G Martinez (martinezrg@mymail.vcu.edu)
Cara C Lewis (lewiscc@indiana.edu)
Bryan J Weiner (weiner@email.unc.edu)

Version: 2  Date: 16 March 2014

Author’s response to reviews: see over
Dear Dr. Sales and Reviewers,

My colleagues and I are submitting a revised manuscript that attempts to address the points made in the first review. We greatly appreciate your feedback and believe that the responses put forth by the reviewers will serve to strengthen the structure and content of the paper, thereby increasing its utility and impact. We have carefully studied the comments and believe we have been able to address the issues raised by the reviewers.

The revised manuscript is sent with this document, which highlights our responses and comments made by the reviewers. Track changes are not presented throughout the manuscript given the great extent of changes made, making a tracked changes version difficult to read. Instead, we have used track changes comments to flag the sections that have been modified or added to address each of the numbered comments below. We hope that this streamlines the second review.

Thank you for this opportunity.

Editor Comments

I agree with the comments and recommendations made by both reviewers, and urge you to be thoughtful in your revision of the manuscript. In particular, the points about ensuring that your comments provide constructive feedback in addition to pointing out flaws in the literature--the issues around self-report are particularly important.

We agree with the editor and reviewers that the issues around self-report require additional discussion. In performing our revisions we have attempted to strengthen the recommendation portions of the manuscript as a whole, but particularly in the self-report section (see our response to Reviewer 1, Comment #8).

In addition, I fully support Dr. Bonds' concerns about how the survey you conducted is reported. While we suggested that you format this paper as a Debate paper rather than a research report, if you introduce information gathered through survey, you need to provide enough information for the reader to understand the value of the information. This could be done either in the text, although you will need to pay attention to the word count limit, or through an additional file.

We believe that there may have been some confusion regarding our upcoming manuscript submission regarding the Seattle Implementation Research
Collaborative Instrument Review Project (SIRC IRP) preliminary outcomes that had been the focus of some email correspondence with the editors. Specifically, we have a separate manuscript we intend to submit to Implementation Science delineating the SIRC IRP methodology and preliminary outcomes. We understand this might be particularly confusing given that we had referenced the SIRC IRP in the manuscript currently under review and provided some preliminary outcomes. In the revised manuscript, we have removed all SIRC IRP preliminary outcomes to avoid confusion.

Upon making our revisions we attempted to provide additional detail regarding the survey administered, including but not limited to information on content development, structure, and sample (full survey made available as an additional file). See also our response to Reviewer 1 Comment #6b.

I find the language used to initiate and motivate the paper somewhat extreme. While I think we would all agree that measurement is a critical issue, I'm not sure that the potentially nihilistic and somewhat adversarial language used to in the paper, which could be read as discrediting much of the research done to date, is fully warranted, and I would encourage you to consider mitigating it somewhat.

We apologize for the extreme nature of the language and have tempered the tone to reflect this concern. We did not intend the language to be adversarial, but bold in an attempt to take a stance on the issues presented and engage a debate; however, we realize that the same goals can be better accomplished through critical yet objective language.

Finally, I was somewhat disappointed in the very brief reference to the Chaudoir systematic review, which I had pointed out in our preliminary emails as an important comparator for the work you are doing; I would encourage you to continue to reflect on the added value of the work ongoing through the SIRC initiative.

We again feel that this comment may be a result of thinking the current manuscript under review was the manuscript that had been discussed over email (i.e., a SIRC IRP preliminary outcomes-focused paper). In removing reference to preliminary SIRC IRP outcomes, we hope that this is less of an issue. We have also tried to clarify the purpose of this manuscript (i.e., not to compare Chaudoir and SIRC), and highlight the Chaudoir findings throughout (in place of SIRC IRP outcomes).

And I am a little concerned about the lack of mention of more recent literature such as Carl May's paper earlier this year, which in its own way synthesizes the literature to date. I would encourage you to consider whether your reference literature needs some updating to be fully current.

After carefully reviewing Carl May’s recent debate piece, we agree it could be an excellent example of a unifying model that importantly addresses the social nature of an implementation effort. We have now included reference to his unifying theory.
in the discussion.

Please also ensure that your revised manuscript conforms to the journal style (http://www.implementationscience.com/info/instructions/). It is important that your files are correctly formatted.

We have checked that the manuscript itself and all additional files and confirmed that they the manuscript conforms to the journal style.

Reviewer #1:

1. (compulsory) An implicit assumption of this paper appears to be that the implementation field should have a single unified theory or conceptual framework and that researchers should share a core battery of implementation measures. The hypothesis that a single unified conceptual framework is appropriate across a wide range of practices was expressed in the Fixsen monograph (Fixsen, Naoom, Blase, Friedman, & Wallace, 2005) and is probably a widely shared view among implementation researchers, but it should be made explicit in this paper.

1a. As I have stated it, this is two-part proposition, both that there should be a unified theory and that many of the measures should be standardized and used in common. If the authors agree with this proposition, then they should state it explicitly at the beginning of the paper.

We appreciate your thoughtful comments. We believe that the creation and utilization of a unified theory has both its advantages and disadvantages, and we did not intend to suggest we are in support of adopting the Consolidated Framework for Implementation Research as a “unifying theory”. We have revised the introduction/background to make clearer the purpose of mentioning this framework given its relative comprehensiveness (see below for more details). We have also encouraged readers to consider Tabak et al.’s review of theories, models, and frameworks throughout. Finally, we have mentioned Carl May’s unified theory of implementation as a potential theoretical model that could advance the field.

1b. If I have misperceived their underlying assumptions, then the authors should make explicit their rationale justifying their enterprise of collecting a repository of implementation measures.

We apologize for the oversight in not making the SIRC IRP rationale clearer in the initial version of the manuscript. In light of this confusion (due to limited details of the SIRC IRP), we have decided to remove mention of this separate project and associated preliminary outcomes to better reflect the goals of this manuscript. That is, we did not intend for this manuscript to center on the SIRC IRP, its importance, or its contribution to the field, but rather to engage the field in dialogue regarding
critical instrumentation issues. These changes have been made throughout the manuscript.

1c. The authors endorse the CFIR, the Proctor implementation outcome framework, and the Powell implementation strategy framework as exemplary and complementary frameworks. My colleagues and I also hold these frameworks in high esteem and the strong endorsement in this paper may contribute to what appears to be a move toward consensus in the field, which would be a good thing.

We appreciate learning your perspective on the matter and agree that these models and frameworks have the potential for noteworthy impact on the advancement of the field.

2. (discretionary) There is little doubt in my mind that a number of implementation factors are common across most if not all implementation efforts. To take a simplistic example, adequate funding is almost universally acknowledged as a critical factor sustaining any practice. But is it really true that a common set of factors transcend all implementation projects? In our work, we distinguish among organizational-level, team-level, and practitioner-level interventions, each of which poses a somewhat different set of implementation challenges (McGovern, McHugo, Drake, Bond, & Merrens, 2013). The authors might clarify the range of interventions for which their measurement repository is intended, or if in fact it aims at a subset of the interventions just described.

That is an excellent point. We did not intend to suggest that all factors are relevant to all studies. We use the CFIR as an example because the CFIR contains a fairly comprehensive listing of variables implicated across implementation efforts and across levels within an implementation (consumer-, practitioner-, organizational-, policy-levels). We have clarified the use of the CFIR merely as an example for selecting constructs important to measure, but juxtaposed it to other leading, widely-used frameworks to address this comment.

3. (discretionary) Even if one assumes that a common theory or framework is appropriate for all implementation projects, there is a second question about the wisdom of using shared standardized measures across studies. Interestingly, one domain that the authors exclude from their project to collect implementation measures is the fidelity domain. They have made this decision because fidelity scales are “necessarily intervention-specific…” (p. 24). But if fidelity measures are specific to the intervention (which is clearly true), then why aren’t other implementation factors also specific? Are the attributes of good supervision identical across interventions? Is acceptability the same? Is penetration? I don’t know the answer, but I do believe that much is sacrificed by reaching for generic measures that by definition asks general questions rather than intervention-specific questions.

The reviewer’s questions are both thought-provoking and appreciated. We agree that evaluation of fidelity is critical. We have made brief mention of the possibility
that a consensus battery may emerge from the SIRC IRP or GEM projects, but simultaneously acknowledge the potential limitations of this approach.

3a. I also note that fidelity scales are the backbone of implementation research, that in my estimation, most important implementation studies have included fidelity measurement, and the absence of fidelity measurement in earlier research was a huge barrier to building a cumulative science (Brekke, 1988). Furthermore, many fidelity scales are exemplary in their grounding in psychometric, wide scale adoption, and standardization – i.e., some of the very problems the authors identify in the implementation measurement literature. By eliminating fidelity scales from their domain of consideration, they eliminate some of the best instruments promulgated in this field.

We agree with the reviewer that the assessment of fidelity in implementation efforts has been crucial in moving the field forward and its assessment is of paramount importance. Because the SIRC IRP data has now been removed, the focus fidelity is now limited and should address this concern.

4. (discretionary) In my estimation, generic implementation measures, including many very well-accepted ones, often ask such general questions that their precision and specificity is compromised. I worry about encouraging researchers to use generic instruments without modification; I would much rather see research tailor their instruments to their research questions if the study requires it. Certainly researchers often err in the opposite direction of mindlessly adopting instruments that supposedly measure some construct of interest but are far-fetched for their particular application (the so-called nominal fallacy). I did not see this set of issues discussed in the paper and recommend that the authors address this.

We apologize for seeming one-sided about the use of generic instruments. We did not intend to suggest that adaptation of instruments is entirely negative or inappropriate. Indeed, adaptation is not only necessary, but it is important in order to expand the use of instruments to different settings, interventions, or populations. We believe the issue arises when instruments are adapted and the psychometric property changes brought about by those adaptations are not carefully studied and reported. It is our hope that the statements made in the overview and recommendation portions of instrument issues #2 and #4 will suffice in expressing our stance that “home-grown” instruments and adaptation of existing instruments are necessary for the development of the field, but the lack of reporting of psychometric qualities and changes resulting from adaptations are potentially detrimental to building an evidence base. We apologize if the tone of the manuscript made it difficult to recognize the intentions of these recommendations.

5. (compulsory) The format of the paper is unconventional. It does not follow the usual structure of Introduction, Methods, Findings, and Results. As a conceptual paper I understand that it does not conform to a research article, but I recommend a more fluid orientation to the paper. The term “debate” introduced in the third paragraph comes out of the clear blue sky. Which debate are the authors referring to?
We had adopted the Implementation Science “Debate” paper format as we believe that the issues raised and recommendations provided are debatable. That is, our intention for this manuscript was to raise these issues and provide potential recommendations to engage the field in a necessary discussion. However, perhaps with the initial format for including the survey results as well as information regarding the SIRC IRP, this language/format is particularly confusing. We hope that with the removal of the SIRC IRP data and the inclusion of more survey development information that this comment has been sufficiently addressed.

6. (compulsory) The informal use of survey data from the SIRC conference is also unconventional and frustrating. It is ironic in a paper criticizing the lack of conceptual frameworks and use of unstandardized measures that the authors present a survey that is not formally introduced through a research study. I was left with the impression that this was a “homegrown” survey instrument (to use the authors’ label). It is possible that I am mistaken, but if so, the authors should make the derivation of this survey instrument clear in their paper.

We appreciate this feedback and have clarified the process through which the survey was developed in the revised manuscript. Also, given the tone in our previous version of the manuscript, we are aware that it appeared as though we did not value the need for “home-grown” instruments. However, upon tempering our tone, we hope the message is made more clearly that home-grown instruments are not inherently “bad”; rather, they are a necessary starting point for understudied areas. To address the reviewers’ concerns, we have added the survey in its entirety as an additional file as well as information regarding survey development (see editor point 4).

6a. The usual methodological questions raised about research quality should also be posed about their survey – for example, how representative is the survey sample?

We apologize for the oversight and have added information on the breakdown of areas/roles from the two listservs that we approached and obtained participation.

6b. Perhaps the most important question regards the choice of survey questions and what questions were not asked in the survey. How did the authors guard against respondent bias? Did the survey include distractor items? How many of the survey respondents were aware of the SIRC project to collect measures? How many attended workshops given by the study authors? It is relatively easy to get the answers you want by posing questions in a certain way.

Your feedback on this issue is appreciated, and we apologize for not putting adequate information about the survey in the initial submission. To be sure, we did not intend to get any sort of particular response when developing the survey. We were genuinely curious to see if other stakeholders perceived the same/similar issues to be present as did we so that the article reflected a broader perspective. We did
not wish to include distractor items in an effort to make the response process easier and more acceptable to respondents, and because we wanted the purpose and focus of the survey to be clear to respondents. With regard to respondent bias, we gave no indication, nor did we have preconceived notions about what instrumentation issues would be perceived as more important than any others. Many of the participants were aware of the SIRC IRP, which might have influenced their perceptions of what instrumentation issues are more important than others (e.g., psychometrics). The limited role of respondent bias will likely be clearer upon review of the actual survey. Additionally, although the first and second author gave a webinar on instrumentation issues in implementation science through the VA program, we do not believe that the content affected the perceptions of survey respondents as the issues discussed in the talk were presented in a different order than they were presented in the survey.

7. (compulsory) On Page 17, the authors note that their survey found that “…48% of the reviewed constructs are considered to be in high need of instruments…” This finding is very difficult to interpret without more information. The authors do not present a list of “reviewed constructs” and do not explain what question(s) this statistic is based on.

We have now removed all SIRC IRP preliminary data and so this confusing statement is no longer in the manuscript.

7a. Another example of an opaque reference to survey results is found on Page 21: “…practical instrumentation was indeed prioritized over and above psychometrics with a ratio of 52:48…”

We have added the survey as an additional file and included a sentence to aid in interpretation of this observation.

7b. There are other examples in the paper. In general, selective reporting of survey results is unacceptable for a scholarly paper. One solution is for the authors to publish the results of their survey study separately and cite that report in subsequent conceptual papers.

We believe that the format and limited information on the survey made it seem as though we were engaging in selective reporting, when this is not the case. By adding the full survey as an additional file, clarifying its role and the data collected, we hope that the findings are more easily understood. We decided to keep the format of presenting the survey data throughout the paper (where it was most relevant) to support the issues discussed. We hope that this was a wise choice as opposed to simply having a “results” section, which is not part of the Debate style paper that Implementation Science uses. Additionally, it appears as though this reviewer misunderstood that the SIRC IRP is not a survey, but it is a different project from which relevant preliminary outcomes were previously drawn to inform the recommendations delineated in the debate. However, all mention of preliminary results of the SIRC IRP have now been dropped.
8. (discretionary) The authors appropriately criticize the overreliance on self-report instruments. What is not clear is what kinds of quantitative measures the authors endorse. Outside of fidelity scales, most widely-used standardized instruments with established psychometric properties are self-report instruments. I was uncertain which instruments or kinds of instruments the authors were holding up as best practices. The paper would greatly benefit by providing positive exemplars in addition to all the admonitions about the shortcomings of implementation measurement. Tell me how to do it right, don’t just hammer me for what I do wrong.

First, we did not intend to imply that self-report measurement is inherently insufficient, which we have clarified in the manuscript. Second, we have elaborated on our recommendations in the self-report section to refer to the recommendations for mixed methods approaches, use of direct observation, multi-informant methodologies, and use of administrative data. Finally, we also included recommendations in the instrument development section for how to optimize the validity and reliability of self-report measures.

9. (discretionary) Figure 2 and elsewhere in the paper the authors introduce a substantial amount of terminology, none of which is formally defined in the paper. The inclusion of many undefined terms appears to be incongruent with the authors’ criticism of a lack of definitional foundation in much of the implementation literature.

We apologize for not defining the terms introduced in the figure. We have made efforts to review the manuscript and define any technical terminology.

Reviewer #2

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.

Thank you for your support and helpful feedback. We agree that the “meat” of the manuscript should focus on the recommendations, and have therefore added more information to each of the recommendation sections to address this concern.

Major compulsory revisions.

10. 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?

We agree that more thoughtful recommendations on how to create definitions are needed. We have elaborated on this process in the revised manuscript.
10a. 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?

We hope that extended discussion of recommendations for readers suffices to address this comment.

10b. 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.

We propose and have added a 4–stage model for identifying what instrument to use: 1. Identify an applicable theory, 2. Determine which constructs in that theory seem relevant to your specific implementation effort, 3. Carefully define those constructs, and 4. Assess construct validity of the instrument to ensure that the items in the instrument tapping proposed constructs are valid.

11. 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?

This is an excellent point. We have added into the manuscript information on power analysis and have addressed the issue of adequate sample sizes in the psychometrics section of the manuscript. Moreover, we have elaborated on the recommendations to make it clear exactly what statistics would need to be reported in order to contribute study information to subsequent meta-analyses.

11a. 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.

**We agree that this is an important point to add. We have added information and relevant sources to provide more helpful information on how to assess and report on internal consistency.**

12. 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?

**We have added additional sources to help readers address this issue.**

13. 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?

**Although we agree that this is an interesting and important issue, we believe that grappling with the difference in kinds of validity is outside the scope of this manuscript. However, we have added additional sources for interested readers and highlighted the importance of exploring criterion-reference validity such as predictive validity, given the utility for the field.**

13a. 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?

**This is a complex issue and the reviewers’ comments have been importantly thought-provoking. We have added a section in the manuscript addressing this issue and provided an exemplar for readers to refer to.**

14. 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:

**We appreciate your feedback that this issue may be overly complex. We have attempted to clarify the issue in the revised manuscript, thinking it an important one for stakeholders to grapple with.**
14a. Who? What are the recommendations for this section?

**We have added additional recommendations based on this feedback that addresses the question of “who” to measure.**

14b. When? “Careful attention must be paid to which constructs are assessed at each stage?” What does this mean exactly? How would this be done?

**We have added additional recommendations to the manuscript and advised stakeholders and model-developers to attend to the work of Proctor et al. (2011) that details when it is important to measure certain implementation outcomes in the implementation process and what measurement at that stage in the process can indicate.**

14c. “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?

**This point is very important and we appreciate the reviewer’s thoughtful comments and suggestions on the matter. The problem with assessment using simply pre-post evaluation is that researchers cannot investigate what variables are occurring or affecting change in between assessment points. As the reviewer mentioned, this may increase participant burden and add time to the research design. We have added a brief piece in the manuscript to address these points.**

14d. 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?

**We hope that the answers to these issues are adequately addressed in the “Use of Frameworks” section of the manuscript.**

14e. Also, I don’t see this particular issue in Table 1. Where did it come from?

**Though this was not explicitly investigated in the survey, this section arose from our work in the field. We hope that by including it, we can engage the field in an interesting discussion.**

15. Instrumentation issue #5. Do the authors have any suggestions for addressing the self-report issue?

**Please see our response to Reviewer 1’s comment #8.**
16. 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.”

We have added more information in the psychometric portion of the paper addressing optimal length for instruments with regard to both practicality and psychometric validation.

16a. No recommendations are provided under “Language.”

We apologize for this oversight. We have added recommendations from test development theories on the “language” issue of the manuscript.

16b. Perhaps suggest piloting instruments with a few subjects using a “think aloud” technique?

We agree that a think-aloud technique would be an acceptable method of assessing the appropriateness of language and have added a recommendation about this in the manuscript.

Minor compulsory revisions

Please provide additional information on the design:

17. “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.

We have added more details to the manuscript on how the instrumentation issues were identified for inclusion in the survey; please see response to Reviewer 1 Comment #6.

18. “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.

We apologize for the lack of clarity on the listservs. We only contacted two listservs, which we deemed to be the most appropriate points of contact for this subject due to the diversity and experience of the stakeholders contained in the listservs. We have obtained more information on the stakeholder makeup of both groups, ABCT DISSIG and the SIRC listserv, and have added information on the broader group membership into the manuscript to provide an indication of representativeness. Please see our response to Reviewer 1 Comment #6a.
19. How many people received a request to complete the survey? What was the participation rate?

This information has now been added.

20. 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.

We apologize for the confusion. There were 12 issues put forth in the survey as well as the option to add “other” issues self-generated by respondents. Therefore, the average could exceed 10 with respect to rank order. This confusion likely arose because we eliminated the bottom two issues from the table because they were identified as not being critical issues in Implementation Science.

Discretionary revisions

21. 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.

Though details pertaining to home-grown and adapted instruments are interwoven throughout the manuscript, we believe it is important to pull these issues out on their own as they remain pervasive in the field.