Author's response to reviews

Title: Psychometric Assessment of Three Newly Developed Implementation Outcome Measures

Authors:

Bryan Weiner (bjweiner@uw.edu)
Cara Lewis (lewiscc@indiana.edu)
Cameo Stanick (cameo.stanick@gmail.com)
Byron Powell (bjpowell@email.unc.edu)
Caitlin Dorsey (dorsey.c@ghc.org)
Alecia Clary (alecias@live.unc.edu)
Marcella Boynton (mhb23@unc.edu)
Heather Halko (heather.halko@umconnect.umt.edu)

Version: 1 Date: 13 Jul 2017

Author’s response to reviews:

Reviewer 1

From what I understand of the design and the results and Table 2, within each vignette the manipulation of the concept to reflect the three measures (acceptability, appropriateness, and feasibility) were all high or all low. So there was not a situation within a vignette where high and low was manipulated between the 3 measures to create some sort of fractured factorial design. Given the correlation between the measures - what are the limitations of this approach? In other words, could of the validity testing and other measurement properties be inflated if the concepts of acceptability, appropriateness, and feasibility reinforced each other to make the highs perceived to be even more high and the lows perceived to be even more low? While still an excellent study a statement about choice and potential drawback about this choice would strengthen this paper. It would also be relevant to the Proctor model - the examples given by the authors speak to circumstances where 1 or 2 of the three concepts - acceptability, appropriateness and feasibility may not be aligned.

Response

We regret any confusion. As we noted in the original submission, due to a study design error, two of the eight possible vignettes were not fielded: one in which acceptability was high, appropriateness low, and feasibility high; and the other where acceptability was low,
appropriateness high, and feasibility low. We have added a sentence in the revised manuscript (lines 328-329) to clarify that all other combinations of high and low levels of the three factors were represented in the six fielded vignettes. We noted in the original submission that this design error limited our ability to explore the discriminant validity of our measures of acceptability and appropriateness, as two of the four possible vignettes in which these constructs were designed to move in opposite directions were not included. See lines 323-327, 373-382, and 481-484 in the revised submission. In particular, we could not explore possible interactions which might indicate that study participants had trouble in distinguishing acceptability and appropriate when both conditions were “high” or both conditions were “low.” That said, Dr. Boynton, the psychometrician involved in the studies, noted that the main effects indicated that there is sufficient basis to state that (1) the measures could differentiate groups with known differences and (2) the 0.77 correlation of the acceptability and appropriateness scales is less than the 0.85 correlation that is used as a rule of thumb for saying that two measures are so highly correlated they are likely measuring the same thing.

Reviewer 1:

Response rate for Study 1 was forgotten - how many individuals had to be invited to find the 62 participants - and by what means were they recruited. This did not happen in the other two studies - I believe an oversight.

Response:

We added the number of implementation scientists and mental health professionals recruited to participate in the study. We also added that they were recruited by email. See lines 190-193 in the revised manuscript.

Reviewer 1:

Suggestion. The authors may wish to unpack the results and separate the more technical phrases from the interpretation phrases into two distinct statement to reduce cognitive load of reader and to consider simplifying the interpretation message in places. While the technical aspects are extremely well written, for some readers of this paper the methods may be quite new.

For example. From this:

Median weighted assignments for all but five items were significantly greater than zero after applying the Hochberg correction for multiple tests, indicating that participants judged the items to reflect to a significantly greater degree the constructs they were intended to measure than they did the other constructs.

To something like this:
Median weighted assignments for all but five items were significantly greater than zero after applying the Hochberg correction for multiple tests. Thus, items were judged to "fit" significantly better to the construct they were intended to measure than to the alternative constructs.

Response:

We implemented this suggestion in lines 268, 367, and 375.

Reviewer 1:

2³ANOVA - I would consider spelling out (again) the three factors and that the levels are high and low.

Response:

We spelled out (again) the three factors and two levels in lines 355 in the revised manuscript.

Reviewer 2:

Lines 234-238: Did you use the same data for the EFA and CFA? Or did you split the sample in half? You may wish to explain/justify in the article. I appreciate the use of CFA since you have an a priori theoretical model. EFA may also be appropriate, but perhaps explain here why you chose to conduct EFA (it becomes clearer in the results section).

Response:

We revised the manuscript to clarify that (a) we used the same data for the EFA and the CFA and (b) we performed EFA prior to CFA to trim the number of items for each scale based on factor loadings and ascertain whether there were additional factors beyond the expected three due to item cross-loadings. See Lines 239-247. We did not employ the split-half approach in Study 1 since we had another sample in Study 2 for replication.

Reviewer 2:

Line 216 and Line 277: The manuscript indicates Table 1 is to be inserted at both of these locations.

Response:

Thank you for pointing this error. In the revised manuscript, we inserted Table 1 at line 225 (what was line 216 in the original submission). We inserted Figure 1 (not Table 1) at line 293 (what was Line 277 in the original submission).
Reviewer 2:

Lines 270-273: RMSEA is more likely to reject fit with small samples. It would be helpful if you added a test to compare change in fit between nested models. I recommend the Schwarz Bayesian Information Criteria, which will at least give you an estimate of how strongly the evidence favors one model over the other (see Raftery, A. Bayesian Model Selection in Social Research. Sociological Methodology, Vol. 25, 111-163. 1995).

Response:

We now report in lines 288-289 the increase in the BIC for the poorer-fitting 1-factor model compared to the 3-factor model.

Reviewer 2:

Lines 330: You might consider the use of McDonald's Omega instead of Cronbach's alpha. Omega relaxes the assumption of tau-equivalence (equal factor loadings) between the items of a factor. Omega is equal to or greater than alpha. The factor loadings appear to be nearly tau-equivalent in figure 2, but not in figure 1.

Response:

Although we agree that McDonald's Omega offers certain advantages, given that it is not commonly used in the literature, its use may not be informative to most readers. The reviewer is correct in asserting that a Cronbach’s alpha has several limitations; however, in the current context, it is only used as a descriptive scale reliability statistic and was in no way used to inform the scale development process. As such, we choose to report the Cronbach’s alpha statistics.

Reviewer 2:

Figures 1 and 2: Are all numbers significant at p<.05? It's difficult to tell if all or none of the numbers are bold. Can you make this clearer?

Response:

Yes, they are. We altered the legend in the Figure 1 and Figure 2 to indicate this.

Reviewer 2:

Lines 331-332: I expect the 3-factor CFA model is the same model identified in study 1, but please clarify.

Response:
We have clarified in line 348 that we estimated in Study 2 the same 3-factor model that we estimated in Study 1.

Reviewer 2:

Lines 334: I think you are calling the 1-factor model a "general concept of social validity", but it feels like you are introducing a new concept. Please clarify or refer the model as an omnibus 1-factor CFA as you do in line 353.

Response:

We eliminated reference to the general concept of social validity since it is likely to be unfamiliar to many readers and is not necessary to justify fitting a 1-factor model. Instead, we refer to it as an omnibus 1-factor model. See line 345.

Reviewer 2:

Line 342-343: Why trim the scale further… because shorter scales are more pragmatic? Do you favor one version of the scale over the other? Did you plan to produce a 4-item scale when you started the project?

Response:

Yes, we trimmed the scales further to produce the briefest measures possible on pragmatic grounds. We did not know when we started the project how few items we would need to develop reliable, valid, and pragmatic measures. However, we are pleased that we can offer researchers three measures that are as brief as possible.

Reviewer 2:

Lines 353-357: Again it would make your argument stronger if you used the Schwarz BIC, which will likely give you "very strong" evidence in favor of the 3-factor model.

Response:

We report in lines 375-378 the increase in the BIC for the poor-fitting 1-factor and 2-factor models compared to the 3-factor model.

Reviewer 2:

Lines 417-418 and 421-422: First you say you use the 5-item scales and then you say you used the 4-item scales. Please clarify.
Response:

Thank you for pointing out this error in the text. We have corrected the text so that it now refers to the 4-item scales derived from Study 2 (see line 438).