Author’s response to reviews

Title: Validity and reliability of measures to assess constructs from the Inner Setting domain of the Consolidated Framework for Implementation Research in a pediatric clinic network implementing HPV programs

Authors:

Timothy Walker (timothy.j.walker@uth.tmc.edu)
Serena Rodriguez (Serena.Rodriguez@UTSouthwestern.edu)
Sally Vernon (Sally.W.Vernon@uth.tmc.edu)
Lara Savas (lara.staub@uth.tmc.edu)
Erica Frost (Erica.L.Frost@uth.tmc.edu)
Maria Fernandez (Maria.E.Fernandez@uth.tmc.edu)

Version: 1 Date: 11 Jan 2019

Author’s response to reviews:

Thank you for considering our manuscript. We have submitted a revised manuscript addressing comments and suggestions, as detailed below:

Reviewer 1 (overall synthesis)

A well-written manuscript which provides important information on the validation of an implementation science framework. Well done for producing such a methodologically robust article. The article will also be of great interest to readers with stern interests in psychometrics!!!

INTRODUCTION

A well written, informative and balanced introduction section. However, authors can further improve clarity by explaining in greater detail the Consolidated Framework for Implementation Research (CFIR). As much as the manuscript focuses on the inner setting domain, it will be important to provide full contextual information for the sake of the readership who may not be
familiar with the CFIR. My understanding is that implementation science is still an evolving field and providing more information on the CFIR will further shed more important information on implementation science.

Response: We added more information about the Consolidated Framework for Implementation Research (CFIR) in paragraph 2 (page 4) of the introduction. This is to provide readers with more contextual information about CFIR as recommended.

METHODS

The section is well written, detailed enough and is easy to follow

It could be helpful if authors could provide the Trial registration number and or link to the trial protocol if available.

Response: This was technically not a trial so it was not registered as such. We realize we referred to the study as a “pre-post intervention trial” in the manuscript. We removed the word “trial” in the first paragraph of the Methods section (page 5) to more accurately represent the type of study.

Can the authors also provide the ethical approval number/reference

Response: We have added the ethics approval number to the first paragraph in the methods section on page 6.

Line 17-19, Page 7, can the authors please provide the meaning abbreviations, not everyone would be familiar with the names of the states

Response: We removed the state abbreviations as recommended and spell out each state in the Measures paragraph (page 7) in the Methods section.
Excellent description of the CFA modelling and justification of the tested models is theoretically sound

RESULTS

There is need to report how missing responses were dealt with

Response: We added a line in the 5th paragraph of the Data Analysis section within the Methods (page 9) to clarify that we used full information maximum likelihood estimation, which accounts for missing data.

DISCUSSION

Authors did well by first restating the study objectives and summarizing key findings, this makes it easy for readers to synthesize the study outcomes given the somewhat lengthy results section

A well-balanced discussion and conclusion section

Reviewer 2: The authors addresses an important topic, investigating the validity and reliability of a measure used in implementation research. The authors report that the Inner setting measures demonstrated good validity and adequate reliability with the exception of available resources. The authors conclude that their results extend previous work by providing additional psychometric evidence to support the use of these Inner Setting measures in pediatric clinics implementing human papillomavirus programs.

The article is well written, the rationale for the study is provided in a clear and logical way and the methods used to assess validity and reliability are, in general, well described. I have some concerns/questions that I would recommend the authors to address.

1) I would suggest to include in the methods section information about the strength (e.g., accepted cut-offs) of the intra-rater reliability analysis as well as for the discriminant validity
analysis in a similar way as for the cut-offs used to address the fit of the models created. In the results section, the authors address this, for example on page 13, lines 19-25 "The ICC(2) values ranged from 0.34-0.67 where they were all below the recommended levels (0.7 to 0.8 and higher) [24,25], indicating a lower than desired level of reliability for group means" Nevertheless, I would recommend to include these cut-offs also in the methods section.

Response: We agree this information would be useful in the methods section. We have added information about the meaning and expected values for these tests in paragraphs 6 and 7 of the Data Analysis section within the methods (page 10).

2) Furthermore, the authors examines the reliability of the measure with an ICC analysis. This provides information about the relative reliability of the measure, however I would also recommend an analysis of the absolute reliability of the measures (e.g., exploring the SEM) and to provide information about the MDC of the measures.

Response: For the reliability testing in this study, we chose an approach that is consistent with what is used in organizational research [1,2]. The purpose of testing ICC(1) and ICC(2) is to understand how much variance is explained at the clinic level and how reliable the mean scores are within a clinic. This provides important information when aggregating individual data to represent a clinic level construct. In addition, this information provides a general sense of reliability that is standardized. We feel this information should be emphasized rather than an absolute measure of reliability that is not traditionally reported when examining reliability of multilevel constructs [1, 2].


3) Provide more information amount which type of ICC(2) measure that was used. For intra-rater reliability analysis are often a two-way mixed absolute agreement model recommended. Was this used?
Response: We used a one-way ANOVA model to calculate both ICC(1) and ICC(2) measures. Components from the ANOVA model (between-group mean square, and within-group mean square) were used to calculate the respective ICCs. We added the specific equations used, along with corresponding citations in the 7th paragraph of the Data Analysis Section within the Methods (page 7). We feel the one-way ANOVA model is appropriate because the clinic is the subject and we have different raters for each respective clinic (or subject).

4) For the reliability analysis, the authors only report the mean ICC values. I would recommend to report also the lower and upper bounds of the reliability analysis. See Koo et al. Journal of Chiropractic Medicine (2016) 15, 155-163

We are reporting the ICCs in a manner consistent with organizational research [3]. Because ICC(1) and ICC(2) are calculated manually using output from a one-way ANOVA model, the upper and lower bounds are traditionally not reported.


a) The authors conclude that the the measure has adequate reliability. Considering the reliability measures, is this actually true? In Table 6, mean ICC values range between 0.34 to 0.67 which would imply poor to moderate reliability. Considering that the lower bounds of the mean ICC values are even lower, I would highly question the reliability of the measure. What are the authors opinion on this?

Response: In our study, we are trying to assess the reliability and validity of measures for clinic-level constructs from individual-level responses. We do not expect each participant within a clinic to rate a construct (e.g. culture) exactly the same. This is because each individual has a unique perception that is influenced by the shared characteristics of their respective clinic.

The ICC(1) values indicate that clinic-level variables are present – otherwise ICC(1) values would be 0, meaning no variance is explained by the clinic level. However, the ICC(1) results also show that these variables are not isomorphic, meaning respondents within a clinic do not provide the exact same scores for a respective construct. Variability between raters is common
with composition processes, where individuals are responding to items about a group (or clinic) level construct [1].

The ICC(2) values are relatively low, in part because the ICC(1) values are low. The more variance explained at the clinic level will lead to more reliable group mean scores.

Overall, we feel the interrater reliability is relatively low, in particular for available resources. However, we feel the values are adequate for other constructs because we do not expect complete agreement. Further, our multilevel CFA tests suggest there are similar factor structures between the individual and organization levels further supporting that the individual responses can be aggregated to assess an organization-level construct.

We updated our explanation of the reliability results in paragraphs 5&6 in the Discussion Section (page 16) to better explain the implications of the reliability findings.

5) Even though the sample size was large, I would recommend to provide information on a sample size calculation for the different measures of validity and reliability.

Response: We agree that sample size considerations are important for this work. This is especially true for the multilevel confirmatory factor analysis (CFA) models. Thus, we added a line in the 2nd paragraph of the Data Analysis section within the Methods (page 8) to indicate the sample size recommendations for these models.

We feel this is sufficient because the required sample size for confirmatory factor analysis (CFA) models is in part conditional on the factor loading strength, where solutions with stable and higher factor loadings (i.e., stronger items) reduce the influence of the sample size [4, 5]. Generally, larger sample sizes will decrease sampling error variance, which can help lead to more stable solutions [6]. However, sample size adequacy for factor analysis models cannot be entirely known until analyses themselves have been conducted. Further, we feel we have an adequate sample to carry out the primary aims of this study.

4) de Winter JC, Dodou D. Factor recovery by principal axis factoring and maximum likelihood factor analysis as a function of factor pattern and sample size. Journal of Applied Statistics
5) de Winter J, Dodou D, Wieringa PA. Exploratory factor analysis with small sample sizes.
Multivariate behavioral research 2009;44:147-81.

6) Hogarty KY, Hines CV, Kromrey JD, Ferron JM, Mumford KR. The quality of factor solutions in exploratory factor analysis: The influence of sample size, communality, and over determination.

6) In the discussion (page 14, lines 4-7) the authors report that: "While the measures demonstrated good validity, there is additional work that can be done to improve reliability, in particular for the available resources measure." Could the authors provide any information on how they think that the reliability of the measure could be improved? Later in the discussion, the authors suggests that "Therefore, future research should focus on investigating factors that could impact reliability such as examining how people in different job types within a clinic respond to these clinic-level constructs."

a. Could there by other factors, in addition to different job types, that could affect the reliability?

b. Considering the large sample size of clinical staff, I would recommend to do a sub-analysis based on job types to explore the impact of job status on the reliability of the measures? This however, requires support from a power analysis.

Response: We appreciate the comment as we feel this section within the manuscript requires a more focused explanation. First, we restated that additional work is needed to examine factors influencing reliability rather than “improve” reliability in the first paragraph of the Discussion Section on page 14.

We also chose to edit paragraphs 5&6 in the Discussion Section (page 16) to further explain how the ICC(1) and ICC(2) values are interrelated and what the practical implications from the results may be. There, we state that the number of people surveyed per clinic could be another factor influencing reliability.
We agree that examining the impact of job type on reliability is an important next step. However, we feel this is worthy of its own study given the purpose of this work was to examine the validity and reliability when aggregating data across all respondents in the sample.

7) How is the tool intended to be used? Is it more on a clinical staff or network level? If the former, than the low reliability measures needs to be further addressed. If the goal is that the constructs should be used in research and practice settings than the low reliability would be a highly limiting factor for the use of the measurement.

Response: The measures are intended to be used by practitioners and researchers to gain a better understanding of clinic-level factors that may influence implementation. The reliability information presented here can inform survey approaches in future work. Therefore, even though the reliability may have been lower than preferred, we feel the collective evidence from validity and reliability testing indicate these are acceptable measures to use for this purpose.

8) I would recommend the authors to follow GRRAS or COSMIN guidelines to improve the reporting of the findings.

Response: We agree that reporting guidelines are important for this work. We chose to use a guideline that focused on multilevel factor analysis [7], which is consistent with the primary focus of this manuscript.