Title: Development of ReproKnow, a Reproductive Knowledge Assessment for Women with Rheumatic Diseases

Authors:

Mehret Birru Talabi (birrums@upmc.edu)
Megan Clowse (megan.clowse@duke.edu)
Susan Blalock (s_blalock@unc.edu)
Galen Switzer (gswitzer@pitt.edu)
Lan Yu (Lan.Yu@pitt.edu)
Alaina Chodoff (chodoffa@upmc.edu)
Sonya Borrero (borrsp@upmc.edu)

Version: 3 Date: 14 Aug 2019

Author’s response to reviews:

Dear Dr. Fitzpatrick,

Thank you for the opportunity to respond to Reviewer 2 and the Statistical Advisor. We offer our responses below:

1. Authors did not explain their power analysis that how they decided to include 158 participants. It can’t be justified by saying preliminary step to validate. Literature suggested that for validation of an instrument, you need a larger sample size preferably 300 or more. As Comfrey and Lee suggested that sample size equal to 100 as poor, 200 as fair, 300 as good, 500 as very good, and 1000 or more as excellent [Comfrey AL, Lee HB. A First Course in Factor Analysis. Hillsdale, NJ: Lawrence Erlbaum Associates; 1992]. It also depends on the variability in the population and types of questionnaire being used as suggested by Osborne and Costello [Osborne JW, Costello AB. Sample size and subject to item ratio in principal components analysis. Pract Assess Res Eval. 2004;9:8].

Thank you for this comment. In scale development, no absolute minimum sample size has been defined for the validation of an instrument. This is a major limitation in the field of scale development. In the “validated” scales that we reference in our paper (citations 15-18 and 25),
the sample sizes include 83, 145, 171, 300, and 354 patients (for reference, ours is 153). A relatively recent review suggested that less than 10% of scale validation studies explicitly mention sample size justification in their methodology [Anthione (2014) Health Qual Life Outcomes]. This may reflect the fact that there are no standards in this area—only recommendations for which individuals have their own preferences.

The Statistical Advisor mentions Comfrey’s suggestions (e.g., 100 is poor, 200 is fair, etc). Dozens of investigators have provided suggestions for the absolute minimum sample size needed for scale validation. We do not intend to enter this controversial debate. However, we prefer Gorsuch’s and Kline’s recommendations, both of which are widely-cited in the literature, and suggest that 100 patients are the absolute minimum number of patients required for scale validation [Gorsuch, R. L. (1983). Factor analysis (2nd ed.). Hillsdale, NJ: Erlbaum; Kline, P. (1979). Psychometrics and psychology. London: Academic Press; Kline, P. (1994). An Easy Guide to factor analysis. Routledge Press].

Additional guidance is provided by Bryant and Yarnold who cite the subjects-to-variables (STV) ratio as being the most widely accepted rule of thumb for sample size [Bryant FB, Yarnold PR. Principal Components Analysis and Exploratory and Confirmatory Factor Analysis in Reading and Understanding Multivariate Statistics. Chapter Four. Grimm and Yarold editors, pgs 99-136). This rule suggests that the STV should be 5 or greater. In our case, the STV is 158/10 – that is 158 participants over 10 variables – STV = 15.8 which is well above the minimum standard of 5.

Ultimately, all investigators in scale development would agree that larger samples are better for validating scales, but it is also clear that there is no single agreed-upon standard. Thus, while our study exceeds the absolute minimum number of patients required for preliminary validation, we do acknowledge that a larger sample size would have been ideal in the Discussion section. Furthermore, we agree with Reviewer 2 that the scale must be validated in other populations, and hope that upon publication of this manuscript, we may encourage others to validate the scale among larger, established cohorts of women with rheumatic diseases.

2. I’m wondering why did authors not describe PCA in method section and its results in the results section. It is important for readers to see the underlying components that are being measured by the survey questions. There could be several steps related to its implementation and goodness-of-fit should be explained properly. For example, when grouping factor loadings, usually we look for values that are ±0.60 or higher, although this varies depending on what the rest of the loadings look like.

Similarly, in case of lower Cronbach’s Alpha, it is not clear whether authors tried “scale if item deleted” option in SPSS. It should have been described in this manuscript.
Reviewer 2 and the Statistical Advisor both acknowledge that the sample size was a limitation, but asked us to report on the principal components analysis. The reason why we did not report on the PCA in our results was actually based on Reviewer 2’s earlier comment that our sample size is too small for PCA. Theoretically, the request to conduct and describe PCA in the context of our sample size, which was suggested to be too small for PCA, seemed somewhat inconsistent. No clear factor solution had emerged from our analysis, which was acknowledged by both the Statistical Advisor and the Reviewer. Given the results of the PCA and factor analyses that suggested that we did not have a single factor solution and given that clear subdomains did not emerge, it remains somewhat unclear to us why internal consistency or additional reliability estimates would be appropriate.

However, we do understand that these steps are part of the general “checklist” for scale validation. Perhaps some readers will look to see if these steps were done. Thus, if preferred by the Editors, we would be happy to add the following sentences to the Methods, Results, and Discussion:

Potential Addition to Methods: Cronbach’s alpha was used to evaluate the internal consistency of ReproKnow; a coefficient of 0.7 or higher is generally considered to be acceptable for established scales, although coefficients of at least 0.6 may be considered for newly-created or preliminary scales [21-23]. We also completed a principal components analysis for nominal level variables in order to assess the dimensionality of the scale, with a confirmatory factor analysis for dichotomous variables using the Hull method to determine the number of factors in the scale (Lorenzo-Seva, Timmerman and Kiers, 2011).

Potential Addition to Results: The Crohnbach’s alpha statistic for ReproKnow was based on patient responses, and was estimated at 0.62, which demonstrates moderate internal consistency. To assess if the moderate internal consistency was secondary to multidimensionality in the scale, we subsequently conducted a principal components analysis (PCA) for nominal level variables. The PCA identified four items on two dimensions that had no clinically meaningful or contextual relationships. Next, we removed these four items from the analysis, and subsequently removed all possible combinations of the items from the analysis. This subanalysis did not change the alpha level. Our factor analysis results were similar to the PCA analysis in a one-factor solution was also recommended using Hull criteria, but that factor explained a minority of the overall variance (37.6%) and several items had either moderately low (<.40, item 4; <.45 items 7 and 9), or very low (< .20, item 3; <.25, item 5) loadings.

Potential Addition to Discussion: ReproKnow’s relatively low internal consistency might be considered a potential weakness. While Cronbach’s alpha is ideal for scales that have multiple response options (e.g., Likert), coefficients may be artificially low for scales with fewer responses [21]. Our interpretation of our findings from PCA and factor analysis were that a
single factor solution did not sufficiently explain the variance in the model, and a multiple-factor solution lacked clinical or conceptual meaningfulness. It is possible that future research with ReproKnow involving larger samples of women will reveal a meaningful latent structure. However, the internal consistency, PCA, and factor analysis results may also reflect that ReproKnow is meant to test a broad range of topics across reproductive health, including pregnancy, pregnancy prevention, lactation, and heritability.

3. Validity of any instrument is crucial. Several approaches to quantify the judgment of content validity across experts are available in literature. For example, the content validity ratio [Lawshe CH. A quantitative approach to content validity. Pers Psychol. 1975;28:563–75] and content validation form [Barrett RS. Content validation form. Public Pers Manage. 1992;21:41–52]. However, authors did not specify any with citation.

We did not evaluate content validity in such a way that actual statistics can be cited. Quantification of content validity is not an absolute requirement for survey development. Haynes, Richard, and Kubany suggest the following for content validation guidelines, “Use population and expert sampling for the initial generation of items and other elements. Carefully structured, opened-ended interviews with persons from the targeted population and experts can increase the chance that the items and other elements are representative of and relevant to the facets of the construct. This process can also suggest additional facets and the need for construct refinement.” (Haynes SN, Richard DCS, Kubany ES: Content validity in psychological assessment: A functional approach to concepts and methods. Psychol Assessment 1995, 7(3):238-247.) Five to seven experts are generally recommended for this step. In our study, ReproKnow’s content was generated, evaluated, and refined based on input from a team of ten experts and three patients without the application of statistical approaches. Again, while statistics may be used, our process is also a widely accepted way of ensuring appropriate content.

Our Methods section discusses our process in greater detail. To further clarify this process in response to the Statistical Advisor, we have made the following additions to the paragraph “ReproKnow Development” [pgs 8-9, lines 171-186]:

“The content of ReproKnow reflected topics addressed in a patient educational pamphlet produced by the ACR about women’s reproductive health that is freely available on its website, and a review article about family planning for women with rheumatic diseases written by several of the current manuscript’s authors [26, 27]. Preliminary questions that addressed heritability of rheumatic diseases, birth outcomes, likelihood of fertility, contraception safety and efficacy, preconception care, pregnancy management, lactation/breastfeeding safety, and medication risk/safety, were developed by one of the principal investigators (M.B.T.). Content validation, as defined by Haynes et al., should include population and expert sampling for the initial generation of items and other elements of the scale. To optimize the content validity of ReproKnow, a
group of ten local and national rheumatologists, obstetrician-gynecologists, internists with formal women’s health sub-specialization, nurses, a pharmacist, and a survey methodologist, reviewed the questions. Based on their input, six questions were extracted, and the remaining questions were refined. Three female reproductive-age patients with SLE were recruited from an outpatient rheumatology clinic to participate in cognitive, “think-aloud” interviews while they used the preliminary tool. Their feedback about the clarity and content of the tool were used to make additional revisions to the questions and response options.”

4. Authors did not report any pilot testing on the intended population [please note that comparison between the fellows and nurses cannot be considered as pilot testing]. Recommendations on sample size for pilot testing vary. Some academicians are staunch supporters of things like a 20 participant per question. The more participants the better, but if all you can get are 60 participants, it may be enough, especially if your survey is short [about 8-15 questions]. Of course, variability in the population of interest play an important role to decide numbers required in the pilot testing.

We did not pilot test widely in the intended population, although we included patient stakeholders in conceptualizing topics of importance, and piloted the test among three patients to assess for readability and comprehension of the test. This is described in our Methods section.

5. The way authors measured construct validity is different than what is suggested in literature. The construct validity of a questionnaire can be evaluated by estimating its association with other variables (or measures of a construct) with which it should be correlated positively, negatively, or not at all [Construct validity in psychological tests. CRONBACH LJ, MEEHL PE Psychol Bull. 1955 Jul; 52(4):281-302]. In practice, the questionnaire of interest, as well as the pre-existing instruments that measure similar and dissimilar constructs, is administered to the same groups of individuals. Correlation matrices are then used to examine the expected patterns of associations between different measures of the same construct, and those between a questionnaire of a construct and other constructs. It has been suggested that correlation coefficients of 0.1 should be considered as small, 0.3 as moderate, and 0.5 as large [Cohen J. Statistical Power Analysis for the Behavioral Sciences. 2nd ed. Hillsdale, NJ: Erlbaum; 1988].

As quoted directly from the citation by Cronbach provided by the Statistical Advisor: “We can use many methods in construct validation. Group Differences: If our understanding of a construct leads us to expect two groups to differ on the test, this expectation may be tested directly... only coarse correspondence between test and group designation is expected. Too great a correspondence between the two would indicate that the test is to some degree invalid, because members of the groups are expected to overlap on the test.” Thus, we do believe our testing of
ReproKnow among different types of users (fellows, nurses, patients) was one way to assess construct validity, and is aligned with the Statistical Advisor’s citation.

Unfortunately, the Advisor’s comment that ReproKnow should be tested against pre-existing instruments cannot be done because there are no similar knowledge assessments broadly for women with rheumatic diseases. This underscores yet another reason why this work is novel.

We agree that our assessment of validity is limited in this paper, as it is for each of the published papers on knowledge assessments that we cite in our manuscript—none of which address all six aspects of construct validity as described in the academic citation provided by Reviewer 2. As described in our Discussion section (lines 381-392), some aspects of validity must be tested in the future with different cohorts of women (e.g., generalizability, ecological); one way to ensure that this happens is to publish our work for other investigators to use among other samples of women.

Again, Dr. Fitzpatrick, the overall purpose of our study is to propose a reproductive knowledge assessment for a clinically vulnerable group of women for whom little educational resources about reproductive health exist. We believe this is important work in a critically understudied area. We were selected to present data from this manuscript at the American College of Rheumatology annual meeting, which underscores that other investigators in rheumatology believe it is clinically relevant work.

We believe that our analysis and preliminary validation is conceptually sound, and certainly within the parameters of other published papers on the subject as cited in our manuscript.

We agree with Reviewer 2 that our initial results are promising. Some of the reviewer’s and Advisor’s comments will be addressed with different samples of women, and with longitudinal follow-up of patients who are exposed to ReproKnow. Again, this requires publication of our manuscript. We appreciate your time and re-review.

Sincerely and with thanks,

The Authors