Reviewer's report

Title: Validity of the ReproQ; a questionnaire measuring the World Health Organization concept of Responsiveness in Perinatal Care

Version: 2 Date: 3 June 2014

Reviewer: Ingeborg Strømseng Sjetne

Reviewer's report:

This paper describes the assessment of a questionnaire that was developed to measure the construct responsiveness in the context of perinatal care. The questionnaire built on WHO's definition of responsiveness, comprising eight domains.

To review and comment on other researchers draws attention and consumes time that I could have used for my own purposes. I am convinced that, as in this case, a group of eight (on the front page) or seven (as in "Author's contributions") authors could have processed this paper a lot further between them before submitting. It is frustrating to point out in writing (in a foreign language) flaws that coauthors could have pointed out just as well and easily, and to have a feeling that much of the flaws are caused by carelessness only. I hope that my comments contribute to the improvement of this paper that describes an important effort in ensuring quality in the care for women's reproductive health. Good luck!

1. Is the question posed by the authors well defined? OK
2. Are the methods appropriate and well described? See comments
3. Are the data sound? OK
4. Does the manuscript adhere to the relevant standards for reporting and data deposition? The study was approved by an ethical committee
5. Are the discussion and conclusions well balanced and adequately supported by the data? See comments
6. Are limitations of the work clearly stated? See comments
7. Do the authors clearly acknowledge any work upon which they are building, both published and unpublished? To my knowledge, yes
8. Do the title and abstract accurately convey what has been found? See comments
9. Is the writing acceptable? Yes, but could be improved on some points, that I have commented on.

- Major Compulsory Revisions
1. I feel rather uncomfortable with some of the terminology in use. I suppose we agree that "Responsiveness" is a term that represents a construct that is agreed upon by definition, for us to be able to discuss and observe a certain phenomenon. The phenomenon has no existence outside our consciousness. It seems to me that several places the term is used to represent the instrument that is presented and assessed in the paper. For example, lines 102 and 105. This should be checked and clarified throughout the text.

Many words could be used to denote the interpersonal/ non-clinical/ non-technical /responsiveness aspects of health care. I think that a true professional integrates both aspects in practice, but I see clearly the good point in discriminating between them for heuristic purposes. On line 74-75 the authors write "non-clinical aspects such as service quality are important too" meaning (I suppose) what Donabedian calls the interpersonal relationship/ interpersonal process. On line 77-78 "health services" seemingly includes also the technical aspects of the care. In general, I suggest you chose carefully what terms to use to denote all the phenomena that are most important in this paper, and stick to them.

2. Definitions should be presented as quotes, line 95-97 is not identical to its source. This is a detail, but of great importance, as the definition presents the main concept of the paper.

3. Six women were invited to judge the feasibility after the expert group had finished their part. I would like to know what the authors think about advantages and disadvantages with regard to including the women at the final stage.

4. Sumscores. I think the description of the transformation is unnecessarily detailed. How did you treat cases with item missing on items included in the domain/scale? (If you chose to keep the details, please tell where the 9 in the formula comes from.)

5. One reason for transforming the sumscores to a 1-10 scale was to "enhance comparability among domains with different numbers of items". With different numbers of items, the contents that is described by the scores is also by necessity different, and I would like to know some more about the value of such comparisons.

6. Is "scale" the same as "domain"? Consistent terminology is crucial to readability.

7. The basis for judging feasibility is somewhat confusing to me, as it seems like response rates and item missing is mixed. Line 219-223. I question the relevance of aggregating the prevalence of item missing per domain/scale, as I would rather view the distribution of item missing as a help to identify problematic single items. The way the authors have done it, many problematic single items may be veiled by the aggregation. I question the statement "in general, missing item rates below 20% can be considered acceptable". To solve potential misunderstandings, I tried without success, to find the primary source of the statement. I think that rather than relying on a preset cut off here, it would be
useful to use discretion, are the proportion of item missing evenly distributed, are there items with notably higher item missing prevalence? If patterns exist, can the patterns be interpreted in light of the subject matter in question? Some confusion may depend on unfamiliar terms, such as "inappropriate non-response rates", "systematic missing rates", and I suggest rewriting this to gain a clearer content.

8. Reliability assessment. Line 228, I suppose the intended meaning is coefficient above 0.70. have you considered using more recent literature on instrument assessment, for example the COSMIN checklist? [1, 2] I am not sure of the information value of the three correlation measures that are presented in the paper.

9. Discriminative validity. Here the authors write that the women "report better outcomes", but responsiveness is what they want to measure…? I do wonder, what is the logic behind hypothesizing that a woman whose child is hospitalized will report poorer responsiveness in the antenatal phase?

10. I miss information about the software that is used for the analyses.

11. 94 women declined the invitation, and this included women with incorrect phone number or address. I spot some carelessness here, please rewrite.

12. Table 1. The information about mean age's confidence interval is redundant. Information about parity adds to 120%. It is not clear what the t-tests are used for. The information in the text does not correspond with the table. The text in lines 258-261 need some revision. I spot some carelessness here.

13. Table 2. See point 7, regarding item missing. The way they are reported here, the missing responses may very well all be in one of the items of a given scale/domain.

Lines 276-277 must be rewritten, the meaning is unclear.

14. Line 281 "Mean transformed scores were relatively high…", compared to what? I would rather say that the distribution is negatively skewed. "Floor effects were observed in 0.6% of women". I.e., on seven of the 24 domains one woman's score was at the negative extreme. Line 285, this interpretation belongs to the discussion. Moreover, I find it quite reasonable that the domain scores differ between the phases, as the number of items (and hence the domain content) differ much between the phases. The averaged cronbach's alfa can be taken out of the text (line 288-289), it is not useful information.

Table 3. The 25th percentile = median on all three phases' scores for dignity. Please check if this is correct.

15. Table 4. Please consider if this table is necessary. In my opinion, this information is summarized effectively in the Cronbach's alfa scores.

16. Table 5. The table and the accompanying text confuses me. The terminology in use in the text and the table directs my understanding towards exploratory
factor analysis, but the authors introduced it as confirmatory factor analysis in the methods section. And why use an exploratory procedure, when the overall objective was to operationalize the eight domains. The procedure and reasoning must be far better presented.

17. Line 330. I can not see that the "presumed structure was grossly confirmed". Line 330 That it holds a "potential to discriminate between quality of care levels" is not substantiated, as far as I see. Also, it is responsiveness, as an aspect of quality of care that was to be measured. Consistent terminology supports readability. Line 337. That the independent value of the domains are supported by human rights is interesting on an policy level, but irrelevant to the assessment of this instrument.

18. Stengths and limitations. In general these sections are characterized by an everyday reasoning and language. I see no need in academic language per se, but in a scientific publication, I would expect to see arguments expressed in elaborate/specific terms.

Example in line 339. "Thirdly, in contrast to patient satisfaction questionnaires, responsiveness [the instrument?] tries to capture the patient's real experience, since literature has shown that expectations strongly influence patient satisfaction".

Line 344-345. The sentence gives no meaning. Line 347-348 "responsiveness (...) is still disturbed by at least some extent of 'subjectivity'' This can hardly be a limitation, given that the overall aim is to measure aspects of care quality from the client perspective. What would an 'objective' alternative look like? Line 348-349. The concept of responsiveness does not include financial barriers, hence it should not be held against it that it does not include financial barriers. And in the same line, collecting qualitative data could produce valuable supplements to data collected by ReproQ, but the aim of the present paper was after all limited to assessing the instrument itself.

Line 370. If you accept items that have a relatively high proportion missing answers because they are "pointing to service events that do not always take place", does this bring about consequences for how you calculate the domain/scale scores? Line 377-380. The contents of these lines are unclear. Line 390-391. I agree. Perhaps the scores for these amenities should not be collapsed? Line 397. Why should inter-item correlations in the antenatal phase be influenced by the phenomena that are suggested? Line 402. Consistent terminology, the authors write experiences her, and I
suspect they should have written responsiveness.....
I do not follow the logic that women whose children were hospitalized should report less responsiveness also in the antenatal phase?

Conclusion.
19. I am not certain about the conclusion, based on the article in its present version. The authors obviously share my doubt, as they describe it as relatively successful, and recommend its use after minor adaptations. What would these adaptations be? This is of high interest for future users of the questionnaire.

Line 444. Contribution to authorship does not correspond with the authors listed on the front page.

References
Line 485, the reference is partly wrong, partly lacking.

20. Abstract. The abstract should be revised in accordance with revisions in the main manuscript.

- Minor Essential Revisions
1. Figure 1: Question about food is still in, in spite of information on line 167.
2. The sample is called a cohort in the abstract, why is that?

Please give the table headings a new consideration, they are a bit scarce to my taste.

- Discretionary Revisions
In general, it benefits easy reading to use precise nouns, rather than replacing them with pronouns (although it may look tedious). Ex line 108, what is it that is "influenced by economic and political influences" and 109, what is "this concept" here? Any reader will understand after going back and read it once more, but rereading will be avoided if you go through the text and put exact nouns in when suitable.


Level of interest: An article whose findings are important to those with closely related research interests
Quality of written English: Needs some language corrections before being published

Statistical review: Yes, but I do not feel adequately qualified to assess the statistics.

Declaration of competing interests:

I work in a public, non-profit organization in Norway that develops and conducts patient surveys on this field and related fields. For the moment I am working on a similar paper about a user experience questionnaire for the same target group, but with a different approach and method.

My employing organization or myself may not gain or lose financially from the publication of this paper, now or in the future.

I declare that I have no competing interests.