Author’s response to reviews

Title: Searching for the optimal number of response alternatives for the distress scale of the Four Dimensional Symptom Questionnaire

Authors:

Jan van Bebber (jan.van-bebber@charite.de)
Johanna Wigman (j.t.w.wigman@umcg.nl)
Rob Meijer (r.r.meijer@rug.nl)
Berend Terluin (berend@terluin.net)
Sjoerd Sytema (s.sytema@umcg.nl)
Lex Wunderink (lex.wunderink@ggzfriesland.nl)

Version: 3 Date: 23 Oct 2018

Author’s response to reviews:

Reviewer reports second 4DSQ paper:

Farzaneh Saeedzadeh Sardahaee, M.D (Reviewer 1): This study reports novel findings that can be both useful and relevant. The manuscript is well referenced and easy to follow. Some editing is needed to make the manuscript more grammatically correct, formal and precise. To give a few examples, "(that is, measurement error)" in line 4 page 5, "various kinds of item" in line 7 page 5 or "does not pay off" in line 25-26 page 5 could be worded differently. Description of methods and results is well written and lays out the setting for later arguments.

This paper is recommended for publication after minor revisions. Please find below a list of points that could be considered by authors in their minor revision.

1- Participants in calibration sample (CS) and convergent validity sample (CV) include adolescents. It would be useful to inform readers whether the tests that were investigated in this study are generally used in adolescent population and if so, which populations, research sample or clinical, or in fact how these tests compare to testing in adults.

>> We are grateful for this suggestion. The calibration sample contains 46 adolescents (age < 18) and the convergent validity sample includes one adolescent only. An article that reports findings
on the measurement invariance of the 4DSQ scales with respect to age (adolescents, 10-17 versus ‘emerging adults’, 18-25) is currently in preparation. We may report that the scale scores are equivalent, although nine out of the total 50 items (all 4 scales) show negligible DIF. We decided to add this information to the paragraph Strengths and limitations in the discussion section of our manuscript.

2- It was not clear to me whether any of participants were excluded, and if so what were the exclusion criteria.

>> Prior to our analyses, the calibration sample contained respondents that had missing values (MVs) on some distress items (N = 776). We decided to remove these respondents because having no respondents with MVs simplified our IRT analyses, and we still had enough respondents to receive stable parameter estimates. We added this information to the paragraph Participants in the Methods section.

3- Sample size in CV sample (55 persons) does not warrant drawing conclusions applicable to larger samples. This could be mentioned as a limiting factor.

>> We are very grateful for this comment as well, and we decided note this limitation in the discussion section (paragraph Strengths and limitations).

Adrian Alacreu-Crespo, Ph.D. student (Reviewer 2): This manuscript evaluates two ways of coding the scale of distress of 4DSQ using the IRT and provide convergent, divergent and predictive validity to this two ways of code the scale. In my opinion the paper is novel, well writted and provide information usefull for the clinical practice. In this sense, I think that paper would be publish with a slight review. Following, I explain some issues that would be revise before publication:

1. First of all, please provide all the information of 4DSQ in instruments, also, make clear that you are going to analyze only the distress scale (provide statistics from past research as alpha’s), although appear in introduction please explain the two ways of coding, and explain what of the four measures of 4DSQ from MISS sample were used to evaluate predictive validity. It was baseline measure? One of the most important concerning things for me is, why you not evaluate the orecision from the other scales? This would improve paper a lot, and the other scales are more or less compatible to test convergent, divergent and predictive validity, althoug of course it
would be more interesting if there exist results with the HML or the BDI for depression scale or the STAI for Anxiety scale.

>> The reviewer is correct that we should provide more information on the psychometric properties of the 4DSQ distress scale in our article. We therefore decided to add a paragraph to the Methods section labeled Psychometric properties 4DSQ distress scale.

   We respectfully disagree with the reviewer that we would not make clear that this article solely focuses on the distress scale of the 4DSQ, because this information is already provided in the title of our manuscript. Nevertheless, we agree that the rationale for doing so is not stated clearly enough. In the Introduction section, we did state that “Most practitioners working with the 4DSQ found the distress scale most useful and important.” We added the following sentence to clear things up: “This, and the fact that the items of this scale are to be used in an adaptive online test battery (van Bebber et al., 2001) is also the reason why this article solely focuses on the distress scale of the 4DSQ.” In addition, we note under Directions for future research that it would be desirable to expand our research to the other three scales of the 4DSQ.

2. I don’t understand the rationale of use the Neuroticim scale of NEO-FFI as discriminant validity. Divergent validity implies that the construct evaluate (distress) not are relate with a construct that a priori should’nt be related with. In this regard, distress and neuroticism was relate in the past (eg. Ploubidis & Frangou, 2011) Moreover your results show relationships between this constructs. Then, that would be another measure of convergent validity?

>> We have to admit that the use of the Neuroticism scale may appear controversial, and the authors of this manuscript discussed this choice many times. Note that in our view, in order to access divergent validity, it would not make sense to use a scale that is a priori, based on theoretical considerations, totally unrelated to the measure under investigation (e.g. an intelligence scale in our case). The measure of choice for assessing divergent validity should tap a construct that is indeed related to the construct under investigation, but to a lesser degree than alternative measures of the same construct under investigation (those that are indeed used to assess the convergent validity of the measure).

3. In the BPL, why you discard the physical functioning statements? Can you provide the rationale in the instruments?

>> We decided to discard the physical functioning items because these seem not as relevant to psychological distress as the items that tap either general or relational (see Table A1 for reference) biographical problems. We added this motivation to the paragraph BPL in the Methods section.
4. With respect to UBOS, Table 3 not show the short-term results. UBOS were asked only in the last follow up? Please explain it in the instruments section.

>> We are grateful that the reviewer mentions this omission. We decided to describe the data collection design more thoroughly in an additional paragraph labeled Measures, measurements, and types of scale scores in the Methods section.

5. In page 12 line 10 you said Table A1 and Table A2 to refer the item score frequencies from first (0-2 coding) and second (0-4) coding; but you commit a mistake, is Table A2 and Table A3. Regarding this tables, it is one of the most important results in the paper, I think that this tables would appear in the results section and not in appendix.

>> We are grateful that the reviewer discovered this mistake in our manuscript and thus corrected the flaw. And we also agree that the content of these tables is of such an importance, that these should be incorporated in the main body of the manuscript as the first two tables.

6. Regarding convergent-validity section (page 13) please provide p values. Furthermore, take into account the coment 2 for interpret this section.

>> We added significance levels to Table 1. We already replied on the suggestion of the reviewer in our response to the second comment.

7. Regarding predictive validity, I have some concerns and proposals. First of all, is the first time in paper when short-term/long-term labels appear, and it people would be lost when they read it. I conclude that short-term prediction speaks about the six-month follow up and long-term about the twelve-month follow up. Then, I think that this issue would be fixed describing more this differentiation in the participants description. As I said in coment 1, what measure of 4DSQ was utilized for perform predictive validity? That would appear before. Another issue, did you compared 4DSQ with the two-month follow up measures? Moreover, as you explained in participants section this participants have elevate levels of distress, then this participants are distressed from the begining and there are not participants without stress. In this sense it would be useful first if you provide the descriptive information from the baseline and the follow ups of the different test. Moreover, it would be interesting calculate the changes in distress and the other scales (follow up - Baseline) and see if distress decline corresponds with decline in the other scales.
a) We are very grateful that the reviewer points to this omission, and thus added a few sentences that clarify i) what we mean by short-term and long-term, and ii) when which measures were assessed just before the paragraph Item Response Theory.

b) We added a table with all the requested descriptive statistics as Table A4. More specifically, Mean, SD, MIN, and MAX for levels of distress at t0 (baseline), t1 (after six month; short-term), and t2 (after twelve months; long-term), Mean, SD, MIN, and MAX for BIOPRO at t1 and t2, and Mean, SD, MIN, and MAX for the UBOS scales at t2.

However, we decided not to compute change scores between the measurement moments, because we were not able to compute change scores (baseline versus t1/t2) for all outcome variables.

8. Regarding the interpretation of results, in the abstract and the conclusion in page 17 you conclude that recoding should be avoided because decrease the precision of test. However, your results not provide complete evidence for say it. In one hand, the model fit for the three items alternative is better than the five items. But, on the other hand, the "sometimes" (that is the 1 in the 3 items model) category is redundant, and three items show less predictive validity in a couple of scales compared to 5-items model. In this regard, that seems mixing results to me, and seems that 3-items is also useful. I’m agree with the recommendation, but the coment in the abstract is to heavy for me, and I propose authors say that is a recomendation instead of should be that is more imperative.

We agree with the reviewer that the results we found in our study are mixed. What we do find is that from a psychometric point of view, rather the middle response option ‘Sometimes’ seems redundant, and that collapsing the three highest response options has a substantive detrimental effect on the measurement precision for elevated levels of distress. That is, we think we have a strong case for advising against the current practice of recoding. It might be that the scale would function even better in case answer-sets with four (or alternatively three) distinct response categories would be utilized. We tried to clarify this distinction by adding the following sentence to the Strengths and limitations paragraph in the discussion section: “The main limitation of this study was that the data of the three-point Likert scale were not obtained using three response alternatives. Thus, we cannot state that using the original five response alternatives is the best way to collect data for the distress items of the 4DSQ in general.”

9. In the discussion you speak about the convergent, discriminant and predictive validity, but you not explain anything about the direction of relationships and the interest of this relationships in the live of patients, it would be interesting interpret this relationships in the discussion.
We are grateful that the reviewers point this out. We decided to explicate our expectations about the directions of the relationships at the end of the Methods section of our manuscript: “For both scale scores, those computed from the original five-point item response scale (0-4) as well as those computed from the recoded three-point item response scale, we expected positive relationships with all criterion measures (GHQ MQ, NEO-FFI, BPL, UBOS, & sick leave), indicating that higher levels of distress correspond to higher scores on the criterion measures.

For all criterion measures except from the NEO-FFI, which we use as an indicator of discriminant validity, higher values of correlation coefficients indicate the more valid scale scores. For the NEO-FFI, a lower correlation is indicative of the more valid scale scores.”

However, we respectfully disagree with the reviewer that we should interpret/explicate the meaning of these measures in the lives of our patients, because a comparison of two alternative scoring schemes is the core of our article, and not so much the relevance of these measures. This is not to say that this would not be relevant as such.