Author's response to reviews

Title: A proposed method to investigate reliability throughout a questionnaire

Authors:

Tore Wentzel-Larsen (tore.wentzel-larsen@r-bup.no)
Tone M Norekvål (tone.norekval@helse-bergen.no)
Bjørg Ulvik (bjorg.ulvik@hib.no)
Ottar Nygård (ottar.nygard@helse-bergen.no)
Are H Pripp (are.hugo.pripp@oslo-universitetssykehus.no)

Version: 2 Date: 22 June 2011

Author's response to reviews: see over
Response to comments by George Ploubidis:

1) The authors employ PCA and classify participants depending on the estimated primary component score. PCA reduces observed items into dimensions and is not a classification technique, although it may have been erroneously used as such. Finite mixture models or Latent Class Analysis would be suitable alternatives in order to classify the participants based on the continuous component derived from PCA.

*PCA have been used for classification purposes in other fields, but after closer examination we agree with the reviewer. Finite mixture models and latent class analysis were assessed, but we found cluster analysis most suitable in our work. The clara function in the R package cluster, clustered the respondents into two groups.*

2) It is very unclear what the authors mean on the bottom of page 4 and this is a crucial point in the manuscript. The authors state: “our proposed method is to estimate ICC for each item between subjects unsupervised classified into groups using PCA”. It is not clear why the groups derived by PCA (which is the wrong method in this instance) are relevant, especially since later in the manuscript it becomes apparent that the cut off point for the classification was not derived empirically.

*PCA is not used as cluster method in the revised version of the manuscript. Instead, we use cluster analysis. Our hypothesis was that random error could differ throughout a scale. Respondents are clustered into two groups based on all the items in the scale. If random error increases throughout the scale, the effect of clustering should on average be lower for items with high degree of random error compared with items with a low degree. Estimating the intraclass correlation coefficient between subjects unsupervised classified into groups assessed the effect of clustering. It was arbitrary chosen to use two clusters.*

3) From the description of the scales in the empirical part of the paper, it becomes clear that these are scored on Likert scale (i.e the items generate ordinal level data). Was this taken into account when the PCS was performed (easy to do with polychoric correlations)?

*It is not relevant in the revised version since cluster analysis is used instead of PCA.*

4) The authors attempt to quantify random error, but fatigue due to questionnaire length is clearly a systematic process. Latest developments within the generalised latent variable modelling framework offer attractive alternatives, such as bifactor multmethod models for example, to the method proposed here. To what extend the proposed method imporoves the existing ones?

*Other effects due to questionnaire length besides random error is discussed in the revised version of the manuscript. We provide reference to a work by Krosnick. Randomly choosing among the response alternatives offered is one of the effects of satisficing. We focused on this, but are aware that fatigue (and thus satisficing) due to questionnaire length may have other effects. However,*
Simulation studies indicated that our proposed method could detect change in random error (and reliability) given that such exists.

Quality of written English: Needs some language corrections before being published

The language is revised.

Response to comments by Aaron Maitland

The authors are posing an important research question. I was very interested in their proposed method for evaluating the reliability of a questionnaire in the absence of test-retest reliability. These designs are often quite expensive or difficult to implement in a practical application. However, I feel that the article has a few weaknesses that need to be addressed.

The theoretical motivation for the paper can be supported more thoroughly from the literature. For example, the author might want to draw on literature by Krosnick (1991) that outlines a theory of satisficing for survey response. It is consistent with the authors’ view that as motivation decreases survey error might increases. However, it is not clear to me that the existing literature supports the view that fatigue necessarily leads to increased random error. In fact, it may be much more systematic. The mixed results led me to question this view. It is possible that one might see more systematic (albeit invalid) variability if respondents resort to strategies like acquiescence bias, nondifferentiation, or extreme response styles.

We agree that the effect could also be more systematic, but have in this work focused on change in random error (and reliability). The work by Krosnick is discussed in the revised version of the manuscript. Krosnick put forward that randomly choosing among the response alternatives offered could be one of the effect of satisficing. Our proposed method targets this effect.

As I understand it in the paper the awareness measure is based on the following four steps: 1) Principle Components Analysis (PCA) of the items 2) Mixed effects modeling with the groups form PCA as the random between groups factor 3) computation of the ICC for each item based on the mixed effects model and 4) linear regression of the items on the ICC. I am unclear exactly how PCA was used to divide the sample into two groups. More specifically, it is not clear to me how one draws the interpretation that one group is positive and the other is negative. Without seeing some kind of empirical results, it is difficult to assess the validity of this assignment. In any event, this assignment is not done without error and this is not discussed in the paper. Steps 2-4 seem logical, but are completely dependent on the PCA results.

Cluster analysis is now used to divide the sample into two groups. It is a method to make subsets in the data. Please see other comments on the use of cluster analysis. The ICCs do not depend on which group is termed ‘positive’.

The principal method in the article is proposed as an alternative to test-retest reliability. The current approach in the article does not include convincing evidence to support this assertion. Preferably,
the authors should include both theoretical and empirical support. The data from the current article does not allow for an empirical evaluation, but the authors might want to show in more detail how the ICC as calculated in the article might be similar to a test-retest coefficient. Otherwise the authors might want to soften the language about this being an alternative for test-retest reliability.

*Since it is not possible to do an empirical evaluation using our data, the language on assessment of test-retest reliability is softened as suggested by the reviewer.*

The authors need to do a better job of defining parameters in their equations. For example, not all of the parameters in the simulation study are well defined. It took a couple of reads for me to understand what each meant. The text should have parentheticals inserted appropriately with adequate explanations of the parameters. For example, “A basic assumption was existence of an unknown underlying factor (fi) partially determining the questionnaire items considered. A unique source of variance (eit)...”

*The explanation of parameters is revised.*