Author's response to reviews

Title: Exploring Differential Item Functioning in the SF-36 by demographic, clinical, psychological and social factors in an osteoarthritis population.

Authors:

   Beth Pollard (bsp@st-andrews.ac.uk)
   Marie Johnston (m.johnston@abdn.ac.uk)
   Diane Dixon (Diane.dixon@strath.ac.uk)

Version: 3 Date: 7 November 2013

Author's response to reviews: see over
Reviewer's report

Title: Exploring Differential Item Functioning in the SF-36 by demographic, clinical, psychological and social factors in an osteoarthritis population.

Version: 2 Date: 1 August 2013

Reviewer: M John G Bankart

Reviewer's report:

Major compulsory revisions

a1. Statistical analysis section. I agree with the logic regarding assessing unidimensionality using large differences between the first and second factors, and small differences between the remaining factors, but no specific rule / method was mentioned. The authors state: “Unidimensionality was supported if there were large difference in eigenvalues between factor 1 to 2 and small difference in eigenvalues between 2 and 3”. The authors should quantify precisely what they mean by ‘large’ and ‘small’. Was there a formal rule for establishing the differences in proportion of variance between factor 1 and factor 2, and between factor 2 and factor 3, which were necessary in order to conclude that there was in fact a state of ‘unidimensionality’ pertaining to a particular factor? If so, the authors should state what this rule was.

We have added our criteria for assessing unidimensionality into the statistical analysis section

‘A widely cited criteria for acceptable unidimensionality is if >=20% variance is explained in first factor (Reckase, 1979). Another commonly reported method for exploring acceptable unidimensionality is by looking at the ratio of the first to second eigenvalue. The first eigenvalue should be significantly higher than the second eigenvalue. If this is 3:1 or 4:1 then there would appear to be a dominant first factor (Anderson, 1994). If both of these criteria were reached unidimensionality was accepted.

a2. Effect of covariates sub-section. As this is an observational study, you cannot assume that covariates other than age were not different between demographic groups such as males and females. Was it possible to check this? You may have adjusted for age, but any DIF effects found could be due to imbalances in other covariates between groups. You should state that you have not ruled out this possibility.

We have added this point into the discussion

‘Additionally although we adjusted for age, it is possible that DIF effects could be due to differences in other covariates between groups and it is possible that there are not real differences in the underlying response probabilities’
a3. Conclusion section, 1st sentence. The authors suggest that they have uncovered DIF items. However, they may not have truly established the presence of DIFs. Not only have they not demonstrated that they have made appropriate adjustments for confounding variables, they have also not demonstrated that between any of the groups they are comparing, there really exist different underlying response probabilities. The authors should acknowledge this fact.

We have added this point into the discussion - see above

Minor essential revisions

b1. Background, paragraph 3. The authors give an example of DIF (measurement bias) in order to illustrate the concept. However, the association between depression and crying may well be different for men and women, as it is traditionally considered to be more socially acceptable for women to cry, and hormones are different in the two groups as well. On this basis, women probably do cry more than men, while controlling for emotional state, although this will be culture dependent to an extent. An item displays DIF only if people from different groups with the same underlying true ability / tendency have a different probability of giving a certain response. With respect to the example given, the two groups probably have different mean underlying tendencies with respect to crying when depressed. The authors should give a different example, where it is less likely that there are real differences between the groups.

We feel that the item about crying illustrates the problem of DIF for the reasons that the reviewer suggests, men and women may respond differently to this item even though they have the same underlying true levels of depression. This individual item is biased for measuring the construct of depression and as such should not be used in a measure of depression. Hence, we have re-phrased this section to make it clearer as follows

‘For example, it has been shown that for the Centre of Epidemiology Scale of Depression (CES-D), women are more likely than men to endorse an item about having crying spells even though they have the same underlying level of depression [2]. Thus, while this item may truly reflect differences between men and women in likelihood of crying, it exhibits gender bias with respect to the measurement of depression and scores for women might be inflated compared to men. Hence, apparent group differences in depression scores may be due to measurement bias rather than true differences’.

b2. Grouping factors section. The contrast between BMI categories, comparing <30 v >30 is currently labelled as normal / overweight v obese. This should be re-labelled as: underweight / normal / overweight V obese. Also, it should be <30 v #30 (or 30+).

We have changed this as suggested
b3. Grouping factors section. The authors state that they used bivariate median splits to dichotomise continuous and ordinal variables such as age. This is potentially problematic, as it can lead to an increase in either type I or type II errors, as well as being associated with other problems, such as inability to assess linearity / departures from linearity (Royston et al, 2006, Stats in Medicine). The authors should acknowledge the weaknesses of using such artificially / statistically derived dichotomisations in creating their categorical variables.

We have added this point into the discussion

‘We created some groups by using median splits and it is possible that other splits may have produced different results.’

b4. Statistical analysis section. Testing assumptions. Why was principal components analysis rather than common factor analysis carried out (why was it deemed appropriate to use all of the variance rather than just the common variance ?) The reasons for this should be made explicit.

We thank the reviewer for querying our use of PCA and agree that common factor analysis would be more appropriate. We have re-run the analysis and changed Table 2 (no changes to conclusions were found).

b5. Ordinal logistic regression model section. The authors should explicitly state what the dependent variable was.

We have clarified this in DIF testing section (ai)

b6. Measures section. Can you state which version of the SF-36 you considered in this study?

We have added that we used the UK version to the measures section

b7. Related to the above point, the authors have not mentioned the situation where sub-groups report the same level of response, but the underlying response levels are in fact different. A major issue concerns how to establish that there are in fact real differences (or not) between sub-groups. This is not discussed in any detail by the authors.

We have added this point into the background section

‘Alternatively, where no group differences are found, if DIF items exist then they might mask true group differences’.

b8. Table 1 (participant characteristics) : this table contains proportions and other information. The proportions should have the associated number in brackets after the %. The authors need to make it clear what the other information means. So for example, for Age (years), is this mean age (SD) ? BMI is given to 2 decimal places, whereas information for other numeric variables is given to 1 decimal place.
We have amended the table as suggested and corrected all to 1 decimal place.

Table 2. The abbreviations ‘var’ and ‘eigenval’ should be written out in full.

We have amended the table as suggested.

Discretionary revisions

c1. The authors might mention that selecting an optimal number of factors (in factor analysis) can also be done using parallel analysis (Horn, 1965). This method has gained in popularity in recent years.

We thank the reviewer for the suggestion but as we did not use this method we have not added it in, but will explore this option in future research.

Reviewer’s report

Title: Exploring Differential Item Functioning in the SF-36 by demographic, clinical, psychological and social factors in an osteoarthritis population.

Version: 2 Date: 18 September 2013

Reviewer: Joel Coste

Reviewer’s report:

Major Compulsory Revisions

The reviewer recommends the authors to clearly state if the main point of the paper is methodological or application to measuring PROs in OA.

We have clarified this by adding to the abstract that the paper was an application of the DIF method to a PRO.

Regarding methodology, the various techniques used in the paper overwhelm the reader a little bit. Rasch model analysis (eg partial credit models) which allows formally testing both undimensionality and differential item functioning (either uniform or non uniform) would be much more elegant in this case. OLR is no less complex and problematical than IRT or Rasch models (see the statistical literature, very critical of OLR). The fact that IRT “requires good model fit” is clearly an advantage over OLR in the reviewer’s mind.

We have added the use of Rasch analysis as an alternative into the discussion where we discuss the alternative method of using IRT models.

Regarding application, the authors may list the hypotheses that deserve attention in the context of clinical research on OA (DIF for sex, joint, side etc) and their implication for design (eg stratification) or analysis of studies. The authors may also discuss more comprehensively the literature on the SF-36 (and on other quality of life measures) behaviour when applied to patients with OA or other
rheumatic or painful conditions.

We have added the highlighted sentence to the discussion.

‘In the study we took the approach of removing the DIF items. However removing items may affect content validity of the measure and comparability with other studies. Using more complex Item Response Theory-based analyses, DIF items do not need to be removed as adjusted scores can be calculated for each subgroup. Alternatively, researchers may choose to stratify by gender, age etc. in the design or analysis of studies using the SF36. If the measure is in development, an alternative to deleting the DIF items, may be to substitute similar but DIF-free items either by re-writing, or choosing an alternative item with similar item properties. Re-writing could be facilitated by the identification of the source of DIF, for example by cognitive interviewing or by reviewing the item by groups of experts.’