Author’s response to reviews

**Title:** Psychometric properties of the Iranian version of 12-item Short Form Health Survey (SF-12): A population-based study

**Authors:**

Ali Montazeri (montazeri@acecr.ac.ir)
Mariam Vahdaninia (vahdaninia@acecr.ac.ir)
Sayed Javad Mousavi (jmousavi@razi.tums.ac.ir)
Sepideh Omidvari (somidvari@ihsr.ac.ir)

**Version:** 2  **Date:** 21 July 2009

**Author's response to reviews:** see over
Dear Dr. Alexandersson

Thank you for your kind e-mail. We found the reviewers’ comments very useful. Thus we have revised the manuscript to comply with their recommendations. Please find the following point-by-point responses as requested:

Reviewer: Dr. Lynette L-Y Lim

Minor essential revisions

Abstract, Results, 3rd sentence: PCS and MCS should be PCS12 and MCS12
Done.

Methods, Translation, last sentence: The actual method used to produce the PCS12 & MCS12 scores should be described.
This was further expanded as recommended:
However to calculate PCS-12 and MCS-12 scores we used the QualityMetric Health Outcomes Scoring Software 2. The software uses all the 12 items to produce scores for the PCS-12 and the MCS-12 and applies a norm-based scoring algorithm empirically derived from the data of a US general population survey [ref ware 1995]. It has been recommended that US-derived summary scores, that assume a mean of 50 and a standard deviation (SD) of 10 for each summary score, be used in order to facilitate cross-cultural comparison of results [7,20].

Discussion, para 2: In the Introduction, the authors stated that the SF-12 was developed for use in large health studies. Questionnaires in such studies are often self-administered. In their Discussion, the authors should stress as a limitation that their findings might not hold when the instrument is used in a self-administered mode. Some self-administered applications of the SF-12 have found very high incompletion rates (e.g. Lim L & Fisher J, Qual Life Res (1999); 8:1-8.)
This was acknowledged as recommend:
Face-to-face administration of the questionnaire allowed the interviewers to collect data without any missing data. Some self-administered applications of the SF-12 have found very high incompletion rates [e.g. see ] (e.g. Lim L & Fisher J, Qual Life Res (1999); 8:1-8.). However, one should note that the findings from this study might not hold when the instrument is used in a self-administered mode.

Discussion, para 1, 2nd sentence: The authors should clarify exactly how their results have shown that the SF-12 can be used for cross-cultural quality of life comparisons.
This was further explained as suggested:
Since the present study used the norm-based scoring algorithms for calculating the PCS-12 and the MCS-12 scores, the results from this study also can be used for cross-cultural quality of life comparisons.

Discussion, para 3, 3rd sentence: The statement “No floor or ceiling effects were observed….” conflicts with the results in Table 2 reporting significant ceiling effects for 5 scales and floor effects for RP and RE.
Thank you for your attention. Table 2 now is corrected. As it can be seen there were no floor or ceiling effects for the PCS-12 and the MCS-12. This also was corrected in the Discussion.

Discussion, para 4, 6th sentence: The authors reported that vitality has been found in other studies to correlate higher with PCS than MCS. They should follow up this observation with its relevance for the Iranian setting.
This was revised and further explained to comply with the recommendation:
As expected the PF, RP, BP and GH subscales correlated higher with the PCS-12 score while the VT, SF, RE and MH more correlated with the MCS-12 score (Table 4). This finding is somewhat different from those that were reported by the Ware et al. where physical functioning, role physical and bodily pain correlated most highly with the PCS and mental health, role emotional, and social functioning correlated most highly with MCS; and vitality, general health and social functioning had a relatively high correlation with both components [9]. However, a number of studies have shown that vitality item has appeared to correlate higher with PCS than with MCS score [24]. It is argued this might be due to cultural differences among people from different countries or simply this might be occurred due to translation problems [26,35]. In addition, it has been reported that even translation of concepts such as social functioning could be difficult in some Asian cultures [e.g. see 6]. As Ware indicates the most important empirical point that should be noted is the fact that scales that load highest on the physical component are most responsive to treatment that change physical morbidity whereas scales loading highest on the mental component respond to drugs and therapies that target mental health [36].

Conclusion, 2nd sentence: "... would minimize problems with multiple health-related quality of life scales" – the authors need to clarify what problems are being minimized.

The conclusion was revised both in Abstract and Discussion:
In general the findings suggest that the SF-12 is a reliable and valid measure of health related quality of life among Iranian population. However, further studies are needed to establish stronger psychometric properties for this alternative form of the SF-36 Health Survey in Iran.

General: There are several typos throughout the text that should be corrected.
Thank you. These were corrected.

Table 3: The abbreviations PF, PF1, PF2 etc are very cryptic. Sufficient information should be provided to enable readers unfamiliar with the SF-12 & SF-36 to understand this table.
This was provided.

Table 3, footnote: Please give justifications for the correlations classified as strong, moderate and weak.
This was clarified in the text (under the subheading Statistical analysis) and reference was provided. Footnote for Table 3 also was revised accordingly:
Correlation values of 0.40 or above were considered satisfactory (correlations ≥ 0.81–1.0 as excellent, 0.61–0.80 very good, 0.41–0.60 good, 0.21–0.40 fair, and 0–0.20 poor) [29].

Table 5: The p-values should be from a trend test instead of the non-specific test from a one-way ANOVA
Linear trend test was performed.

Discretionary revisions
Abstract, Results, last sentence: The statement "... strong correlations were observed..." is rather vague. The authors could give a more precise statement of their findings.
This was revised as recommended:
In addition, correlations between the SF-12 scales and single items showed that the physical functioning, role physical, bodily pain and general health subscales correlated higher with the PCS-12 score, while the vitality, social functioning, role emotional and mental health subscales more correlated with the MCS-12 score lending support to its good convergent validity.

Methods, The questionnaire: It would help readers unfamiliar with the SF-12 & SF-36 to have the statements and responses of the 12 items spelt out.
This was revised as suggested:
The 12-item Short Form Health Survey (SF-12) is a shorter alternative of the SF-36 instrument that includes 12 questions and 8 scales: physical functioning (PF-2 items...
on limitations doing moderate activities and climbing several flights of stairs), role limitations due to physical problems (RP-2 items on less accomplishment than one would like to achieve and limitation in kind of work or other activities), bodily pain (BP-1 item on pain interference with one’s normal work), general health (GH-1 item on general health perception), vitality (VT-1 item on having energy), social functioning (SF-1 item on interference of physical health or emotional problems with one’s social activities), role limitations due to emotional problems (RE-2 items on less accomplishment than one would like to achieve and not being careful in doing activities as usual) and perceived mental health (MH-2 items on feeling calm or peaceful and feeling sad or blue). Response categories for items vary from 2- to 6-point scales and raw scores for items are ranging from 1 to 6.

Reviewer: Dr. Ulf Jakobsson
An interesting paper with focus on psychometric evaluation of the SF-12 instrument. The study have a sufficient number of respondent included in the study and the sample seems fully adequately chosen. However, the authors only perform some of the possible analyses that should be conducted to fully evaluate the instruments psychometric properties. For example, to only use Cronbach’s alpha to assess reliability do not give any firm knowledge about the instruments reliability. Furthermore, the authors have only evaluated construct validity in the study and all other types of validity measures (such as face validity, content validity, criterion validity) is not considered/discussed at all. I suggest that e.g. corrected item-total correlations, analyses of discriminant and convergent validity, analyses of known-group validity between other groups than only age and gender are added (and maybe also add a discussion about face & content validity).

We have tried to comply with recommendations as indicated in the following point-by-point responses.

Major Compulsory Revisions
(1) The background is too sparsely written and do not have any argumentation for the need of this study. Furthermore, there is no critical discussion about the results from previous studies, for example the authors state that the instrument is a valid and reliable measure, but this is can be questioned. All results from previous studies do not indicate this.

This was further expanded and new references were provided.

(2) In the method section, (a) the instrument is too sparsely described. More information about e.g. the scoring method and development is needed. (b) Furthermore, the paragraphs under the heading “The questionnaire” and “Translation” should be merged.

a. This was further explained. Also development of the questionnaire was further explained in the Background.

b. Done. However, this was changed to: The questionnaire and scoring.

(3) Why an explorative factor analysis instead of a confirmatory factor analysis is used is unclear (the factor structure of the SF-12 is well-known). The authors need to more explicit explain and discuss the reason to choose an EFA in favor to an CFA. Furthermore, more information is needed about how the EFA was done (including extraction method and how various “goodness-of-fit tests” was applied). To be noticed, the measurement model (both SF-12 & SF-36) is somewhat complex with cross-loadings which is hard to identify with an EFA.

Our intention was to publish another paper with CFA analysis. However, to comply with the recommendation we included this information in the present paper.

(4) Table 3. Figures regarding eigenvalues, communalities, and explained variance are missing.

These are now provided in a separate table as requested (now Table 5a).
The use of Pearson’s correlation do not seem appropriate to use when correlating various item with the summary scores since the items are nominal and ordinal data. A corrected item-total correlation should be used instead (based on appropriate correlation matrix).

To comply with recommendation now a new table (Table 4) containing item-scale correlation corrected for overlaps was provided and instead of Pearson's correlation Spearman's correlation was used.

(6) The discussion is in general too weak in the discussion of the results as well as the methodology and study limitations. It should also more thoroughly discuss the limitations related to the chosen methods for psychometric evaluation. Why only these methods used and what are the consequences of this.

The Discussion was reorganized and further improved as suggested.

(7) The conclusions (both in the discussion section and in the conclusion) about that the instrument is found to be valid and reliable needs to be revised because the results from the study do not fully support the statements made by the authors (e.g. the floor & ceilings affect is not acceptable).

The conclusion was revised both in Abstract and Discussion sections:

In general the findings suggest that the SF-12 is a reliable and valid measure of health related quality of life among Iranian population. However, further studies are needed to establish stronger psychometric properties for this alternative form of the SF-36 Health Survey in Iran.

Minor Essential revisions

(8) The analyses of floor and ceiling effect is not mentioned in the method section and should be inserted under the subheading Statistical analysis.

The first sentence under subheading Statistical analysis was revised to comply with recommendation:

In addition to descriptive statistics (including floor and ceiling effects), according to International Quality of Life Assessment (IQOLA) Project to assess the psychometric properties of the Iranian version of SF-12 several tests were performed.

(9) In the discussion section, the authors state that there is no problem with floor and ceiling effect. However, this is not supported by the results.

This was corrected.

(10) Table 2 & 4 & 5, the mean values should not be used for the subscales (due to several subscales is only based on one item and all items in SF-12 is either nominal or ordinal data).

The mean values for subscales were deleted as suggested.

(11) Table 4 & 5, the Student t-test and ANOVA should not be used for comparisons regarding the subscales (due to several subscales is only based on one item and all items in SF-12 is either nominal or ordinal data).

To comply with recommendation Table 4 and 5 now were changed to Table 3 and only contains comparison of summary scores between males and females and among different age groups and people with different educational status.

(12) I think that the authors should be a little cautious with the statements about that the results could serve as normative data (for the entire Iranian population).

The sentence was revised to comply with recommendation:

…. the result of this study might be considered as Iranian normative data for the 12-item Short Form Health Survey (SF-12) and perhaps could be used as a basis for comparison with specific populations in the future studies.

Also please see the following statement that follows the above sentence:

However one might argue that a sample from the urban capital is not necessarily representative of the entire country. In general this is true but since Tehran has become a multicultural metropolitan area it has been suggested that a sample from the general population in Tehran at least could be regarded as a representative sample of urban population in Iran [26].

Discretionary Revisions
Based on that the psychometric evaluation of the instrument is limited to only construct validity, the authors may consider changing the title in accordance to this. If the authors chose to not change the title, then consider removing “a population-based study” in the title since it do not add any important information.

The title was changed to:
The Iranian version of 12-item Short Form Health Survey (SF-12): factor structure, reliability and construct validity

Readers will benefit from a revision of the layout in the tables (so they will be easier to read, all necessary information is included, and the data is presented properly).

The layouts of tables were revised as suggested.

Again we are very grateful to you and both reviewers and hope you find the corrections satisfactory.

Kind regards
Ali Montazeri