Reviewer's report

Title: Health related quality of life measured by SF-36: a population-based study in Shanghai, China

Version: 2 Date: 30 April 2008

Reviewer: Giovanni Apolone

Reviewer's report:

The paper describes the results of a cross-sectional survey carried out in a random, representative sample of Chinese people living in Shanghai to collect information about the reliability, psychometric and correlation validity of the Mandarin Chinese version of the SF-36 Health Survey, previously translated by an independent team.

Operationally, in the context of a vis-a-vis interview, people filled in the questionnary (self-administration) and reported information about socio-demographic (sex, age, schooling, etc) and medical (chronic diseases) variables. A subsample of cases also took part in a test-retest exercise. Standard psychometric and correlational analyses were carried out to test the relation between and within items and scales, as well as between scales scores and other external variables. Univariate and multivariable analyses were performed to describe and test the magnitude and statistical relevance of associations.

Results were compared with what expected and with other results from previous Chinese or internationals studies.

The study is sound, and the paper is well structured and reports and discusses in a fair way available evidence. Analysis carried out are those recommended. Findings may help implement better this questionnaire in the Chinese setting and also add valuable information for the international context. Nevertheless, a few points that hamper the paper quality should be considered.

Major compulsory revisions

1) As Authors have carried out this study with the primary purpose to test the feasibility and "actual" validity of this questionnaire in the real setting/world, it is quite disturbing that they did not perform any kind of analysis to evaluate the frequency and the pattern of the non-responder status. As a matter of fact the prevalence of non-responders was quite high (only 919 cases were actually used for the final analysis, yielding a 23% proportion of non-responders). This fact reinforce the need to describe the comparative difference between responders and non-responders and to identify the risk factors of the non-responder status. In addition, given the fact that it was possible to help/assist people, this event also should be quantified and evaluated as it is well known that people asking/receiving help are a quite peculiar sub-sample and help may distort results. An analysis of the pattern of missing values at item level may also help
identify critical/disturbing items.

2) Results document that some scales had quite low reliability estimates, such as the SF. Although in other countries and languages this fact was already documented, it requires a deeper analysis and discussion as other findings, such as the low performance of the MH scale (one of the best performer in other settings), suggest that there might be some problems in the very conceptualization of mental health in this sample.

3) Some multivariable analysis using as dependent variable the GH scale (that has a generic meaning and could be related to either a physical or mental concept of health) and as independent factors the other scales could improve the understanding (better than the factors analysis) of the complex relation between physical and mental scales and global health.

4) Somewhat disturbing is the fact that female in this sample reported almost always higher scores than males, across relevant subgroups. Despite the fact that multivariable analysis reduced the magnitude of this phenomenon, this fact should be carefully studied and discussed as in other settings it was observed the opposite relation.

5) The poor global performance of SF is also very disturbing. Although noted in other settings, in this context together with other findings it suggests that something of the mandarin version of the SF-36 is wrong, at least in this sample. The scale is formed by two very similar questions and usually the item-item correlation is high and problem comes from convergence-discriminative analyses. In this case the item-item correlation is quite substantial (0.70), but the reliability coefficients are low and apparently no problems are evident in terms of correlations with items from other dimensions (range 0.110-0.504). Authors suggest that an explanation of the poor reliability may be the reverse orders of the two items. It could be the case but the high item-item correlation does not support completely this explanation.

Minor essentials

1) It is not clear what kind of chronic diseases were evaluated. Only physicals? Do Authors have some information about the presence of mental/psychologic problems in the sample studied?

2) Did Authors use the first or second release/version of the SF-36 HS? Did they use the RAND or IQOLA rules and algorithms?

3) Scores from scales are quite high, when compared to other studies. Authors should discuss these findings in a deeper way, as it could be another indicator of a difference between Chinese and non-Chinese “culture” and conceptualization of health.

4) Reporting and analyzing the summary scores (PCS and MCS) estimates overall and across groups may help understand some of the phenomena above identified.

5) Authors should add in the discussion a section discussing the study limitations due to the study design (cross-sectional nature, lack of longitudinal data,
presence of a distortion due to the type of administration/assistance, etc).

**Level of interest:** An article whose findings are important to those with closely related research interests

**Quality of written English:** Acceptable

**Statistical review:** No, the manuscript does not need to be seen by a statistician.

**Declaration of competing interests:**

'I declare that I have no competing interests'