

Author's response to reviews

Title: Further Validation of the Multidimensional Fatigue Inventory in a U.S. Adult Population Sample

Authors:

Jin-Mann S Lin (dwe3@cdc.gov)
Dana J Brimmer (dyv4@cdc.gov)
Elizabeth M Maloney (evm3@cdc.gov)
Ernestina Nyarko (gyh8@cdc.gov)
Rhonda BeLue (rz10@psu.edu)
William C Reeves (wcr1@cdc.gov)

Version: 2 **Date:** 18 October 2009

Author's response to reviews: see over

Please find our responses to the reviewers' comments itemized below. We added two additional references to the Reference section. Thank you for your positive response to our manuscript.

Reviewer #1

Reviewer's report:

This is an excellent article as is save for two possible non-essential criticisms.

DISCRETIONARY I believe that the article would require less time and energy to read if the negative correlations etc. were dropped and only the sig. or near sig. values were reported. This is partly a matter of style.

Reply: In the current version we only report the significant or near significant values in the tables; these include the significant negative correlations which result from correlations between SF-36 subscales (higher values indicate better functioning) and MFI-20 subscales (higher scores indicate worse (more) fatigue).

DISCRETIONARY I don't believe I saw much on external validity/ generalizability. I realize that this is a measurement article and generalizability isn't the point. Nevertheless the authors should address the fact that the subjects were limited to Georgia in the same study in other regions could be different. In other word the effect of regional lifestyle.

Reply: As the reviewer pointed out, this is a measurement article and technically generalizability is not the objective in this paper. However, we agree with the reviewer that our data limits the generalizability to other regions and we have added this limitation to the paper.

COMPULSORY Sample size was adequate but was the experiment-wise error addressed. It should be because there were so many statistical tests, thus the possibility of Type II errors. This is the only point of contention. If the authors can adequately counter this point I will be satisfied with that.

Reply: Thank you for pointing out the possibility of Type II errors due to multiple statistical tests. In our study, we applied a 0.01 alpha level of statistical significance to adjust for multiple testing instead of the popular standard level of 0.05. We acknowledge that our approach increases our confidence that we are not falsely rejecting a true null hypothesis (a Type I error), but increases the risk of failing to reject a false null hypothesis (a Type II error), and therefore reduces the statistical power. However, the statistically significant results observed in this study are of practical significance. For example, the group mean differences in our study are greater than the generic minimal clinically important difference (MCID) of 2 points in the MFI-20 subscales across the pre- and post-radiotherapy comparison and occupational productivity anchor in Purcell's study. The possibility of a Type II error should, however, be considered. We have acknowledged this as one of our study limitations.

Reviewer #2

Overview:

This is a nice report of a test of the measurement characteristics of the MFI in a population-based sample from the USA (state of GA). The authors are correct in introducing this as an important extension to prior uses of the instrument (typically for cancer patients). Special strengths of this manuscript include the large population-selected sample, careful clinical determination of groups, selection of comparison measures, report of reading levels of the instrument (page 9), and correction of p-values for multiple testing. It is a pleasure to read a report with these details. Current expectations of testing (validation, scaling, item function, etc) would recommend some additional analyses above the ones presented here. While this report is acceptable in terms of classical test theory, aspects such as item function based on Rasch methods, and examining the scale based on clinically meaningful differences would be important future steps. The authors might describe how the criteria were accepted from some review set (e.g., for disability; Andresen, 2000; COSMIN see Mokkink et al below).

Specific comments:

1. For correlations analyses, the p-value is relatively useless – it is very strongly related to sample size. The Table suggests the authors have used a correlation value of 0.50 and higher. Perhaps a range of levels would be useful (e.g., Andresen, 2000), allowing for a pattern of correlations to be described. Perhaps less text and more about patterns would accompany the detailed tables. Currently the text is hard to track because of the detail. Theoretically important relationships (the SF-36 vitality scale is the most obvious) would be the key to describe, as well as why one would expect others (e.g., mental health and/or physical health) to be higher, lower, etc. Basically, more overview on patterns of the text, but reduced text, would be easier to understand.

Reply: We thank the reviewer for the insightful discussion and references on how to improve the flow of the paper. We have made changes to the text to reflect a better flow pattern and have added to the Discussion section as well.

2. Additional analyses that meet current expectations of testing might include item response theory (Rasch), floor/ceiling effects, and clinically (and patient) important differences to understand if the group mean differences are “important.” Eventually, showing change in scores for treatment outcomes, or recovery would be important. The authors mention retest reliability (and that would be useful in the future), and these others might be added as future analyses/tests.

Reply: Thank you for this suggestion for additional analyses. We have added floor/ceiling effects in Table 2 and found no significant floor/ceiling effects in our study sample. Although our cross-sectional data prevent us from examining responsiveness and minimal clinically important difference (MCID) in fatiguing illness, we used the generic MCID of 2 points across the pre- and post-radiotherapy comparison and occupational productivity anchor [Purcell et al. 2009] to determine if the group mean differences in our study are important in fatiguing illness. Our

results demonstrated a desirable MCID (greater than 3 points) across the CFS-like, intermediate (Chronically Unwell), and Well groups. We have acknowledged the limitation of “clinically important differences” in the “Strengths and Limitations.”

3. The response report (page 6) is hard to decipher. Who are the 19,381 screened persons? If these are the “called” sample, then response would be only 32.4%.

Reply: We have clarified this confusion in the current paper – see paragraph below. “Based on the 19,381 persons from the household screening interview, 8,910 adults were randomly selected for detailed telephone interviews: 5,623 individuals completed the detailed telephone interview, 1,874 refused to participate, 141 were further confirmed to be ineligible, and 1,272 were excluded due to physical or mental inability to participate in the study, inability to be contacted, language barriers, or because they had died. This yielded an overall response rate of 75%.” [5623/(8910-141-1272)=75%].

4. The factor loading of 0.30 seems low – typically this is set at 0.40? As it is, the factor loading information is pretty scattered and hard to interpret (not the authors’ fault- this is how the instrument comes out). This is where Rasch might really help. Also, it is possible this is a single-score instrument, with less ability to have subscales anyway (based on less clear patterns of the reliability analysis also).

Reply: We thank the reviewer for pointing out this error and in fact the factor loading of 0.30 in text was a typo. In the old Table 5, we used the typical cut-off of 0.40. In the current version, we chose a higher cut-off of 0.50 so that the factor loading information is less scattered and easier to read. We agree with the reviewer that Rasch analysis might help for the patterns of constructs. However, the factor loading information is much clearer after using the desired cut-off of 0.50 for factor loadings (Table 5).

Editorial and presentation suggestions:

1. The discussions and analyses that compare men and women should be referred to as “gender” since they are not hormone-sex-based and the terms gender/men/women used consistently in the manuscript and tables.

Reply: The variable, sex, was the self-reported response to the question, “Are you female or male?” We therefore used sex as female or male consistently throughout the manuscript and tables.

2. Tables are easiest to understand when they “stand alone” meaning there are sufficient details in the title, variable labels, and footnotes to understand the population, analysis techniques, and results. The table that accompany this manuscript could use some editing for complete instrument names (or footnotes that spell them out), and for details (Table 3 would not be understandable for gender differences without reading and rereading the test).

Reply: We have added more details in the title, variable labels and footnotes to help readers to understand the population, analysis techniques, and results. The complete instrument names have been included in the tables.

3. Tables with descriptives do not need the details of percentiles scores, range, etc. What would be useful is % at extreme (floor or ceiling) levels.

Reply: In general, we agree with the reviewer that tables with descriptive statistics do not need the details of percentile scores, range, etc. However, only one study has reported norms of MFI-20 subscales, and it was for a German population. We sought to provide the details as age, sex-norms; especially group-specific norms for fatiguing illnesses for the U.S. population. However, we deleted the values for 90 percentiles and added percentages at extreme (floor or ceiling) levels (Table 2). We also described the floor/ceiling effects in the Results section. No significant floor/ceiling effects were detected in the whole study sample.

4. Were the cases with very sparse cells (e.g., missing marital status) used in the analysis (e.g., Table 1)?

Reply: Yes, we used cases with missing values since we did not hypothesize the associations of demographics with MFI subscales in this paper, except for age and sex for which we controlled in the analysis.

5. Please use consistent use of past tense, where appropriate. E.g., page 9, end of reading level analysis, "...respondents were (not are)...".

Reply: Thanks to the reviewer for pointing out the inconsistent use of past tense. We have changed "are" into "were".

6. What is the meaning of $|r|$ vs. r in correlations? The notation is not familiar.

Reply: $|r|$ means the absolute value of estimated correlation r and it was explained in the place where it first appeared. For most of the subscales except for SF-36 subscales, the higher scores indicate worse symptom or health status whereas the higher SF-36 subscale scores indicate better functioning status. To compare the magnitude of correlations between MFI and other subscales, we used the range of $|r|$ in text when negative correlations occurred. However, we reported the r correlations in Table 6.