# Measuring Fatigue in Parkinson's Disease:
# A Psychometric Study of Two Brief Generic Fatigue Questionnaires

Peter Hagell, RN PhD[1,2,3], Arja Höglund, RN BSc[4], Jan Reimer, RN[3], Brita Eriksson, RN[5], Ingmari Knutsson, RN[6], Håkan Widner, MD PhD[3], David Cella, PhD[7]

[1] Department of Health Sciences, Lund University, Lund, Sweden
[2] The Vårdal Institute, the Swedish Institute for Health Science, Lund University, Lund, Sweden
[3] Department of Neurology, University Hospital, Lund, Sweden
[4] Department of Neurology, Karolinska University Hospital Huddinge, Stockholm, Sweden
[5] Department of Neurological Rehabilitation, Arvid Carlsson Institute of Neuroscience, Sahlgrenska University Hospital, Göteborg, Sweden
[6] Divison of Neuroscience and Locomotion, Department of Neurology, Faculty of Health Sciences, Linköping, Sweden
[7] The Center on Outcomes, Research and Education (CORE), Evanston Northwestern Healthcare and Northwestern University, Evanston, IL, USA

Financial support:
The study was supported by the Swedish Parkinson Foundation, the Swedish Research Council, the World Federation of Neuroscience Nurses, the Skane County Council Research and Development Foundation, the Vårdal Institute, the Department of Nursing at Lund University, and Inge & Elsa Andersson's Parkinson's disease Research Foundation.

Corresponding author:
Peter Hagell
Division of Gerontology and Caring Sciences
Department of Health Sciences
Lund University
P.O. Box 157
SE-221 00 Lund
Sweden
*Tel*: +46 46 222 1937
*Fax*: +46 46 222 1934
*E-mail*: Peter.Hagell@med.lu.se

**ABSTRACT**

This study evaluated and compared the measurement properties of the 13-item FACIT-Fatigue Scale (FACIT-F) and the 9-item Fatigue Severity Scale (FSS) in 118 consecutive Parkinson's disease (PD) patients, using traditional and Rasch measurement methodologies. Both questionnaires exhibited excellent data quality and reliability (coefficient alpha ≥0.9), acceptable rating scale functionality, and discriminated between fatigued and non-fatigued patients. Factor and Rasch analyses provided general support for unidimensionality of both the FACIT-F and FSS, although they do not appear to measure identical aspects of fatigue. No signs of differential item functioning (DIF) were found for the FACIT-F whereas potential age DIF was detected for 2 FSS items. These results support the measurement validity of both questionnaires in PD, although the FACIT-F displayed better measurement precision and modest psychometric advantages over the FSS. Availability of psychometrically sound fatigue measures that are applicable across disorders provides a sound basis for advancing the understanding of this common and distressing complaint.

**Key words:** fatigue, Parkinson's disease, questionnaires, reliability, validity

Running title: Measuring Fatigue in PD

Fatigue can be defined as an overwhelming sense of tiredness, lack of energy, and feeling of exhaustion (1) and is a common complaint in a range of medical conditions, including many neurological disorders (1-3). In a clinical context, fatigue is considered to be a multidimensional concept with physical, emotional, cognitive, and social aspects. This can pose a challenge to measurement. In Parkinson's disease (PD), fatigue has been reported in 40-65% of patients and although many consider it to be one of their most disabling symptoms, it often remains undetected in clinical practice (4-8). Its cause remains unclear. For example, while some reports suggest an association between fatigue and the underlying parkinsonism (5, 9), others do not (10, 11).

One reason for such mixed results may relate to how fatigue has been assessed. Studies have tended to use fatigue scales from generic health status questionnaires or approaches not validated in PD (4-8, 10-12). Although generic health status questionnaires have been validated broadly, their subscales are brief and often lack sufficient detail for accurately measuring individuals. For example, the Energy subscale of the Nottingham Health Profile (NHP-EN) (13) has been found to be only a coarse measure of fatigue (14). Other employed instruments have been validated in other patient groups but remain untested in terms of their measurement properties in PD. This is a limitation because traditional psychometric properties are sample dependent. Thus their performance in specific applications is important to consider in the context of accumulated experience with an instrument (15). There is thus a need for fatigue measures with documented reliability and validity in specific patient populations, such as PD, that also allow for comparisons with other patient groups and healthy control populations (3).

We therefore sought to identify and validate an available fatigue questionnaire for use in PD. It was considered that such an instrument should have been successfully applied with well documented good measurement properties in various populations, and be brief and easy to incorporate into clinical research protocols without inducing substantial respondent burden (16). Availability of the instrument in several languages was also desired. Based on these criteria and a review of the literature, we undertook a validation of the 13-item Functional Assessment of Chronic Illness Therapy - Fatigue Scale (FACIT-F), which is part of the larger FACIT measurement system and currently is available in 46 languages (www.facit.org). The questionnaire was originally developed to assess anemia-associated fatigue (17), but has since been used and validated also in other patient groups, such as rheumatoid arthritis (RA) and various forms of cancer, as well as in the general United States population (18-20). In addition, it has been thoroughly documented regarding its responsiveness and minimally

important difference (MID) (19, 21, 22), and recent work (23) has also begun linking FACIT-F scores to levels of physical impairment and activity limitations, which has potential to further facilitate score interpretation. Furthermore, it can be administered via a range of modes, including interview and touch-screen computer administration, and novel modes such as computer-assisted telephone and web-based administration are under development (24).

Although it did not meet the above criteria regarding measurement properties, the Fatigue Severity Scale (FSS) was also considered due its relatively wide usage among people with PD (25) and the apparent lack of documented measurement properties in this disorder. The FSS was originally developed and tested for use in multiple sclerosis and systemic lupus erythematosus (26), and has also been evaluated in, e.g., postpoliomyelitis syndrome and chronic hepatitis C (27, 28). Here we assess the measurement properties of the FACIT-F, and compare it with the FSS, as applied to people with PD.

## METHODS

### *Design*

The study was designed as a cross-sectional, multi-center psychometric comparative study.

### *Patients*

One hundred and twenty four consecutive Swedish speaking patients with clinically diagnosed PD from four Swedish movement disorder clinics were invited to participate. Four patients declined participation, one did not have time to participate, and one was found not to meet inclusion criteria, leaving a total of 118 participants (n=30, 26, 30 and 32 from the respective study sites; no significant differences in dropout rates across study sites). Exclusion criteria were ongoing infections, psychiatric drug adverse reactions and clinically significant co-morbidities (including depression and cognitive impairment), as determined by patients' attending neurologist and the study assessor at the time of assessment. Patients participating in other ongoing studies were also excluded. All participating patients signed informed consent. The study was approved by the local research ethics committees. Patient characteristics are provided in Table 1.

### *Procedures*

Clinical assessments were performed by one experienced assessor (a PD specialized nurse) at each participating center. Before initiating data collection, all raters underwent

standardized video based training (29, 30) regarding clinical assessments according to the Unified PD Rating Scale (UPDRS) (31) and the Hoehn & Yahr staging of PD (32). This was followed by independently conducted ratings of patient video sequences where all assessors rated the same video sequences. Inter-rater concordance was assessed by means of Kendall's coefficient of concordance, which was ≥0.85 for both Hoehn & Yahr and UPDRS scores.

Patients were assessed clinically by means of parts I (mentation), III (motor score) and IV (complications of therapy) of the UPDRS, the Schwab & England activities of daily living scale, Hoehn & Yahr, and the Mini-Mental State Exam (33). All assessments were performed during the "on" phase (i.e., periods of good drug response and no or minimal PD-related disability). Hoehn & Yahr and Schwab & England were also estimated for the "off" phase (i.e., periods of poor drug response and increased PD-related disability) based on patient reported history and medical records. Patients then completed the fatigue questionnaires (see below) and questionnaires tapping sleep quality, daytime sleepiness, depression, anxiety, perceived adjustment to illness, and illness-related distress. Demographic data were collected by patient interview and from their medical records. Patients were also asked to complete a second copy of the FACIT-F at home one week later together with a question on whether their perceived level of fatigue had changed (according to a 5-grade scale, "much better" – "better" – "unchanged" – "worse" – "much worse") since the clinic assessment. Because the FACIT-F was the primary target instrument, and to minimize respondent burden and maximize retest response rates, the FSS was not included in this aspect of the protocol. Data regarding aspects beyond the measurement properties of the FACIT-F and FSS will be reported separately.

*Patient-reported fatigue questionnaires*

FACIT-F (17) consists of 13 items (Table 2) that are responded to by affirming one of five Likert-type response categories ("not at all" – "a little bit" – "somewhat" – "quite a bit" – "very much"). The instrument yields a summed total score ranging between 0 and 52 (52 = no fatigue).

The FSS (26) consists of 9 items (Table 2) and a 7-grade non-defined Likert-type scale anchored by "completely disagree" (=1) and "completely agree" (=7) at the respective ends. A total FSS score is calculated as the mean response across the 9 items, yielding a score range between 1 and 7 (7 = more fatigue).

The NHP-EN (13) was used to identify the presence of fatigue (5, 14). Patients who affirmed one or more of its three dichotomous ("yes"/"no") items (Table 2) were classified as fatigued.

All questionnaires had previously been translated into Swedish according to established standardized methods (34-36).

*Analyses*

The FACIT-F and FSS were evaluated regarding data quality, reliability, floor- and ceiling effects, construct validity, precision of scores, unidimensionality, rating scale functionality, and differential item functioning, using traditional and Rasch measurement methodologies (15, 37, 38).

*Data quality, reliability, floor- and ceiling effects*. Data quality relates to the usefulness of an instrument, and is considered high when the percentage of missing data is low.

Reliability was assessed by Cronbach's coefficient alpha, a measure of item interrelatedness and an estimate of reliability. The FACIT-F was also assessed for test-retest reliability by means of the intra-class correlation (ICC) coefficient, by comparing first and second administration scores among patients who reported stable fatigue. Reliability coefficients should not be below 0.7 and preferably >0.8 (37, 39).

Floor- and ceiling effects represent the percentage of respondents obtaining the lowest and highest possible raw scores, respectively. The threshold for acceptable floor- and ceiling effects was set at 15% (40).

*Construct validity and precision of scores*. Two aspects of construct validity, convergent and known-groups validity, were assessed. Convergent validity was evaluated by the correlation between scores on the NHP-EN and the FACIT-F and FSS. Strong correlation coefficients (≥0.6-0.7) were hypothesized. In evaluating known-groups validity, significantly ($P<0.05$) different FACIT-F and FSS scores between patients classified as fatigued and non-fatigued according to the NHP-EN were expected, and was interpreted as support for known-groups validity.

The precision by which the FACIT-F and FSS distinguished between these groups was explored by comparing the respective *t*-statistics, 95% confidence interval (95% CI) widths, and effect sizes (difference between means / full sample SD) following score transformation to 0-100 scales (100 = more fatigue). The effect size expresses differences in standard deviation units, where values of 0.2, 0.5 and 0.8 are regarded as small, moderate and large,

respectively (41). Using the respective *t*-statistics, the FACIT-F and FSS were also compared regarding their relative efficiency by calculating the squared *t*-statistics ratio using the smallest *t*-value as the denominator (42).

*Unidimensionality*. The extent to which items tap a single underlying latent construct, i.e., unidimensionality, is a basic assumption for the use of summed rating scales and was evaluated by two approaches. First, instruments were subjected to principal component exploratory factor analysis (EFA). The number of factors to extract was determined by parallel analysis (43). For both the FACIT-F and the FSS, 1000 sets of parallel random data were generated, followed by independent parallel EFAs of the empirical and random data matrices. For each consecutive empirical eigenvalue that exceeded the 95$^{th}$ percentile of the distribution of random data eigenvalues, a factor was retained (43). To support unidimensionality, the FACIT-F and FSS should render only one such factor each.

Second, the FACIT-F and FSS were subjected to analyses according to the Rasch rating scale model (38). According to this model, the probability of a person giving a certain response to an item is a logistic function of the difference between the level of the underlying construct represented by the item and that possessed by the respondent. The model yields separate measures for each person, item, response category, and transition point between categories on a common logit (log-odd units) metric, which measures at the interval level and ranges from minus infinity to plus infinity (with mean item difficulty set at zero). A fundamental Rasch model assumption is that each item contributes to the measurement of a single underlying latent construct. Unidimensionality was thus assessed by determining each item's information-weighted and outlier-sensitive goodness-of-fit (INFIT and OUTFIT, respectively) expressed as mean-square (MNSQ) and standardized statistics. MNSQ is the ratio between observed and predicted variance and has an expected value of 1. For polytomous rating scales, MNSQ values ≤1.4 are considered appropriate (44). The standardized fit statistic (ZSTD) is an approximate *t* statistic, suggesting significant deviation from expected variation at the 0.05 alpha level if >2.0. Items with INFIT or OUTFIT MNSQ >1.4 and a corresponding ZSTD >2.0 were thus considered misfitting, suggesting deviation from unidimensionality.

*Rating scale functionality*. Functioning of the FACIT-F and FSS response scales was evaluated by means of the Rasch rating scale model, which allows examination of basic rating scale assumptions. Rating scales were thus assessed regarding the following aspects and criteria (45): rating scale category counts (there should be a minimum of 10 observations for each category in order to allow stable estimations); average rating scale category measures

(should be ordered in an expected manner, and each category should appear as the most probable outcome at some point on the underlying latent continuum); transition threshold points between categories (should be ordered in an expected manner); and category OUTFIT and INFIT MNSQ (should be <2.0 and preferably <1.5).

*Differential item functioning (DIF)*. DIF is present when an item displays different statistical properties in various subsets of respondents. The presence of DIF was explored between age groups (as defined by the median) and genders by comparing separate Rasch derived item calibrations from the respective sub-samples by means of two recommended criteria (46, 47). According to these, DIF is present if an item displays (a) a DIF contrast (i.e., the difference between the separate item calibrations) of more than 0.5 logits, or (b) a *t*-test determined statistically significant difference between the separate calibrations. Presence of items displaying DIF indicates that these items have different meaning across subsets of respondents and challenge the validity of pooling and comparing data across such subgroups.

Variables were checked regarding assumptions underlying the use of parametric and non-parametric statistics and described and analyzed accordingly. Analyses were performed using SPSS version 12 (SPSS Inc., Chicago, IL) and WINSTEPS version 3.55 (winsteps.com, Chicago, IL). The alpha-level of significance was set at 0.05. P-values are 2-tailed.

**RESULTS**

No patients expressed any difficulties understanding or completing the questionnaires. Descriptive and psychometric statistics of the FACIT-F and FSS are summarized in Table 3.

*Data quality, reliability, floor- and ceiling effects*. Both instruments yielded good data quality with few missing item responses (Table 3; no significant differences across study sites). Scale scores could be computed for all subjects according to the FACIT-F and for all but 5 patients (4.2%) for the FSS.

Coefficient alpha reliabilities were ≥0.9 (Table 3), indicating little random measurement error. One-hundred and seven patients (91%) returned the second FACIT-F after an average of 8 days, of whom 87 (81%) reported stable levels of fatigue since they completed the first questionnaire. Test-retest reliability based on scores from stable patients was 0.85 (Table 3). Scores from stable patients who responded within 2 weeks (n=81) and during the second week after the initial assessment (n=70) yielded test-retest ICC coefficients of 0.85 and 0.84, respectively.

Floor- and ceiling effects were minimal for both questionnaires (Table 3), thus supporting their appropriateness for this sample of people with PD.

*Construct validity and precision of scores*. The FACIT-F correlated strongly with both FSS and NHP-EN scores (Table 4). FSS scores correlated less strong with NHP-EN scores but the coefficient surpassed the predefined lower bound for support of convergent validity. The correlation (*r*) between Rasch calibrated person logit FACIT-F and FSS measures was 0.71, indicating that they measure related but not identical aspects. Both scales were able to correctly discriminate between patients classified as fatigued and non-fatigued (Table 3) with high statistical significance ($P<0.0001$), thus providing further support for their construct validity.

The effect size for the difference between fatigued and non-fatigued patients was 1.23 for the FACIT-F and 1.09 for the FSS. Unpaired *t*-tests yielded *t*-statistics (95% CI widths) of -8.447 (10.94) for the FACIT-F and -6.779 (17.1) for the FSS. The relative efficiency of the FACIT-F was 1.55 over the FSS, indicating advantages of the FACIT-F over the FSS in terms of measurement precision.

*Unidimensionality*. Parallel analyses EFAs of the FACIT-F and FSS resulted in one factor from each with empirical eigenvalues exceeding those from random data, which supports the unidimensionality of both scales. The eigenvalues of the first two empirical FACIT-F factors were 6.14 and 1.3, and from the corresponding random data they were 1.74 and 1.54, respectively. Eigenvalues of the first two FSS factors were 6.0 and 0.8, and from the corresponding random data they were 1.58 and 1.39, respectively. Rasch analyses identified one item in each questionnaire that failed to meet the predefined unidimensionality criteria. Item An8 of the FACIT-F ("I need to sleep during the day") had an INFIT MNSQ of 1.56 (ZSTD, 3.8) and an OUTFIT MNSQ of 2.03 (ZSTD, 5.6). In the FSS, item 1 ("My motivation is lower when I am fatigued") had an INFIT MNSQ of 1.84 (ZSTD, 5.0) and an OUTFIT MNSQ of 2.38 (ZSTD, 6.6).

*Rating scale functionality*. Rating scale analyses supported basic rating scale assumptions for both the FACIT-F and FSS. Category endorsement frequencies ranged between 79 (5%) and 435 (29%). Average rating scale category and transition step measures were all monotonically ordered in an expected manner in both questionnaires (Fig. 1). Category INFIT and OUTFIT MNSQ values ranged between 0.93-1.08 and 0.80-1.27, respectively, for the FACIT-F. For the FSS, category INFIT and OUTFIT MNSQ values were between 0.78-1.31 and 0.74-1.36, respectively.

*Differential item functioning (DIF)*. Explorative analyses of DIF by gender and age groups did not suggest any signs of DIF in the FACIT-F (Fig. 2A-B). Similarly, there were no signs of DIF by gender among FSS items (Fig. 2C). However, *t*-tests (but not DIF contrasts)

indicated DIF by age for item 1 (DIF contrast = 0.42 logits, $t$ = -2.48, P = 0.015) and item 8 (DIF contrast = 0.44 logits, $t$ = 2.6, P = 0.011) of the FSS (Fig. 2D). Following Bonferroni correction for multiple comparisons, these differences did not remain significant (P = 0.135 and 0.099, respectively). This provides preliminary support for the validity of comparing and pooling FACIT-F and FSS data across age and gender groups among people with PD.

## DISCUSSION

The 13-item FACIT-F and the 9-item FSS are among the most widely used brief fatigue questionnaires in clinical medicine and both have compared favorably to other fatigue instruments in previous evaluations (19, 27, 48, 49). However, this study appears to be the first head-to-head comparison between the two, and the first documentation of their measurement properties in PD. Both instruments were found to perform well and data provide support for their validity and reliability as self-report fatigue instruments in PD.

Several multidimensional instruments have been developed to address various expressions and/or consequences of fatigue (50, 51). Although such tools play a role in clinical assessment, they can be hampered by an increase in respondent burden, and it is not clear that conceptual multidimensionality of item content in the realm of fatigue cannot be captured in an assessment that has essentially unidimensional measurement properties. Brief, efficient measurement is particularly useful in clinical trials and other clinical and research applications where patient burden is a concern (16). This aspect becomes particularly salient when considering patient populations that are expected to experience troublesome levels of fatigue.

The study sample represented all five stages of PD according to Hoehn & Yahr, although the overall severity was somewhat skewed towards less severe stages, and enrollment criteria excluded patients with clinically significant co-morbidities, such as depression. This poses some limits to the generalizability of results and may underestimate the occurrence of fatigue in PD. However, the primary purpose of the study was not to provide a representative picture of the prevalence of fatigue in PD, but to assess the measurement properties of the FACIT-F and compare it to those of the FSS. Nevertheless, the magnitude and frequency of fatigue found in this sample were very similar to those previously reported in PD (4, 5, 7). Some differences in age and indices of PD severity were observed across study centers, probably reflecting somewhat different clinic profiles. However, and more importantly, there were no differences regarding fatigue questionnaire response rates, fatigue scores or rates of people classified as fatigued across study sites.

Data quality was high for both instruments, which, together with the lack of reported difficulties and ambiguities, indicates adequate patient perceived acceptability of the questionnaires (15). Both instruments also demonstrated excellent reliability with coefficient alpha values ≥0.9, indicating that scores can be considered sufficiently reliable to be used at an individual patient level (37, 40). Evaluations of test-retest reliability of the FACIT-F yielded results compliant with a sufficiently high level of reproducibility to support its feasibility for use in clinical research (37, 39). These properties are important because insufficient reliability does not only compromise interpretability of scores, but also adversely affects, e.g., their correlations with other measures and sample size requirements in clinical trials (39).

Floor and ceiling effects were small for both FACIT-F and FSS scores. This is an important aspect of effective outcome assessment because large floor and ceiling effects may compromise responsiveness, i.e., the ability of an instrument to detect change. Due to its cross-sectional design, responsiveness could not be addressed in the present study. The responsiveness of the FACIT-F has, however, been thoroughly evaluated among, e.g., people with anemia and RA, where it has shown an MID of 3-4 raw score points, corresponding to about 1 standard error of measurement (SEM) and a moderate effect size (19, 21, 22). Interestingly, these experiences are in close resemblance with observations reported here. The SEM (= SD x √1-alpha) associated with FACIT-F in this study was thus 3.13, which could indicate that a similar MID (i.e., 3-4 raw score points) may apply also to PD (52). While prospective evidence still is lacking for PD, availability of this type of consistent information facilitates the interpretation and clinical meaning of scores and aids in designing clinical trials. In contrast, evidence regarding the responsiveness of the FSS appears to be lacking (50, 51). Thus, the FSS has often (53-55), but not always (56), failed to detect change where other parallel fatigue measures have succeeded. While it is recognized that the observations reported here regarding known-groups differences in fatigue scores cannot be interpreted in terms of responsiveness, the narrower 95% CI of the difference score and larger ES of the FACIT-F, as compared to the FSS, are nevertheless in accordance with the experiences reviewed above. Furthermore, the relative efficiency of 1.55 in favor of the FACIT-F suggests that this instrument would be about 50% more efficient in detecting differences in clinical PD trial fatigue outcomes as compared to the FSS.

*A priori* expectations regarding convergent and known-groups validity of both the FACIT-F and the FSS were met, thus providing support for the construct validity of the questionnaires. Whereas the FACIT-F correlated relatively strongly with both the FSS and the

NHP-EN, the FSS correlated weaker (albeit still within the predefined acceptable range) with the NHP-EN. This is probably due to the somewhat different emphasis of the questionnaires, illustrated also by the correlation between the Rasch derived FACIT-F and FSS person logit measures (where about 50% the variance in one could be explained by that in the other). While the NHP-EN focuses on feelings of reduced energy, FSS items emphasize functional impact of fatigue (50). The FACIT-F, on the other hand, appears to cover both aspects.

This may also have contributed to the somewhat stronger EFA-derived second factor of the FACIT-F, as compared to the FSS. Item fit to the Rasch model, on the other hand, indicated one misfitting item in each questionnaire, of which the degree of misfit was greater for the FSS item. However, taken together, and until more data are available, we do not consider these observations to call for any questionnaire revisions, but both appear to define reasonably unidimensional underlying and partly overlapping concepts.

From a substantive point of view, the observed item misfits may, however, suggest that needing to sleep during the day (item An8, FACIT-F) and fatigue-related lack of motivation (item 1, FSS) do not adhere as much as the other items in the respective scales to the same underlying predominant concept. The misfitting FACIT-F item may suggest that fatigue (as defined by the remaining items) differs from daytime sleepiness, a notion that has been suggested elsewhere (25). While motivational processes have been suggested as an important aspect of fatigue in PD and other brain disorders (2), the observed misfit of the motivation item of the FSS could be due to the functional emphasis of this questionnaire. Thus, while the item may be relevant to fatigue, it may not behave in harmony with the content of the other FSS items. Furthermore, although perceived lack of motivation may not be caused by fatigue (as implied by FSS item 1), central motivational processes may still be an important contributor to the experience of fatigue (2).

Rating scale functionality is a prerequisite for valid interpretation of resulting scores and refers to its shared meaning among respondents (45). If the rating scale does not function the way it is assumed to among the people using it (due to, e.g., ambiguous distinction between categories), responses may become arbitrary and the resulting scores dubious. In this study, rating scale functioning was acceptable for both the FACIT-F and the FSS. However, while FACIT-F categories displayed monotonically ordered thresholds of about equal distances, FSS category transition thresholds were somewhat unequally spaced. One reason for this may be the lack of rating scale category definitions in the FSS. For example, a score of "3" may have different meanings to different respondents.

The lack of differential item functioning (DIF) among FACIT-F and FSS items support their measurement validity across genders and age groups in PD. Whereas we are unaware of any previous DIF evaluations of the FSS, Lai et al. (57) reported presence of modest DIF between cancer patients and general US population representatives for three FACIT-F items according to the 0.5 logit DIF contrast definition. In addition, these items also displayed various degrees of misfit (57). However, Lai et al. (57) did not assess DIF by gender or age. It is therefore possible that there is a lack of age and gender DIF also in cancer and general population samples, although the associated misfits argue against this. Furthermore, the sample size of this study was considerably smaller than that by Lai et al. (57), who analyzed responses from about 1000 people in each sample, which increases measurement precision. Nevertheless, while the DIF results reported here should be viewed as tentative, they support the lack of DIF and hence the validity of pooling and comparing data across genders and age groups in PD for both the FACIT-F and probably also the FSS. However, they also illustrate the need for further DIF evaluations in larger samples, not only across gender and age groups, but also across diagnostic and language-/culture groups.

Recently, Brown and coworkers (58) proposed a PD-specific fatigue questionnaire, the 16-item Parkinson Fatigue Scale. While such condition- *and* symptom-specific tools can be of value, they may also carry limitations. For example, in contrast to using cross-validated generic instruments, results do not allow for direct comparisons across patient groups or with healthy control populations. This is potentially problematic because of the largely unknown cause(s) of fatigue and the need to identify effective therapy, which both well may be non-disease specific (2, 3). However, the relative merit of the Parkinson Fatigue Scale and more generic tools such as the FACIT-F and FSS will need to be assessed empirically.

The current study illustrates the need for additional work to fully assess the relative merits and limitations of the FACIT-F and FSS. First, regretfully the FSS was not analyzed regarding test-retest reliability, which therefore needs to be assessed in future studies. Furthermore, there is a clear need for longitudinal data in order to empirically assess responsiveness and elucidate whether circumstantial indications favoring the FACIT-F in this regard also are supported empirically. It is also possible that the FACIT-F performed better than the FSS due to differences in translation into Swedish. Testing in other languages would be enlightening. The lack of information regarding the responsiveness and MID of the FSS will also need to be addressed in order to establish its relative merits in relation to other fatigue questionnaires, such as the FACIT-F. Finally, more and larger samples are needed to

support the present findings. Preferably, such studies should cover more than one country and diagnostic group in order to allow for assessment of DIF by culture and diagnosis.

In conclusion, this study provides support for the reliability and validity of two widely used brief generic fatigue questionnaires, the FACIT-F and FSS, as applied among people with PD. When comparing the two, the FACIT-F exhibits better measurement precision and modest but consistent psychometric advantages over the FSS. These observations offer a starting point for evidence-based measurement of fatigue in PD. Availability of psychometrically sound fatigue measures that are applicable and valid across disorders provides a sound basis for advancing our understanding of this common, distressing and under-recognized symptom.

**Acknowledgements**

The authors want to thank all participating patients for their cooperation and Drs. J. Ahlberg, N. Dizdar, H. Edwall, B. Johnels, J. Lökk, S. Pålhagen, and T. Willows for assistance with patient recruitment.

**REFERENCES**


1. Krupp LB, Pollina DA. Mechanisms and management of fatigue in progressive neurological disorders. Curr Opin Neurol 1996;9:456-460.
2. Chaudhuri A, Behan PO. Fatigue and basal ganglia. J Neurol Sci 2000;179:34-42.
3. Swain MG. Fatigue in chronic disease. Clin Sci (Lond) 2000;99:1-8.
4. Herlofson K, Larsen JP. Measuring fatigue in patients with Parkinson's disease - the Fatigue Severity Scale. Eur J Neurol 2002;9:595-600.
5. Karlsen K, Larsen JP, Tandberg E, Jorgensen K. Fatigue in patients with Parkinson's disease. Mov Disord 1999;14:237-241.
6. Scott B, Borgman A, Engler H, Johnels B, Aquilonius SM. Gender differences in Parkinson's disease symptom profile. Acta Neurol Scand 2000;102:37-43.
7. Shulman LM, Taback RL, Rabinstein AA, Weiner WJ. Non-recognition of depression and other non-motor symptoms in Parkinson's disease. Parkinsonism Relat Disord 2002;8:193-197.
8. van Hilten JJ, Weggeman M, van der Velde EA, Kerkhof GA, van Dijk JG, Roos RA. Sleep, excessive daytime sleepiness and fatigue in Parkinson's disease. J Neural Transm Park Dis Dement Sect 1993;5:235-244.
9. Alves G, Wentzel-Larsen T, Larsen JP. Is fatigue an independent and persistent symptom in patients with Parkinson disease? Neurology 2004;63:1908-1911.
10. Abe K, Takanashi M, Yanagihara T. Fatigue in patients with Parkinson's disease. Behav Neurol 2000;12:103-106.
11. Lou JS, Kearns G, Oken B, Sexton G, Nutt J. Exacerbated physical fatigue and mental fatigue in Parkinson's disease. Mov Disord 2001;16:190-196.
12. Weintraub D, Newberg AB, Cary MS, et al. Striatal dopamine transporter imaging correlates with anxiety and depression symptoms in Parkinson's disease. J Nucl Med 2005;46:227-232.
13. Hunt SM, McKenna SP, McEwen J, Backett EM, Williams J, Papp E. A quantitative approach to perceived health status: a validation study. J Epidemiol Community Health 1980;34:281-286.
14. Hagell P, Whalley D, McKenna SP, Lindvall O. Health status measurement in Parkinson's disease: validity of the PDQ-39 and Nottingham Health Profile. Mov Disord 2003;18:773-783.

15. McHorney CA, Ware JE, Jr., Lu JF, Sherbourne CD. The MOS 36-item Short-Form Health Survey (SF-36): III. Tests of data quality, scaling assumptions, and reliability across diverse patient groups. Med Care 1994;32:40-66.
16. Scientific Advisory Committee of the Medical Outcomes Trust. Assessing health status and quality-of-life instruments: attributes and review criteria. Qual Life Res 2002;11:193-205.
17. Yellen SB, Cella DF, Webster K, Blendowski C, Kaplan E. Measuring fatigue and other anemia-related symptoms with the Functional Assessment of Cancer Therapy (FACT) measurement system. J Pain Symptom Manage 1997;13:63-74.
18. Cella D, Lai JS, Chang CH, Peterman A, Slavin M. Fatigue in cancer patients compared with fatigue in the general United States population. Cancer 2002;94:528-538.
19. Cella D, Yount S, Sorensen M, Chartash E, Sengupta N, Grober J. Validation of the Functional Assessment of Chronic Illness Therapy Fatigue Scale relative to other instrumentation in patients with rheumatoid arthritis. J Rheumatol 2005;32:811-819.
20. Cella D, Zagari MJ, Vandoros C, Gagnon DD, Hurtz HJ, Nortier JW. Epoetin alfa treatment results in clinically significant improvements in quality of life in anemic cancer patients when referenced to the general population. J Clin Oncol 2003;21:366-373.
21. Cella D, Eton DT, Lai JS, Peterman AH, Merkel DE. Combining anchor and distribution-based methods to derive minimal clinically important differences on the Functional Assessment of Cancer Therapy (FACT) anemia and fatigue scales. J Pain Symptom Manage 2002;24:547-561.
22. Patrick DL, Gagnon DD, Zagari MJ, Mathijs R, Sweetenham J. Assessing the clinical significance of health-related quality of life (HrQOL) improvements in anaemic cancer patients receiving epoetin alfa. Eur J Cancer 2003;39:335-345.
23. Mallinson T, Cella D, Cashy J, Holzner B. Giving meaning to measure: Linking self-reported fatigue and function to performance of everyday activities. J Pain Symptom Manage 2006; in press.
24. Hahn EA, Cella D. Health outcomes assessment in vulnerable populations: measurement challenges and recommendations. Arch Phys Med Rehabil 2003;84:S35-42.
25. Friedman JH, Chou KL. Sleep and fatigue in Parkinson's disease. Parkinsonism Relat Disord 2004;10 Suppl 1:S27-35.
26. Krupp LB, LaRocca NG, Muir-Nash J, Steinberg AD. The fatigue severity scale. Application to patients with multiple sclerosis and systemic lupus erythematosus. Arch Neurol 1989;46:1121-1123.

27. Horemans HL, Nollet F, Beelen A, Lankhorst GJ. A comparison of 4 questionnaires to measure fatigue in postpoliomyelitis syndrome. Arch Phys Med Rehabil 2004;85:392-398.
28. Kleinman L, Zodet MW, Hakim Z, et al. Psychometric evaluation of the fatigue severity scale for use in chronic hepatitis C. Qual Life Res 2000;9:499-508.
29. Goetz CG, Stebbins GT, Chmura TA, Fahn S, Klawans HL, Marsden CD. Teaching tape for the motor section of the unified Parkinson's disease rating scale. Mov Disord 1995;10:263-266.
30. Klawans HL, Goetz CG, Tanner CM. Common movement disorders: a video presentation. Philadelphia: Lippincott Williams & Wilkins, 1988.
31. Fahn S, Elton RL, Committee motUD. Unified Parkinson's Disease Rating Scale. In: Fahn S, Marsden CD, Calne DB, Goldstein M, eds. Recent Developments in Parkinson's Disease, Vol. 2. Florham Park: MacMillan Healthcare Information, 1987: 153-163.
32. Hoehn MM, Yahr MD. Parkinsonism: onset, progression and mortality. Neurology 1967;17:427-442.
33. Folstein MF, Folstein SE, McHugh PR. "Mini-mental state". A practical method for grading the cognitive state of patients for the clinician. J Psychiatr Res 1975;12:189-198.
34. Bonomi AE, Cella DF, Hahn EA, et al. Multilingual translation of the Functional Assessment of Cancer Therapy (FACT) quality of life measurement system. Qual Life Res 1996;5:309-320.
35. Hunt SM, Alonso J, Bucquet D, Niero M, Wiklund I, McKenna S. Cross-cultural adaptation of health measures. European Group for Health Management and Quality of Life Assessment. Health Policy 1991;19:33-44.
36. Wiklund I, Romanus B, Hunt SM. Self-assessed disability in patients with arthrosis of the hip joint. Reliability of the Swedish version of the Nottingham Health Profile. International Disability Studies 1988;10:159-163.
37. Nunnally JC, Bernstein IH. Psychometric theory. New York: McGraw-Hill, Inc., 1994.
38. Wright BD, Masters GN. Rating scale analysis. Chicago: MESA Press, 1982.
39. Fleiss JL. The design and analysis of clinical experiments. New York: Wiley, 1986.
40. McHorney CA, Tarlov AR. Individual-patient monitoring in clinical practice: are available health status surveys adequate? Qual Life Res 1995;4:293-307.

41. Cohen J. Statistical power analysis for the behavioral sciences, 2nd ed. . Hillsdale: Lawrence Earlbaum and Associate, 1988.
42. Fayers PM, Machin D. Quality of life: Assessment, analysis and interpretation. West Sussex: John Wiley & Sons, Ltd., 2000.
43. O'Connor BP. SPSS and SAS programs for determining the number of components using parallel analysis and velicer's MAP test. Behav Res Methods Instrum Comput 2000;32:396-402.
44. Wright BD, Linacre JM. Reasonable mean-square fit values. Rasch Measurement Transactions 1994;8:370.
45. Linacre JM. Optimizing rating scale category effectiveness. J Appl Meas 2002;3:85-106.
46. Draba RE. The identification and interpretation of item bias. Chicago: MESA Press, 1977. Report No.: 26.
47. Scheuneman JD, Subhiyah RG. Evidence for the validity of a Rasch model technique for identifying differential item functioning. J Outcome Meas 1998;2:33-42.
48. Chipchase SY, Lincoln NB, Radford KA. Measuring fatigue in people with multiple sclerosis. Disabil Rehabil 2003;25:778-784.
49. Hwang SS, Chang VT, Kasimis BS. A comparison of three fatigue measures in veterans with cancer. Cancer Invest 2003;21:363-373.
50. Comi G, Leocani L, Colombo B, Rossi P. Pathophysiology and treatment of fatigue in multiple sclerosis. In: Abramsky O, Compston DAS, Miller A, Said G, eds. Brain disease – Therapeutic strategies and repair. London: Martin Dunitz Ltd., 2002: 389-394.
51. Schwid SR, Covington M, Segal BM, Goodman AD. Fatigue in multiple sclerosis: current understanding and future directions. J Rehabil Res Dev 2002;39:211-224.
52. Wyrwich KW, Bullinger M, Aaronson N, Hays RD, Patrick DL, Symonds T. Estimating clinically significant differences in quality of life outcomes. Qual Life Res 2005;14:285-295.
53. Kos D, Kerckhofs E, Nagels G, et al. Assessing fatigue in multiple sclerosis: Dutch modified fatigue impact scale. Acta Neurol Belg 2003;103:185-191.
54. Krupp LB, Coyle PK, Doscher C, et al. Fatigue therapy in multiple sclerosis: results of a double-blind, randomized, parallel trial of amantadine, pemoline, and placebo. Neurology 1995;45:1956-1961.
55. Wingerchuk DM, Benarroch EE, O'Brien PC, et al. A randomized controlled crossover trial of aspirin for fatigue in multiple sclerosis. Neurology 2005;64:1267-1269.

56. Rammohan KW, Rosenberg JH, Lynn DJ, Blumenfeld AM, Pollak CP, Nagaraja HN. Efficacy and safety of modafinil (Provigil) for the treatment of fatigue in multiple sclerosis: a two centre phase 2 study. J Neurol Neurosurg Psychiatry 2002;72:179-183.
57. Lai JS, Cella D, Chang CH, Bode RK, Heinemann AW. Item banking to improve, shorten and computerize self-reported fatigue: an illustration of steps to create a core item bank from the FACIT-Fatigue Scale. Qual Life Res 2003;12:485-501.
58. Brown RG, Dittner A, Findley L, Wessely SC. The Parkinson fatigue scale. Parkinsonism Relat Disord 2005;11:49-55.
59. Reimer J, Grabowski M, Lindvall O, Hagell P. Use and interpretation of on/off diaries in Parkinson's disease. J Neurol Neurosurg Psychiatry 2004;75:396-400.

**TABLE 1** Patient characteristics (n=118) [a]

| | |
|---|---|
| Gender (men / women) [b] | 64 (54%) / 54 (46%) |
| Age (years) [c] | 63.9 (9.6) |
| Time since PD diagnosis (years) [c] | 8.4 (5.7) |
| Daily dopaminergic anti-PD medication [d, e, f] | 780 (518-1110) |
| Hoehn & Yahr stage of PD (during "on") [d, g] | II (II-III) |
| Hoehn & Yahr stage of PD (during "off") [d, g] | III (II-III) |
| Schwab & England ADL score (during "on") [d, h] | 90 (90-100) |
| Schwab & England ADL score (during "off") [d, h] | 90 (80-90) |
| UPDRS motor score (during "on") [d, i] | 17 (10.5-27) |
| MMSE score [d, j] | 29 (28-30) |

[a] No statistical differences across study sites except for age (younger participants from one center), Hoehn & Yahr and Schwab and England scores during "off" (more severely affected participants from one center).

[b] n (%).

[c] Mean (standard deviation).

[d] Median (interquartile range).

[e] Expressed as total levodopa equivalent dose: 100 levodopa equivalents = 100 mg standard levodopa = 133 mg controlled-release levodopa = 10 mg bromocriptine = 5 mg ropinirole = 1 mg pramipexole = 1 mg cabergoline = 2 mg apomorphine. For patients who received a COMT-inhibitor, the sum of standard levodopa and 0.75 times the dose of controlled-release levodopa was multiplied by 1.3 (59).

[f] Non-dopaminergic treatment consisted of selegiline (n=16), amantadine (n=11), anticholinergics (n=1) and neurosurgical interventions (n=8); one patient was not yet on any medical anti-parkinsonian therapy.

[g] Range, I-V (I = mild unilateral disease; II = Bilateral disease without postural impairment; III = Bilateral disease with postural impairment, moderate disability; IV = Severe disability, still able to walk and stand unassisted; V = Confined to bed or wheelchair unless aided) (32).

[h] Range, 0-100 (100 = normal ADL functioning).

[i] Range, 0-108 (0 = no signs of parkinsonism).

[j] Range, 0-30 (30 = normal cognition).

PD, Parkinson's disease; ADL, activities of daily living; UPDRS, Unified Parkinson's Disease Rating Scale; MMSE, Mini-Mental State Exam.

**TABLE 2.** Fatigue self-report questionnaires used in the current study

| NHP-EN items | | FACIT-F items | | FSS items | |
|---|---|---|---|---|---|
| No. | Content (abridged) | No. | Content (abridged) | No. | Content (abridged) |
| 1 | Tired all the time | HI7 | Feel fatigued | 1 | Motivation lower when fatigued |
| 12 | Everything is an effort | HI12 | Weak all over | 2 | Exercise brings on fatigue |
| 26 | Soon out of energy | An1 | Listless (“washed out”) | 3 | Easily fatigued |
| | | An2 | Feel tired | 4 | Fatigue interferes with physical functioning |
| | | An3 | Trouble starting things because tired | 5 | Fatigue causes frequent problems |
| | | An4 | Trouble finishing things because tired | 6 | Fatigue prevents sustained physical functioning |
| | | An5 | Have energy | 7 | Fatigue interferes with duties and responsibilities |
| | | An7 | Able to do usual activities | 8 | Fatigue is among my three most disabling symptoms |
| | | An8 | Need to sleep during the day | 9 | Fatigue interferes with work, family or social life |
| | | An12 | Too tired to eat | | |
| | | An14 | Need help doing usual activities | | |
| | | An15 | Frustrated by being too tired to do things I want to do | | |
| | | An16 | Have to limit social activity because tired | | |

**TABLE 3** Descriptive and psychometric statistics for the FACIT-F and FSS

| | FACIT-F [a] | FSS [b] |
|---|---|---|
| Score mean (SD) | 34.2 (9.9) [c] | 3.9 (1.6) [c] |
| Score median (IQR) | 35.5 (26.8-42) [c] | 3.8 (2.6-5.2) [c] |
| *Data quality* | | |
| Missing item responses (%) | 0.9 | 0.8 |
| Computable scale scores (%) | 100 | 95.8 |
| *Reliability* [d] | | |
| Cronbach's alpha (min, max if item deleted) | 0.90 (0.89, 0.91) | 0.94 (0.93, 0.94) |
| | 0.92 (0.91, 0.93) [e] | - |
| Test-retest ICC [f] | 0.85 | - |
| Floor/ceiling effects (%)[g] | 1.7 / 0 | 2.5 / 2.5 |
| *Known-groups validity* [h] | | |
| Non-fatigued patients (n=61; 52%) [c] | 41 (35.6-45) / 40.1 (7.2) | 3.1 (2-4.1) / 3.1 (1.2) |
| Fatigued patients (n=57; 48%) [c] | 27 (22-34.1) / 28 (8.4) [i] | 5.1 (3.8-6.1) / 4.8 (1.5) [i] |

[a] Score range, 0-52 (0 = more fatigue).
[b] Score range, 1-7 (7 = more fatigue).
[c] No significant differences in FACIT-F or FSS scores, or in proportions of patients classified as fatigued and non-fatigued, across the four study sites.
[d] Should be <0.7 and preferably >0.8 (37, 39).
[e] Second administration.
[f] One-way random intra-class correlation coefficient for single measures.
[g] Should be <15% (40).
[h] Median (IQR) / mean (SD) FACIT-F and FSS scores compared between fatigued and non-fatigued patients (as defined by the NHP-EN, see Methods). Levene's test revealed lack of homoscedasticity (i.e., non-equal variances) between FSS scores but not between FACIT-F scores.
[i] P<0.0001 (Mann-Whitney *U*- and unpaired *t*-tests), as compared to patients classified as non-fatigued.

FACIT-F, Functional Assessment of Chronic Illness Therapy - Fatigue Scale; FSS, Fatigue Severity Scale; SD, standard deviation; IQR, interquartile range; ICC, intra-class correlation.

**TABLE 4** Convergent validity among FACIT-F, FSS and NHP-EN scores [a]

| | FACIT-F | FSS |
|---|---|---|
| FACIT-F | 1 | |
| FSS | -0.767 [b] | 1 |
| NHP-EN | -0.703 [b] | 0.624 |

[a] Spearman correlations. Coefficients ≥0.6-0.7 were expected in order for convergent validity to be supported.
[b] Negative coefficients due to opposite scoring directions.

FACIT-F, Functional Assessment of Chronic Illness Therapy - Fatigue Scale; FSS, Fatigue Severity Scale; NHP-EN, the Energy (EN) subscale of the Nottingham Health Profile (NHP).

**Legends to Figures**

Fig. 1.

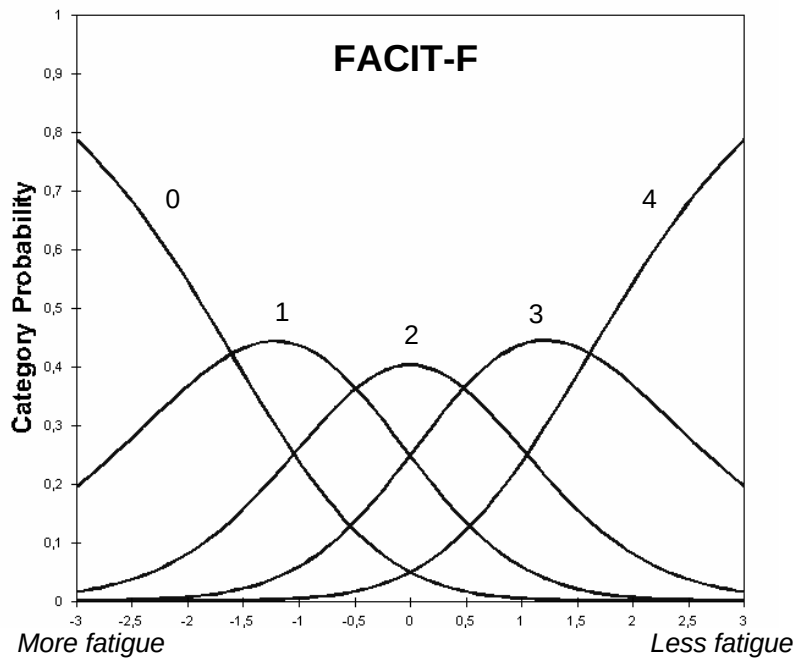
Rating scale functioning of the (A) FACIT-F and (B) FSS. Curves show the probability of each category (y-axis) relative to the logit difference between person and item measures (x-axis). Response categories should be ordered in an expected manner and emerge as more and more probable as one moves along the fatigue continuum (x-axis). Rating scale categories should thus appear as an ordered even succession of "hills" across the latent fatigue continuum, where each category is modal over a certain range. Transition steps between categories should also be ordered in an expected manner. Categories never emerging as modal and category step disordering contradict rating scale assumptions. FACIT-F, Functional Assessment of Chronic Illness Therapy - Fatigue Scale; FSS, Fatigue Severity Scale.

Fig. 2.

Differential item functioning (DIF) of the (A-B) FACIT-F and (C-D) FSS according to (A, C) gender and (B, D) age (defined by the median, i.e., 64 years). The FSS was reversed to yield measures in the same direction as the FACIT-F (high values = less fatigue). Separate Rasch item logit calibrations were performed for men and women and for younger and older respondents, and plotted against one another with 0.5 logit trace lines. Circled item plots indicate item calibrations with statistically significant ($P<0.05$; *t*-test) deviations (not corrected for multiple comparisons) between subsets of respondents (FSS items 1 and 8; see main text for details). DIF, differential item functioning; FACIT-F, Functional Assessment of Chronic Illness Therapy - Fatigue Scale; FSS, Fatigue Severity Scale.

Fig. 1

**A**


FACIT-F
Category Probability
0
1
2
3
4
More fatigue
Less fatigue

**B**

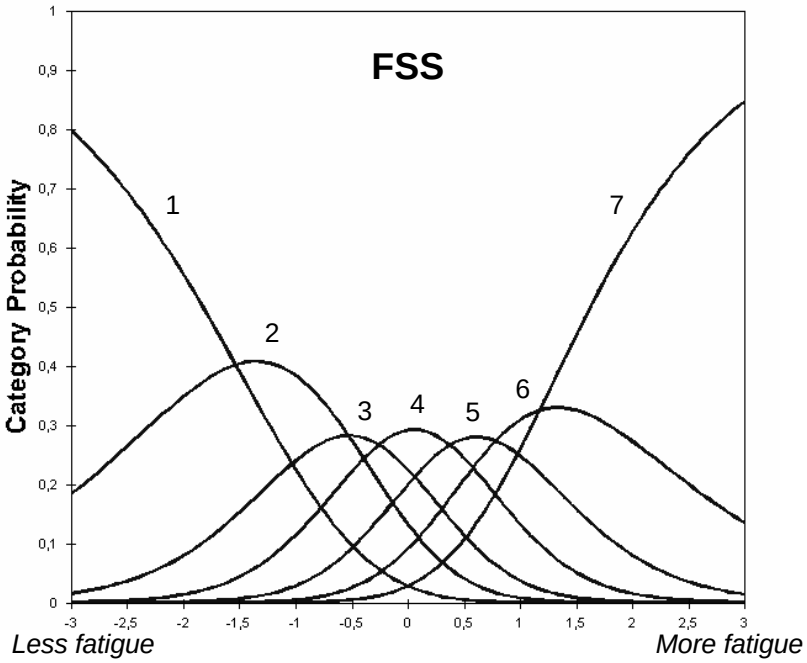
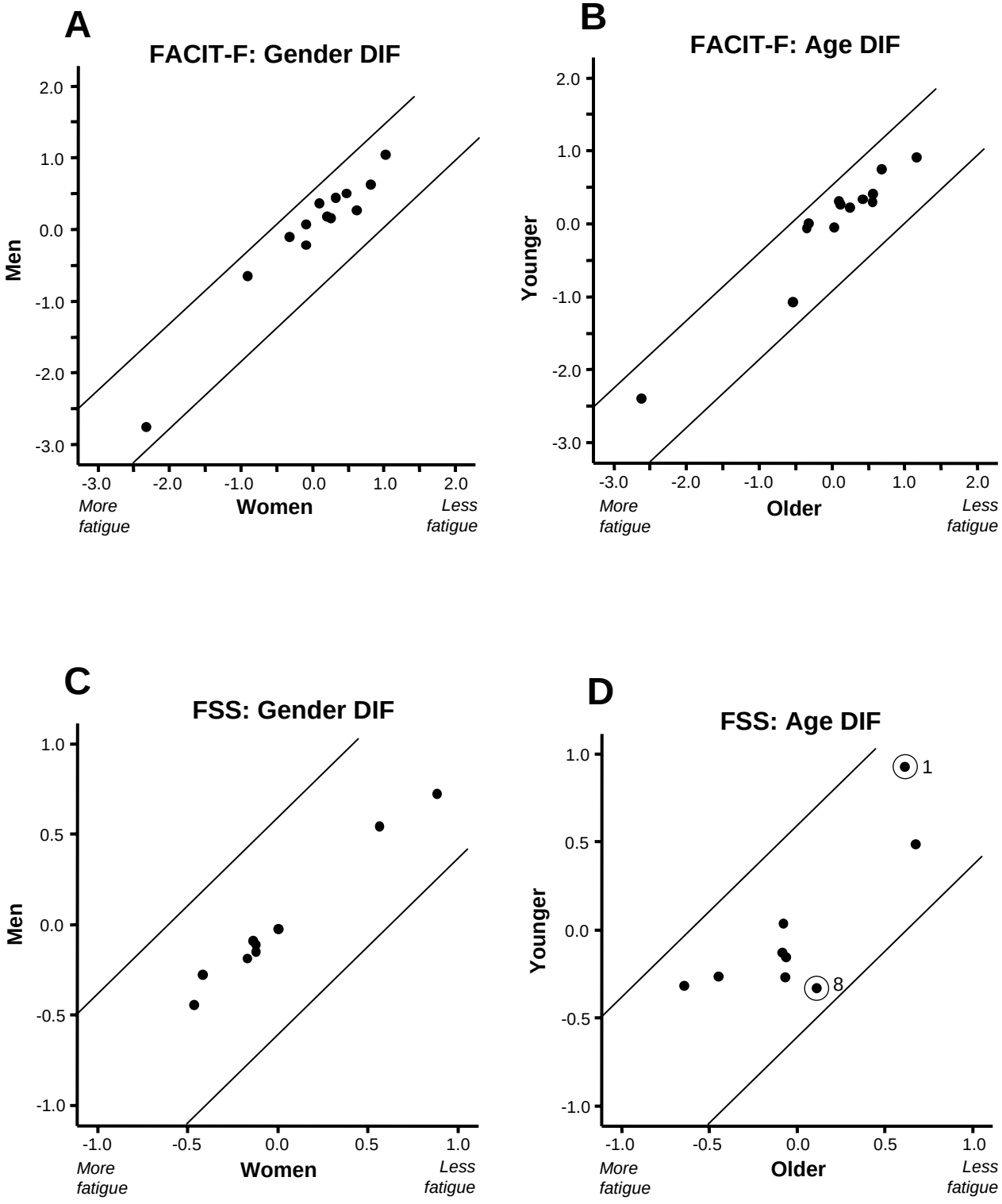

FSS
Category Probability
1
2
3
4
5
6
7
Less fatigue
More fatigue

Fig. 2


A
FACIT-F: Gender DIF
Men
Women
More fatigue
Less fatigue
B
FACIT-F: Age DIF
Younger
Older
More fatigue
Less fatigue
C
FSS: Gender DIF
Men
Women
More fatigue
Less fatigue
D
FSS: Age DIF
Younger
Older
More fatigue
Less fatigue
1
8