Fatigue in Parkinson's Disease: Measurement Properties of a Generic and a Condition-specific Rating Scale.

Nilsson, Maria H; Jonasson, Stina; Hagell, Peter

Published in:
Journal of Pain and Symptom Management

DOI:
10.1016/j.jpainsymman.2012.11.004

2013

Fatigue in Parkinson's disease: Measurement properties of a generic and a condition-specific rating scale

Maria H Nilsson, PhD [1], Stina Bladh, MSc [1,2], Peter Hagell, PhD [3]

[1] Department of Health Sciences, Lund University, Lund, Sweden
[2] Department of Neurology, Skåne University Hospital, Lund, Sweden
[3] The PRO-CARE Group, School of Health and Society, Kristianstad University, Kristianstad, Sweden

Keywords: Fatigue, Parkinson's disease, Rasch model, rating scales, reliability, validity

Running title: A generic and a PD-specific fatigue scale

Corresponding author:
Peter Hagell
School of Health and Society
Kristianstad University
SE-291 88 Kristianstad
Sweden
Tel: +46 44 204056
Fax: +46 44 204043
E-mail: Peter.Hagell@hkr.se

ABSTRACT

Context: High quality fatigue rating scales are needed to advance understanding of fatigue and determine the efficacy of interventions. Several fatigue scales are used in Parkinson's disease, but few have been tested using modern psychometric methodology (Rasch analysis).

Objectives: To examine the measuring properties of the generic Functional Assessment of Chronic Illness Therapy - Fatigue (FACIT-F) scale and the condition-specific 16-item Parkinson Fatigue Scale (PFS-16) using Rasch analysis.

Methods: Postal survey data (n=150; 47% women; mean age, 70) were Rasch analyzed. PFS-16 scores were tested according both to the original polytomous and the suggested alternative dichotomized scoring methods.

Results: The PFS-16 showed overall Rasch model fit whereas the FACIT-F showed signs of misfit, which probably was due to a sleepiness-related item and mixing of positively/negatively worded items. There was no differential item functioning by disease duration but by fatigue status (greater likelihood of needing to sleep or rest during the day among people classified as non-fatigued) in the PFS-16 and FACIT-F. However, this did not impact total score based estimated person measures. Targeting and reliability (≥0.86) was good, but the dichotomized PFS-16 showed compromised measurement precision. Polytomous and dichotomized PFS-16 and FACIT-F scores identified 6, 3 and 4 statistically distinct sample strata, respectively.

Conclusion: We found general support for the measurement properties of both scales. However, polytomous PFS-16 scores exhibited advantages compared to dichotomous PFS-16 and FACIT-F scores. Dichotomization of item responses compromises measurement precision and ability to separate people, and should be avoided.

INTRODUCTION

Fatigue is common and troublesome in Parkinson's disease (PD) (1). Its etiology remains speculative and specific therapy is lacking (1-3). In order to advance understanding of fatigue and determine the efficacy of interventions there is a need for high quality fatigue rating scales. A recent systematic review recommended four scales for rating fatigue in PD: the Multidimensional Fatigue Inventory (MFI), the Fatigue Severity Scale (FSS), the Functional Assessment of Chronic Illness Therapy - Fatigue Scale (FACIT-F), and the PD-specific 16-item Parkinson Fatigue Scale (PFS-16) (4). These scales were all developed according to classical test theory (CTT) principles (5, 6), whereas modern test theory (particularly the Rasch model) is preferable to CTT in rating scale development and evaluation (5, 7, 8).

The relative merits of scales should preferably be determined in empirical head-to-head comparisons. To that end, the FACIT-F has exhibited better measurement precision than the FSS in PD (9). Similarly, a comparison between the FSS and the PFS-16 found both to be adequate, but reliability was somewhat better for the PFS-16 (10). The generic FSS and FACIT-F appear to be the only fatigue scales among those identified as recommended (4) that have been Rasch analyzed in PD (9). It is therefore unknown to what extent the PFS-16 meets the more rigorous demands of the Rasch model, and evidence is limited regarding its potential advantages over a generic fatigue scale (4).

Here we report a Rasch based head-to-head comparison of the measurement properties of the FACIT-F and the PFS-16 in PD.

METHODS

Participants and procedure

An anonymous postal survey was sent to all members registered as having PD in a regional branch of the Swedish PD Association. The study was conducted in accordance with the Declaration of Helsinki and all participants provided written consent.

Instruments

The PFS-16 consists of 16 items (Appendix 1) with five polytomous response categories ('strongly disagree', 'disagree', 'do not agree or disagree', agree', and 'strongly agree') (11). Responses were scored from 0 (strongly disagree) to 4 (strongly agree), yielding a summed total score ranging from 0-64 (64=more fatigue). This is equivalent to the original 1-5 scoring method (4, 12). An alternative scoring method has also been proposed (11), where item responses are dichotomized ('agree' and 'strongly agree'=1; all other responses=0), giving a total score of 0-16 (16=more fatigue). Both scoring methods require complete responses to produce total scores. Here we refer to the polytomous (0-4) scoring as PFS-16p and to the dichotomized scoring as PFS-16d.

The FACIT-F consists of 13 items (Appendix 1) with five response categories, scored 0-4 ('not at all', 'a little bit', 'somewhat', 'quite a bit', and 'very much') (13). The total score ranges from 0-52 (52=less fatigue) and requires completion of more than 50% (≥7) of the items (www.facit.org).

In addition, the Energy section of the Nottingham Health Profile (NHP-EN; total scores, 0-100; 100=worse) (14) was used to identify the presence of fatigue; people who affirmed one

or more of its three dichotomous ("yes"/"no") items were classified as fatigued (9). Respondents rated their perceived PD severity as mild, moderate or severe.

Analyses

All analyses were conducted separately for the FACIT-F and the polytomous and dichotomized PFS-16 scoring versions. To ease interpretation relative to the PFS-16, total FACIT-F scores were reversed (52=more fatigue).

Data completeness and Rasch model fit

Data completeness was studied by calculating the percentage of missing item responses; up to 10% missing data has been suggested as acceptable (15).

Scales were analyzed regarding fit to the (partial credit) Rasch model (5, 16-18). The Rasch model separately locates persons and items on a common logit metric that is centered at the mean item location, which is set at zero. The probability of a certain item response is a function of the difference between the levels of the measured variable (e.g. fatigue) represented by the item and the person, respectively. Whether rating scales yield valid measurement depends on the extent to which data fit the Rasch model (5, 16, 18).

Since no single aspect of model fit is neither necessary nor sufficient on its own (5, 17), a variety of fit indices was considered interactively. Overall and individual item fit is supported by non-significant Bonferroni corrected chi-square statistics. Individual item chi-square values are also used as an order statistic; value(s) substantially larger than those of other items in a scale suggest misfit. Standardized residuals (discrepancies between observed and expected responses) also provide clues regarding the presence and nature of misfit. Large

(>2.5) positive residuals suggest multidimensionality (the item represents a different variable than the scale as a whole), whereas large negative residuals signal response dependency (suggesting item redundancy). Fit statistics were complemented by visual examination of item characteristic curves (ICCs) of expected and observed responses (5, 17).

Differential item functioning (DIF) is an additional aspect of fit to the Rasch model that concerns measurement invariance across relevant subgroups of people (5, 16, 17). That is, DIF analyses assess whether subgroups of people with similar levels on the measured construct respond systematically different to items. The presence of DIF suggests that an item does not work the same way in such subgroups. When DIF is uniform (i.e. item responses differ uniformly between subgroups across levels of the measured construct) this can be adjusted for by splitting the item into two new items, one for each subgroup (5, 17). In this study, we tested for DIF by disease duration (longer vs. shorter, as defined by the median duration, <7 vs. ≥7 years) and fatigue status according to the NHP-EN, i.e. whether people were classified as fatigued (n=86) or not (n=58). The clinical significance of any observed DIF was studied by assessing how DIF influenced the estimated person locations (logit measures). Items without DIF in the original scale were first anchored by their item locations from the DIF-adjusted scale to assure that the two sets of person estimates were on the same metric (19). The two sets of person locations were then compared and correlated to assess the influence of DIF on people's estimated fatigue levels.

Targeting

Targeting was analyzed by examining how well the distribution of PFS-16 and FACIT-F scores accorded with the levels of fatigue in the sample (5). Floor and ceiling effects were determined as the proportions of participants scoring minimum (floor) and maximum

(ceiling), and should not exceed 15% (20). Targeting was also studied using Rasch analyses (5). For a well-targeted scale, the mean sample location approximates the mean item location (i.e., zero). Furthermore, the person-to-item threshold distributions were examined to determine whether item response thresholds were evenly spread along a similar range as the persons, and if there were notable gaps in the distributions of item response thresholds (indicating compromised measurement precision).

Reliability and measurement precision

Reliability was estimated by the Person Separation Index (PSI), which is conceptually analogous to coefficient alpha (5, 17, 21). The separation ratio (√PSI/[1-PSI]) was also computed. This statistic is freed from the ceiling effect of traditional reliability indices and can be used to estimate the number of distinct sample strata (groups of people separated by at least three errors of measurement) that a scale distinguishes (18). In addition, measurement precision was assessed by examining the standard errors (SE) associated with various Rasch estimated person locations based on total scale scores.

Response category functioning

Rasch analysis allows for examination of whether response categories work as intended, i.e. if they reflect an increasing amount of the measured variable (5, 17). We thus examined the thresholds between adjacent response categories (i.e. the points where there are 50/50 probabilities of scoring, e.g., 1 or 2 and 2 or 3) in the PFS-16p and the FACIT-F.

Analyses were conducted using PASW Statistics 18 and RUMM2030 (22). The significance level was set at 0.05 (two-tailed).

RESULTS

Of 237 mailed questionnaires, 191 were returned. Of these, 41 were returned blank. Characteristics of the remaining 150 participants (conservative response rate, 63%) are presented in Table 1.

Data completeness and Rasch model fit

Data completeness was satisfying for both the PFS-16 (0-2% missing responses/item) and the FACIT-F (1-8% missing responses/item). The proportions of people without any missing item responses were 90% for the PFS-16 and 85% for the FACIT-F.

Rasch analyses suggested overall model fit for PFS-16p and PFS-16d scores, but not for FACIT-F scores (Table 2). At the individual item level, fit residuals of items 1 and 14 of the PFS-16 were larger than expected. These departures were not significant and their associated chi-square values (Table 2) were not considerably different from those of other scale items (range PFS-16p, 0.38-6.50; range PFS-16d, 0.42-7.66). Three FACIT-F items (items 7-9) displayed signs of misfit (Table 2). Of these, only the chi-square value for item 7 departed substantially from those of the other items in the scale (range items 1-6 and 10-13, 1.04-10.76). These observations were supported by inspection of the ICCs (data not shown).

DIF analyses did not reveal any DIF by disease duration but two instances of uniform DIF by fatigue status. Item 1 of the PFS-16p ("have to rest during the day") and item 9 of the FACIT-F ("need to sleep during the day") were both associated with higher scores, i.e. greater need to rest/sleep during the day, among people classified as non-fatigued. The observed DIF did not appear to bias the total scores, as DIF-adjusted person locations did not differ ($P \geq 0.66$) from the non-DIF-adjusted locations from either the PFS-16p (mean

difference, 0.01 logits) or the FACIT-F (mean difference, 0.07 logits), and the two sets of location estimates correlated strongly (intraclass and Pearson correlations, >0.99 in both instances). Both the DIF-adjusted and non-DIF-adjusted versions of the two scales differentiated between people classified as fatigued and non-fatigued ($P<0.0001$).

Targeting

All three scale scores spanned their full (or almost full) possible score ranges, with average total scores close to scale midpoints (Table 1). Floor/ceiling effects were ≤5% except for the PFS-16d, which had a 19% floor effect (Table 2).

Negative Rasch derived mean person locations (Table 2) suggest that all three scales represent more fatigue than that experienced by the sample. This tendency was more pronounced for the FACIT-F and the PFS-16d (Table 2; Fig. 1). Item thresholds of the PFS-16p were relatively evenly spread along a range of 8.4 logits that roughly covered the sample distribution, except for those with the lowest and highest levels of fatigue (Fig. 1a). The PFS-16d covered a narrower range and displayed notable gaps, i.e. locations along the fatigue continuum not represented by the scale (Fig. 1b). The FACIT-F (Fig. 1c) covered a similar range as the PFS-16p but exhibited gaps toward the lower end of the continuum (representing lower levels of fatigue).

Reliability and measurement precision

Reliability (PSI) was ≥0.86 for all scale scores; the dichotomized PFS-16d exhibited the lowest reliability and the PFS-16p the highest (Table 2). This pattern was enhanced with the separation ratio, i.e. when removing the ceiling effect of the traditional 0-1 reliability range (Table 2). Accordingly, the uncertainty (standard errors) associated with PFS-16d scores

were larger than that associated with PFS-16p and FACIT-F scores (Fig. 2). These data imply that the PFS-16p was able to separate the sample into six distinct strata of people according to their levels of fatigue, whereas the PFS-16d and FACIT-F identified three and four strata, respectively (Table 2).

Response category functioning

All PFS-16p item response categories had thresholds ordered in an expected manner. This applied also for the FACIT-F except for item 7, which showed disordered thresholds between categories 2-to-3 and 3-to-4 (Fig. 3).

DISCUSSION

This study presents a Rasch based head-to-head comparison of the FACIT-F and the PFS-16 in PD. We found general support for the measurement properties of all three tested scale scores, but the PFS-16p displayed consistent advantages, and similarly to previous findings (12) we failed to obtain convincing evidence for the appropriateness of the PFS-16d. Furthermore, anomalies were disclosed that have implications beyond the scales studied here.

We found signs of overall and item level Rasch model misfit for the FACIT-F. Specifically, items 7-9 displayed fit statistics suggesting that they may not represent the same variable as the rest of the scale. In contrast to the rest of the scale, items 7 and 8 are positively worded. Indeed, evidence suggests that positively and negatively worded items are not treated equivalently by respondents and may represent different dimensions (23). This is in accordance with our observations and may also explain why the response categories of FACIT-F item 7 did not work as expected (a similar pattern, albeit marginally ordered, was also found for item 8). Signs of multidimensionality for FACIT-F item 9 (and, to a lesser

degree for PFS-16 item 1) are consistent with previous observations (9), and supports the notion that fatigue is a separate entity from sleepiness (1, 2). Further evidence for this was obtained from our DIF analyses, which showed that people classified as fatigued had a systematically lower probability of reporting that they needed to sleep or rest during the day. These observations imply that the FACIT-F may be improved by rewording so that all items are negatively worded, and that items representing sleepiness should be avoided in fatigue scales.

Targeting and reliability were generally good but raised some concerns regarding the PFS-16d, which exhibited the largest floor effect. This is not surprising since dichotomization of well-functioning response categories leads to loss of information, clustering of people and compromised ability to detect differences. This was illustrated here in that the PFS-16d exhibited compromised reliability and ability to separate the sample into distinct strata, compared to the PFS-16p. PFS-16d item locations also displayed gaps in their coverage of the latent fatigue continuum. This reduces measurement precision (5), which is reflected in enlarged standard errors compared to the original polytomous scoring of the same scale. These findings illustrate the clinical importance of rating scale measurement properties.

The study sample came from a patient association, which introduces potential uncertainties regarding diagnosis and generalizability. However, our observations are similar to those from previous clinic based fatigue studies in PD (2, 12), and the Rasch model is independent of sample distributions (5, 16). Although within general limits for stable estimates (24), another limitation is the relatively small sample size. Sample size requirements in Rasch analysis relate to targeting and increase with poorer targeting (24). In this study, items were well covered by the sample, which increases the confidence in results (5). However, additional

studies in larger samples are warranted for firmer conclusions. Finally, the design prevented us from examining responsiveness and test-retest reliability. However, previous studies have found acceptable test-retest reliability of both the FACIT-F and the PFS-16 in PD (9, 11, 12).

In conclusion, although we found room for improvements this study supports the measurement properties of the PFS-16p, which demonstrates advantages compared to the generic FACIT-F, although there was general support also for the latter. Importantly, and in agreement with earlier observations (12), our data argue against using the dichotomized PFS-16 scoring method.

Disclosures and Acknowledgements

The authors declare that they have no conflict of interest.

The study was supported by the Swedish Research Council, the Swedish Parkinson Academy, the Swedish Parkinson Foundation, the Ribbing Foundation in Lund, the Swedish Council for Working Life and Social Research, and the Faculty of Medicine at Lund University. It was accomplished within the Basal Ganglia Disorders Linnaeus Consortium (BAGADILICO) research group at Lund University, and within the context of the Centre for Ageing and Supportive Environments (CASE) and the strategic research area MultiPark, Lund University, Sweden. Neither of the funding sources was involved in the conduct of the study or development of the submission.

The authors wish to thank the responders for their collaboration and the Swedish Parkinson's Disease Association for assistance with data collection.

Appendix 1 Fatigue rating scales tested in the current study

PFS-16 items		FACIT-F items	
No.	Content (abridged)	No.	Content (abridged)
1	Have to rest during the day	1	Feel fatigued
2	Life restricted by fatigue	2	Weak all over
3	Tired more quickly than other people	3	Listless (“washed out”)
4	One of my three worst symptoms	4	Feel tired
5	Feel completely exhausted	5	Trouble starting things because tired
6	Reluctant to socialise	6	Trouble finishing things because tired
7	Takes longer to get things done	7	Have energy
8	Feeling of heaviness	8	Able to do usual activities
9	Could do more if not tired	9	Need to sleep during the day
10	Everything is an effort	10	Too tired to eat
11	Tired much of the time	11	Need help doing usual activities
12	Totally drained	12	Too tired to do things I want to do
13	Difficult to cope with everyday activities	13	Have to limit social activity
14	Tired even when I haven’t done anything		
15	Do less in my day than I would like		
16	So tired I want to lie down wherever I am		

PFS-16, 16-item Parkinson Fatigue Scale; FACIT-F, Functional Assessment of Chronic Illness Therapy – Fatigue scale.

REFERENCES

1. Friedman JH, Brown RG, Comella C, et al. Fatigue in Parkinson's disease: a review. Mov Disord 2007;22:297-308.

2. Hagell P, Brundin L. Towards an understanding of fatigue in Parkinson disease. J Neurol Neurosurg Psychiatry 2009;80:489-492.

3. Pavese N, Metta V, Bose SK, Chaudhuri KR, Brooks DJ. Fatigue in Parkinson's disease is linked to striatal and limbic serotonergic dysfunction. Brain 2010;133:3434-3443.

4. Friedman JH, Alves G, Hagell P, et al. Fatigue rating scales critique and recommendations by the Movement Disorders Society task force on rating scales for Parkinson's disease. Mov Disord 2010;25:805-822.

5. Hobart J, Cano S. Improving the evaluation of therapeutic interventions in multiple sclerosis: the role of new psychometric methods. Health Technol Assess 2009;13:iii, ix-x, 1-177.

6. Ware JE, Jr., Gandek B. Methods for testing data quality, scaling assumptions, and reliability: the IQOLA Project approach. International Quality of Life Assessment. J Clin Epidemiol 1998;51:945-952.

7. Hobart JC, Cano SJ, Zajicek JP, Thompson AJ. Rating scales as outcome measures for clinical trials in neurology: problems, solutions, and recommendations. Lancet Neurol 2007;6:1094-1105.

8. Massof RW. The measurement of vision disability. Optom Vis Sci 2002;79:516-552.

9. Hagell P, Höglund A, Reimer J, et al. Measuring fatigue in Parkinson's disease: a psychometric study of two brief generic fatigue questionnaires. J Pain Symptom Manage 2006;32:420-432.

10. Grace J, Mendelsohn A, Friedman JH. A comparison of fatigue measures in Parkinson's disease. Parkinsonism Relat Disord 2007;13:443-445.

11. Brown RG, Dittner A, Findley L, Wessely SC. The Parkinson fatigue scale. Parkinsonism Relat Disord 2005;11:49-55.

12. Hagell P, Rosblom T, Palhagen S. A Swedish version of the 16-item Parkinson fatigue scale (PFS-16). Acta Neurol Scand 2012;125:288-292.

13. 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.

14. Hunt SM, McKenna SP, McEwen J, et al. A quantitative approach to perceived health status: a validation study. J Epidemiol Community Health 1980;34:281-286.

15. Saris-Baglama RN, Dewey CJ, Chisholm GB, et al. SF health outcomes™ scoring software user's guide. Lincoln: QualityMetric Inc., 2004.

16. Andrich D. Rasch models for measurement. Beverly Hills: Sage Publications Inc., 1988.

17. Andrich D, Sheridan B, Luo G. Interpreting RUMM2030. Perth: RUMM Laboratory Pty Ltd., 2009.

18. Wright BD, Masters GN. Rating scale analysis. Chicago: MESA Press, 1982.

19. Wann-Hansson C, Klevsgard R, Hagell P. Cross-diagnostic validity of the Nottingham Health Profile Index of Distress (NHPD). Health Qual Life Outcomes 2008;6:47.

20. McHorney CA, Tarlov AR. Individual-patient monitoring in clinical practice: are available health status surveys adequate? Qual Life Res 1995;4:293-307.

21. Andrich D. An index of person separation in latent trait theory, the traditional KR-20 index, and the Guttman scale response pattern. Education Research and Perspectives 1982;9:95-104.

22. Andrich D, Sheridan BE, Luo G. RUMM2030: Rasch Unidimensional Models for Measurement. Perth: RUMM Laboratory Pty Ltd., 2009.

23. Locker D, Jokovic A, Allison P. Direction of wording and responses to items in oral health-related quality of life questionnaires for children and their parents. Community Dent Oral Epidemiol 2007;35:255-62.

24. Linacre JM. Sample size and item calibration (or person measure) stability. Rasch Meas Trans 1994;7:328.

Figure legends

Fig. 1

Distribution of the locations of people (upper panels) and item response category thresholds (lower panels) on the common logit metric (x-axis; positive values = more fatigue). Panel A: PFS-16p; item thresholds range 8.4 logits (from -3.9 to 4.5 logits). Panel B: PFS-16d; item thresholds range 5.4 logits (from -1.9 to 3.5 logits). Panel C: FACIT-F; item thresholds range 9.1 logits (from -4.5 to 4.6 logits)

Fig. 2

Standard errors (SE; y-axes) associated with various total scores (person logit locations; x-axes) on the PFS-16p (panel A), PFS-16d (panel B) and FACIT-F (panel C). It is seen that scores in the middle of the respective ranges have better precision and that dichotomization of PFS-16 scores (panel B; lowest SE, 0.612) results in compromised precision (more measurement error) compared to the original polytomous scoring (panel A; lowest SE, 0.307), which is similar to that of the FACIT-F (panel C; lowest SE, 0.3).

Fig. 3

Response category probability curves depicting the probabilities (y-axis) of observing each response category relative to various person locations on the fatigue continuum (x-axis; positive values = more fatigue), with threshold locations centered at zero. Panel A: Response categories displayed disordered thresholds between categories 2-to-3 and 3-to-4 for FACIT-F item 7 (“have energy”). This suggests that the response categories do not work as expected;

the point where there is a 50/50 probability of responding in either of categories 2 or 3 represents more fatigue than that where there is a 50/50 probability of responding in either of categories 3 or 4. Furthermore, response category 3 is never the most probable response. Panel B: Marginally ordered response category thresholds between categories 1-to-2 and 2-to-3 for FACIT-F item 8 ("able to do usual activities").

Fig. 1

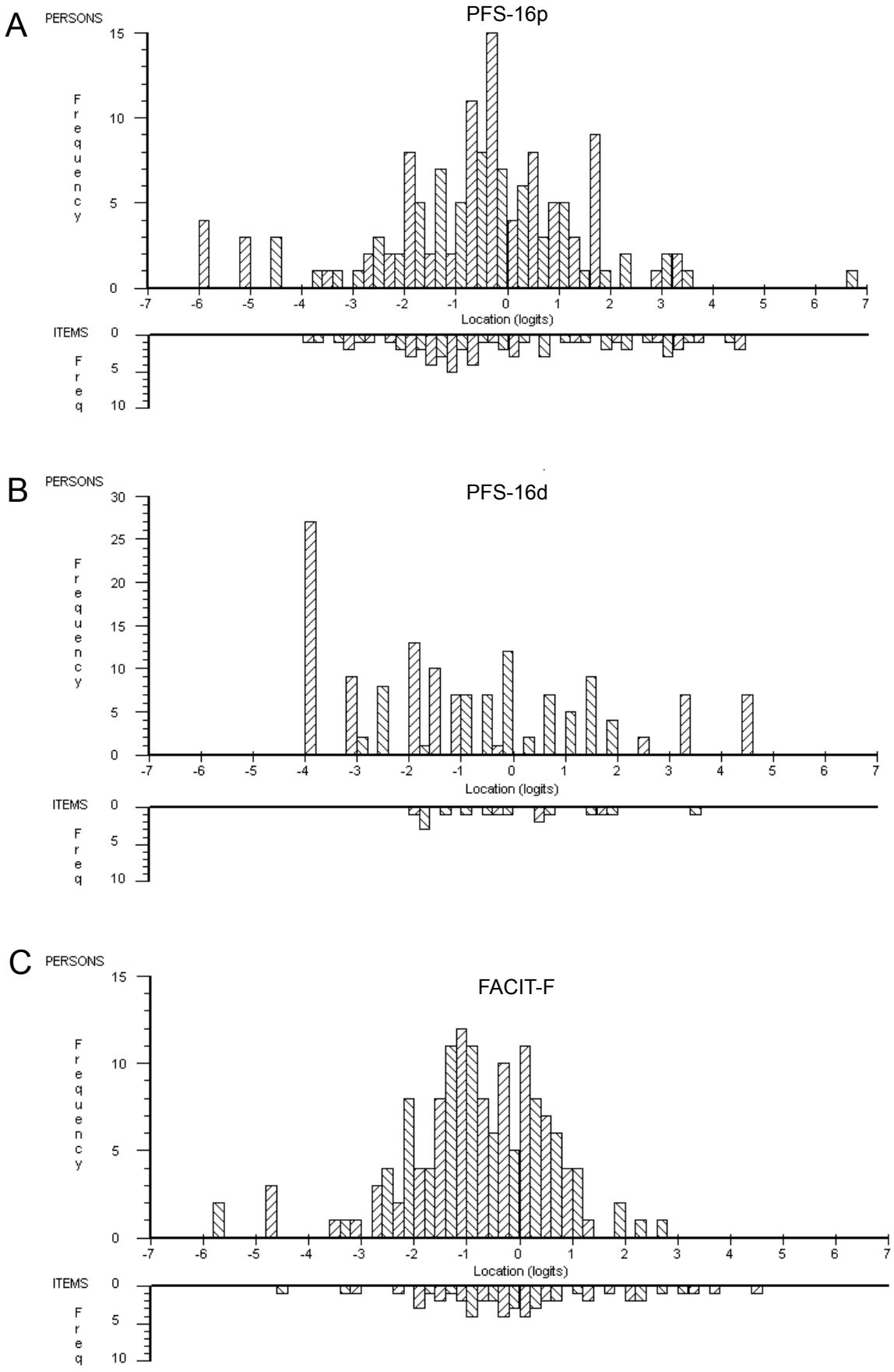

A
PERSONS
PFS-16p
Frequency
Location (logits)
ITEMS
Freq
B
PERSONS
PFS-16d
Frequency
Location (logits)
ITEMS
Freq
C
PERSONS
FACIT-F
Frequency
Location (logits)
ITEMS
Freq

Fig. 2

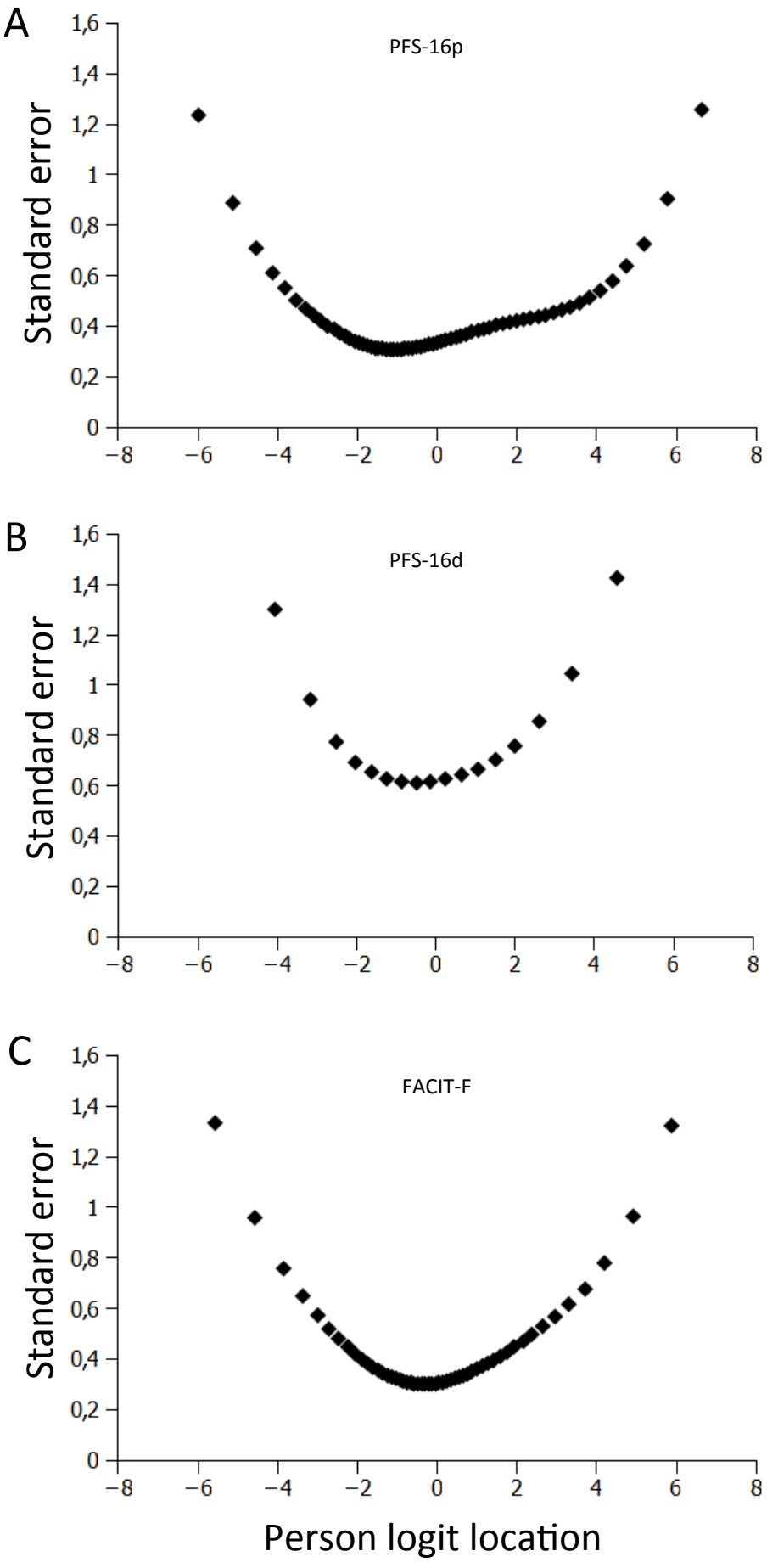
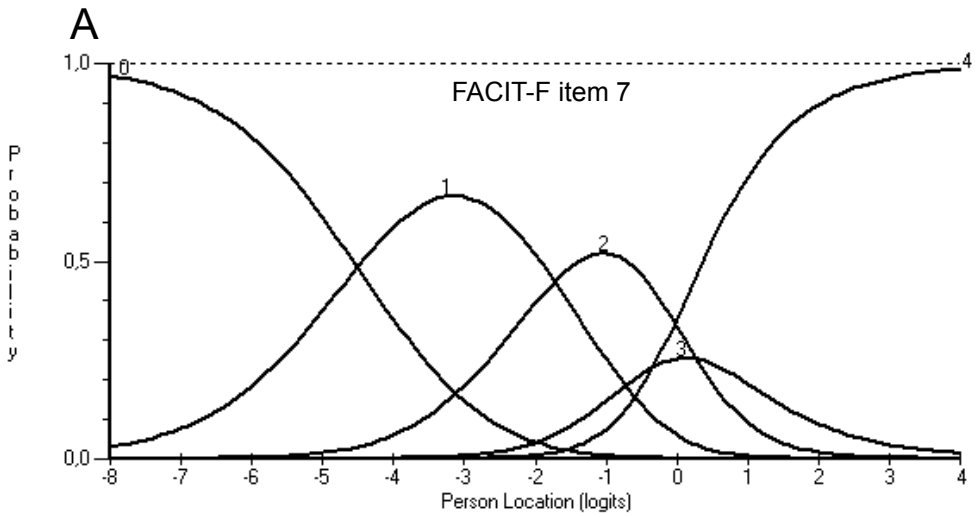

A
PFS-16p
Standard error
1,6
1,4
1,2
1
0,8
0,6
0,4
0,2
0
−8
−6
−4
−2
0
2
4
6
8
B
PFS-16d
Standard error
C
FACIT-F
Standard error
Person logit location

Fig. 3


A
FACIT-F item 7
Probability
1,0
0,5
0,0
0
1
2
3
4
-8
-7
-6
-5
-4
-3
-2
-1
0
1
2
3
4
Person Location (logits)

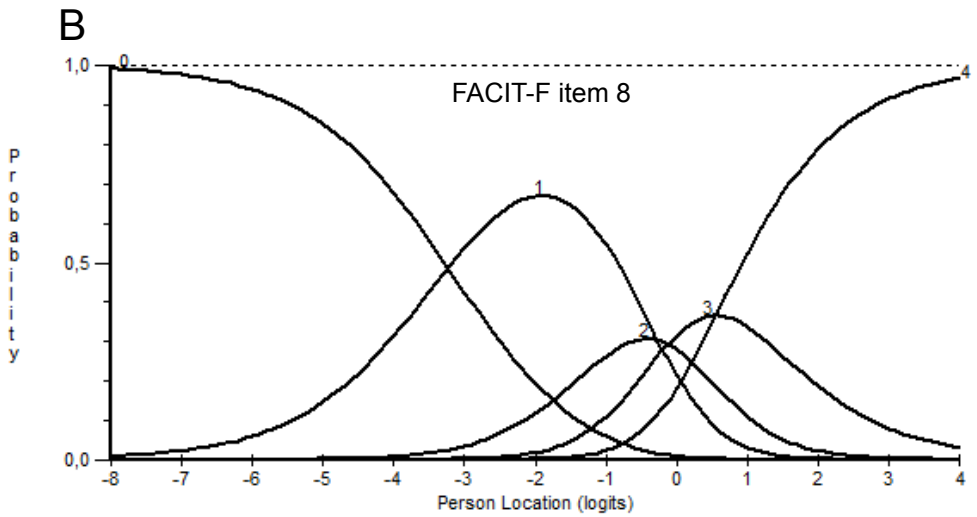

B
FACIT-F item 8
Probability
1,0
0,5
0,0
0
1
2
3
4
-8
-7
-6
-5
-4
-3
-2
-1
0
1
2
3
4
Person Location (logits)

TABLE 1 Sample characteristics (n = 150)

Female, n (%)	70 (47)
Age (years), mean (SD)	70 (9)
PD duration (years), mean (SD)	8 (5)
Taking anti-PD medication, n (%)	147 (98)
Perceived PD-severity, n (%)	
Mild	35 (24)
Moderate	93 (62)
Severe	21 (14)
NHP Energy, median (q1-q3)[a]	33 (0-67)
PFS-16p, mean (SD) / median (q1-q3) / min-max[b]	30.2 (15) / 32 (19-41) / 0-64
PFS-16d, mean (SD) / median (q1-q3) / min-max[c]	6.2 (5.2) / 5 (1-11) / 0-16
FACIT-F, mean (SD) / median (q1-q3) / min-max[d]	20 (10.7) / 18 (12-29) / 0-46

[a] Possible score range, 0-100 (100 = worse).

[b] Possible score range, 0-64 (64 = worse).

[c] Possible score range, 0-16 (16 = worse).

[d] Possible score range, 0-52 (52 = worse).

SD, standard deviation; PD, Parkinson's disease; q1-q3, 1st-3rd quartile; NHP, Nottingham Health Profile; PFS-16p, Parkinson Fatigue Scale (original polytomous scoring); PFS-16d, Parkinson Fatigue Scale (dichotomized scoring); FACIT-F, Functional Assessment of Chronic Illness Therapy - Fatigue Scale.