Qual Life Res (2008) 17:541–548
DOI 10.1007/s11136-008-9341-6

Rasch analysis of the short form 8-item Parkinson's Disease Questionnaire (PDQ-8)

Franco Franchignoni · Andrea Giordano · Giorgio Ferriero

Accepted: 31 March 2008 / Published online: 16 April 2008

Abstract
Purpose To evaluate the Italian version of the 8-item Parkinson's Disease Questionnaire (PDQ-8)—a subset of PDQ-39 (a 39-item health-related quality of life instrument for subjects with Parkinson's Disease [PD])—through classical psychometric techniques and Rasch analysis.
Methods Two convenience samples (100 PD subjects each) were observed consecutively from 2004 to 2006. One group completed the PDQ-8 nested within PDQ-39, the other, the stand-alone PDQ-8.
Results Once verified that the two independent samples came from the same population and showed consistent item calibrations using differential item functioning analysis, the two groups were combined. Cronbach's alpha was 0.72. According to Rasch analysis, the response scale of PDQ-8 could be simplified into a 3-category rating scale. After that, all the PDQ-8 items fitted the construct that the scale was intended to measure. Item separation reliability of PDQ-8 was 0.98 and person separation reliability was 0.70. Principal component analysis on the standardized residuals suggested a minor departure in the data from Rasch criteria (multidimensionality) and some marginal inter-item dependency.
Conclusions The PDQ-8 embedded in the PDQ-39 presented psychometric properties similar to the stand-alone PDQ-8. Our results, while consistent with previous classical psychometric analyses, add information on the meaningfulness of PDQ-8 in people with PD. In particular, a simplification of its rating scale is recommended. Moreover, additional analyses should be performed in order to further check unidimensionality and local dependence, and try to improve item selection and scaling properties of the questionnaire. In order to use the PDQ-8 for clinical decision-making in reference to individuals, its reliability should first be increased.

Keywords PDQ-8 · PDQ-39 · Parkinson's disease · Quality of life · Psychometrics

Introduction

Parkinson's Disease (PD) is a chronic progressive neurodegenerative disease that affects patients in terms of both physical and psychosocial functioning and can lead to increasing disability and a deterioration in quality of life [1]. Thus, assessment of health-related quality of life (HRQoL) is of crucial importance for research on clinical interventions in PD [2, 3].

The Parkinson's Disease Questionnaire (PDQ-39) is the most widely used disease-specific measure of self-perceived health and HRQoL in PD patients [4]. The instrument has been translated into many languages, widely validated both in the UK (where it was developed) and elsewhere, and used in a broad variety of studies and randomized controlled trials [5, 6]. Recently, in order to reduce the respondent burden, a short form of the PDQ-39—the PDQ-8—was created, selecting from each of the eight sub-scales the item with the strongest item-to-total correlation [5]. Studies using classical test theory methods have found that the PDQ-8 is

F. Franchignoni (✉) · G. Ferriero
Unit of Occupational Rehabilitation and Ergonomics, Fondazione Salvatore Maugeri, Clinica del Lavoro e della Riabilitazione, IRCCS, Via Revislate 13, 28010 Veruno, NO, Italy
e-mail: franco.franchignoni@fsm.it

A. Giordano
Bioengineering Service, Salvatore Maugeri Foundation, Clinica del Lavoro e della Riabilitazione, IRCCS, Veruno, NO, Italy

representative of the PDQ-39, both when scores on the short form are nested within the PDQ-39 (PDQ-8/39) [6–8] and when PDQ-8 is scored as an independent instrument [9]. However, some concerns have been raised about the structure of response categories of the PDQ-8 and PDQ-39 [10], and the need expressed for further quantitative evaluations of the psychometric properties of the questionnaires [11].

In recent years, there has been a growing trend to use Rasch analysis to facilitate the development and validation of questionnaires [12]. Rasch analysis provides psychometric information that is not obtainable through classical test theory [13, 14], examining, inter alia, the functioning of rating scale categories, the validity of a measure by evaluating the fit of individual items to the latent trait, and if the pattern of item difficulties is consistent with the expectations of the construct (and hence provides an adequate description of the range and hierarchical relationship of the variable). Indeed, Rasch analysis has been recommended as a method to assess scaling properties in addition to the traditional psychometric criteria used in surveys and questionnaires for disability outcomes research [15].

The purpose of this study was to evaluate the Italian version of the PDQ-8 (both as a part of the PDQ-39 and as an independent instrument) using both classical psychometric techniques and Rasch analysis in order to investigate a wide range of measurement requirements (quality of the rating categories, unidimensionality, construct validity, reliability indexes, etc.).

Materials and methods

Subjects

Participants were two convenience samples of ambulant patients with PD (each group consisting of 100 subjects) consecutively observed at the Scientific Institute of Veruno from 2004 to 2006. Table 1 shows the main demographic, clinical, and functional characteristics of the two groups. The diagnosis of PD was made according to the United Kingdom PD Society brain banking criteria [16]. Patients scoring on the Mini-Mental State Examination (MMSE) below 24 were excluded. All patients were tested in the morning, at 60–120 min after their first morning drug intake, usually corresponding to a time of good performance. Local Ethics Committee approval of the study was obtained. Participants, informed of the experimental protocol, provided written consent prior to participation.

Assessment

PDQ-39 and PDQ-8 Subjects in the first group were assessed by the 39-item Parkinson's Disease Questionnaire (PDQ-39). The PDQ-39 is a disease-specific self-completed instrument designed to measure aspects of health-related quality of life that are relevant to patients with PD [4, 5]. It contains 39 items covering 8 domains: mobility (10 items), activities of daily living (6 items), emotional well-being (6 items), stigma (4 items), social support (3 items), cognition (4 items), communication (4 items), and bodily discomfort (3 items). Questions refer to how often patients have experienced difficulties due to PD in the preceding month. Subjects respond on a 5-point ordinal scale: 0 = *never*; 1 = *occasionally/rarely*; 2 = *sometimes*; 3 = *often*; 4 = *always*. The scores of the 8 items of the PDQ-8 were calculated from the following PDQ-39 answers: # 7 "getting around", # 12 "dressing", # 17 "depression", # 25 "embarrassment in public", # 27 "close relationships", # 31 "concentration", # 35 "inability to communicate", # 37 "cramps". These PDQ-8 scores, nested within PDQ-39, are referred to as PDQ-8/39.

Subjects in the second group were assessed by the distinct 8-item form of the instrument, PDQ-8. In both groups, the scores of the eight items were summed and then expressed as a percentage of the maximum attainable score, thus obtaining a summary index score ranging from 0 to 100. Lower scores indicate better health status.

Impact on Participation and Activity questionnaire Both groups completed the first part of this generic self-report instrument (IPA-I) [17] designed to quantify perceived limitations in participation and autonomy in relation to 25 different life situations across the following subscales: autonomy indoors (7 items), family role (7 items), autonomy

Table 1 Main socio-demographic and clinical characteristics of the study population

	Group 1 ($n = 100$) (PDQ-8/39)	Group 2 ($n = 100$) (PDQ-8)	All ($n = 200$)
Age (in years)[a]	72 (±7)	71 (±8)	72 (±7)
Years of disease[a]	7 (±5)	7 (±5)	7 (±5)
Gender (M/F)	41/59	44/56	85/115
PDQ-8[b]	34 (19–50)	36 (25–53)	34 (22–50)
IPA-I[b]	72 (58–83)	73 (63–84)	73 (61–84)
UPDRS-ADL[b]	16 (12–20)	15 (12–20)	16 (12–20)
UPDRS-ME[b]	23 (18–28)	23 (18–28)	23 (18–28)
HY[b]	3 (2–3)	3 (2–3)	3 (2–3)
SE[b]	80 (70–80)	80 (70–80)	80 (70–80)

[a] Mean (SD)

[b] Median score (25th–75th percentile)

PDQ-8 8-item Parkinson's Disease Questionnaire, *IPA-I* Impact on Participation and Activity questionnaire part I, *UPDRS-ADL* UPDRS part II-Activities of Daily Living, *UPDRS-ME* UPDRS part III-Motor Examination, *HY* Hoehn and Yahr scale, *SE* Schwab and England's ADL scale

outdoors (5 items), social life and relationships (6 items) [18]. For each question there are 5 response options: 0 = *very good*, 1 = *good*, 2 = *fair*, 3 = *poor*, and 4 = *very poor*. The total score was obtained by summing the individual item ratings. Higher scores denote more restriction in participation.

The Unified Parkinson's Disease Rating Scale version 3.0 Part II "Activities of Daily Living" (UPDRS-ADL) consists of 13 items, and part III "Motor Examination" (UPDRS-ME) has 14 items (only the highest score per item was taken into account in items assessing signs in different parts of the body). Each item is rated on a 5-point rating scale (0–4) where higher scores correspond to higher disability or impairment [19].

The modified Hoehn and Yahr scale (HY) A 7-level staging with a higher score indicating more advanced disease [19].

The Schwab and England's ADL scale (SE) This scale reflects the patient's ability to perform daily activities in terms of speed and independence [19]. It is scored 0–100 in 10-point increments, with 0 describing a bedridden patient with altered vegetative functions and 100 a completely independent subject.

A psychologist was available to assist patients in completing the PDQ-39, PDQ-8, and IPA-I when any kind of help was needed (e.g. due to visual problems). The Italian versions of both instruments were obtained using a forward/backward translation method and a procedure of linguistic and stylistic adaptation [20]. The only semantic difficulty found in producing the Italian version of the PDQ-8 was with the response category 'occasionally', which was translated as 'raramente', a term similar to those adopted by the US ('rarely') and Swedish ('sällan') versions [10, 21, 22]. A neurologist recorded the UPDRS-ADL, UPDRS-ME, HY, and SE.

Statistical analysis

Classic test theory statistics Descriptive statistics of the two samples (PDQ-8/39 and PDQ-8) were calculated, including measures of central tendency (mean, median) and spread (standard deviation, 25th–75th percentile).

Then, the two samples were compared in terms of demographics and scores on the assessment instruments (Mann–Whitney U-test) to test the null hypothesis that the two independent samples came from the same population.

Internal consistency of the PDQ-8 (pooled data, see Results) was analysed by calculating Cronbach's coefficient alpha and the item-total correlation. Alpha values $\geq$0.70 are recommended for group level comparison, whereas a minimum of 0.85–0.90 is desirable for individual judgments [23, 24]. For item-total correlation each item should correlate with the total score with r > 0.20 [23, 25].

All correlations in this study were calculated as Spearman's ρ (r_s), corrected for ties.

To test the construct validity of the PDQ-8, we correlated the following variables: PDQ-8 with IPA-I, hypothesizing a good correlation ($r_s > 0.50$) between a disease-specific quality of life questionnaire and a measure of autonomy and participation; and PDQ-8 with clinical PD-specific measures (UPDRS-ADL, UPDRS-ME, HY, and SE), hypothesizing a moderate degree of relationship (r_s = 0.30–0.50).

Rasch analysis Rasch analysis is an original item-response theory based on latent-trait modelling. It provides a statistical model that prescribes how data should be in order to comply with theoretical requirements of measurement and it estimates, amongst other things, how much the modelled measure is supported by the actual observed scores (the so-called "data-model fit") [13, 14]. The matrix of single raw scores for each subject underwent Rasch analysis (rating scale model) using the WINSTEPS software (WINSTEPS Rasch Measurement v. 3.58.1, Linacre JM). The rating scale model specifies that a set of items shares the same rating scale structure and provides average measures and thresholds for categories for the entire instrument. Thresholds—also called step calibrations—are ability levels at which the response to either of two adjacent categories is equally likely.

A sequence of analyses was carried out. As a first step, a differential item functioning (DIF) analysis was performed to search for differences, possibly due to context effects, between the measures obtained using the stand-alone PDQ-8 compared with the PDQ-8 nested within PDQ-39 (PDQ-8/39). Pairwise item-by-item difficulty DIF tests between the two sets were computed (two-sided *t* test for the difference between means, null hypothesis being that the two estimates were the same, except for measurement error) after anchoring on the scale structure obtained for all subjects according to the Rasch model [26, 27]. Bonferroni correction (alpha = 0.05/8) was applied as a protection against type I error.

Having verified that the two independent samples came from the same population and showed consistent item calibrations (i.e. no item bias), we combined the two groups and the pooled data were fitted to the Rasch model (see Results).

To analyse the functioning of the rating scale, the following criteria were used to judge the performance of the response categories [28]: (1) at least ten cases per category; (2) even distribution of category use; (3) monotonic increase in both average measures across rating scale categories and thresholds; (4) category outfit mean square values less than 2 (see the following paragraph); (5)

threshold differences higher than 1.4 logit units and lower than 5. When necessary, categories were collapsed following specific guidelines, and different patterns of categorization were compared, looking not only at the above indicators of category diagnostics but also at the solution maximizing the person separation and reliability indices (see below) [14, 29].

Following this, validity was analysed by evaluating the fit of individual items to the latent trait as per the Rasch model and examining if the pattern of item difficulties was consistent with the model expectancies. Depending on the string of ordinal raw scores, the Rasch model estimates goodness-of-fit (or simply "fit") of the observed data to the model-expected data. If the differences between observed and expected scores are not too large, it is said that "the data fit the model" (see below), and this is seen as equivalent to proving the theoretical construct validity and adequacy of the scale [12, 14]. Information-weighted (infit) and outlier-sensitive (outfit) mean-square statistics (MnSq) for each item were calculated to test if there were items which did not fit with the model expectancies. Both of these fit statistics are expected to approach to 1. In accordance with the literature, we considered MnSq > 0.7 and < 1.3 as an indicator of acceptable fit. Items outside this range were considered underfitting (MnSq ≥ 1.3, suggesting presence of unexpectedly high variability) or overfitting (MnSq ≤ 0.7, indicating a too predictable pattern) [14]. For Rasch analysis it is reported that a sample size of about 100 persons will estimate item difficulty with an alpha of 0.05 to within ± 0.5 logits [30]. The Rasch analysis provides estimates of the level of difficulty achieved by each item ('item difficulty') and of the location of each individual subject along the continuum ('subject ability', representing the global amount of trait in the individual). Item difficulty and subject ability are expressed—on a common interval scale—in logit units, a logit being the natural logarithm of the ratio (odds) of mutually exclusive alternatives (e.g. pass vs. fail, or higher response vs. lower response) [13, 14]. Logit-transformed measures represent linear measures (i.e. the intended amount of the trait). Conventionally, 0 logit is ascribed to the mean item difficulty.

In addition, reliability was evaluated in terms of "separation" (G), defined as the ratio of the true spread of the measures to their measurement error [13, 14]. The item separation index gives an estimate (in standard error units) of the spread or "separation" of items along the measurement construct; the person separation index gives an estimate of the spread or separation of persons along the measurement construct. This index reflects the number of "strata" of measures which are statistically discernible, defined as segments whose centres are separated by distances greater than can be accounted for by measurement error alone (number of distinct strata $= [4G + 1]/3$) [13]. A separation of 2.0 is considered good and enables the distinction of three strata (i.e. groups). A related index is the reliability of these separation indices, providing the degree of confidence that can be placed in the consistency of the estimates. Coefficients range from 0 to 1: coefficients >0.80 are considered as *good* and >0.90 *excellent* [14].

Finally, principal component analysis on the standardized residuals was performed as a further confirmation of the unidimensionality of each scale (proportion of variance attributable to the first residual factor compared with that attributable to Rasch measures) and of the local independence of each item (i.e. the independence of item measures from extraneous variables, once their belonging to the shared construct has been ascertained) [26].

Results

Classic test theory statistics

Median scores of the two groups were similar (PDQ-8/39: 34; PDQ-8: 36). The minimum score of 0 occurred once in both PDQ-8/39 and PDQ-8, but the maximum score of 100 was not observed.

No significant difference was found between the two samples—assessed respectively with PDQ-8 and PDQ-8/39—in terms of demographics (age, years of disease, gender) or scores on the assessment instruments (for each one $P > 0.05$ on the Mann–Whitney U-test).

Following this test and the DIF analysis (see below), the data from PDQ-8/39 and PDQ-8 were pooled and further analyses were performed. In the PDQ-8 (pooled data), the Cronbach's coefficient alpha was 0.72. The item-total correlation coefficients (r_s) ranged from 0.24 to 0.59. The item "...had problems with your close personal relationships" showed the lowest item-total correlation, whereas "...had problems getting around in public" showed the highest. The correlation of PDQ-8 with IPA-I was $r_s = 0.67$ and that with the clinical PD-specific measures ranged from $r_s = 0.38$ (HY) to $r_s = 0.44$ (UPDRS-ADL).

Rasch analysis

Figure 1 shows the hierarchy of item difficulty contrasted across the two groups (PDQ-8/39, PDQ-8) used in the DIF analysis. All items lay within the control band and DIF contrasts ranged from -0.19 to 0.27 ($P > 0.15$, n.s.), showing no or negligible context effects. Under these conditions, it seems reasonable to argue that PDQ-8 and PDQ-8/39 are measuring the same underlying trait and are equivalent. Therefore, the data sets were combined for further analyses.

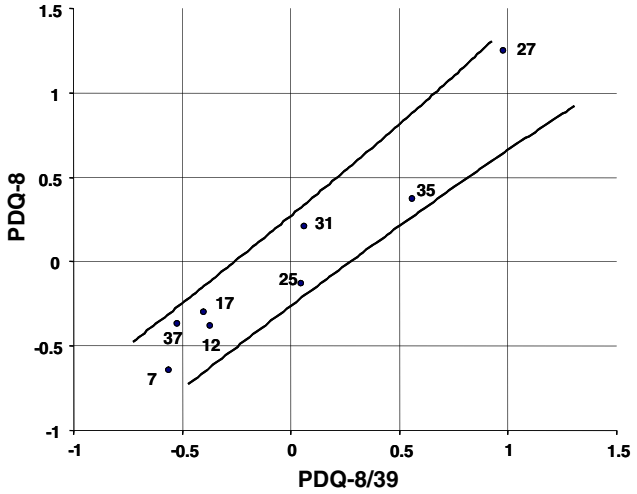

Fig. 1 Analysis of differential item functioning. The hierarchy of item difficulty (in logit units), computed after anchoring on the scale structure obtained on all subjects according to the Rasch model, is contrasted across the two groups (PDQ-8/39, PDQ-8). Items are represented by their number in PDQ-39. The two lines represent the 95% confidence bands along the identity line

The rating scale diagnostics showed that the three central levels (1 = *seldom/rarely*; 2 = *sometimes*; 3 = *often*) did not comply with the criteria for category functioning (average measures, thresholds, etc.). The model meeting all the established criteria and with the best person separation and reliability was the one that collapsed them into a single category, thus obtaining a new 3-level rating scale.

After this rating-scale modification, all eight PDQ-8 items fitted the Rasch model (MnSq between 0.7 and 1.3) (Table 2).

Figure 2 shows the distribution of subjects' ability and item difficulty measures of PDQ-8 after data pooling ($n = 200$) and Rasch transformation of subject and item scores. There was a fairly even spread of items along the variable, and subject ability showed a normal distribution. Ability levels spanned from −3.98 to 1.83 (average measure = −1.16); the levels of HRQoL for our sample were, on average, higher than the mean difficulty of the PDQ-8 items (set by convention at 0 logits). The mean error estimate for the subject ability levels was 0.73. Item difficulty estimates spanned from −1.29 to +2.26 logits (each item estimate can be considered as the balance point for the response distribution across item categories); the range of category step calibrations (thresholds) was approximately from –3.00 to 4.00. Principal component analysis on the standardized residuals showed that: 7.9% of the unexplained variance was explained by the first residual factor (eigenvalue 1.6), whereas the variance explained by the estimated measures was 61%; and in 8 out of 28 cases the correlation between the item residuals was between −0.20 and −0.40. This finding suggests a minor departure in the data from Rasch criteria (in terms of multidimensionality) and some marginal inter-item dependency.

The reliability indices were—after the phase of rating-scale modification—as follows: item separation index was 6.99 and item separation reliability was 0.98; person separation index was 1.53 and person separation reliability was 0.70. The items were distributed into 9 difficulty strata, but the PDQ-8 was able to distinguish only two levels of subject 'quality of life' (good vs. bad) in this study sample.

Discussion

This study is the first validation of the PDQ-8 questionnaire using both classic psychometric test theory and Rasch

Table 2 Difficulty estimates for each of the eight items of PDQ-8 (i.e. the mean value of the difficulty measures of the thresholds along that item), with standard errors (S.E.), infit and outfit mean-square statistics (MnSq), and lower and upper category thresholds ($n = 200$)

	PDQ-8					
	Measure	S.E.	Infit MnSq	Outfit MnSq	Thresholds	
Item					Lower	Upper
... had problems getting around in public? (#7)	−1.29	0.14	0.98	0.96	−2.97	0.39
... had difficulty dressing yourself? (#12)	−0.69	0.14	1.21	1.19	−2.37	0.99
... felt depressed? (#17)	−0.57	0.14	0.81	0.85	−2.25	1.11
... felt embarrassed in public due to having Parkinson's disease? (#25)	0.02	0.14	1.15	1.13	−1.66	1.70
... had problems with your close personal relationships? (#27)	2.26	0.19	1.16	1.06	0.58	3.94
... had problems with your concentration, e.g. when reading or watching TV? (#31)	0.43	0.14	0.88	0.92	−1.25	2.11
... felt unable to communicate with people properly ? (#35)	0.99	0.15	0.84	0.76	−0.69	2.67
... had painful muscle cramps or spasms? (#37)	−0.74	0.14	0.99	1.02	−2.42	0.94

The higher the item estimate, the less likely it is for any subject to gain a high score (i.e. more frequent problems) in that item. Alongside each item is indicated (in brackets) its number in the PDQ-39. For all items the frame question was "Due to having Parkinson's disease, how often during the last month have you..."

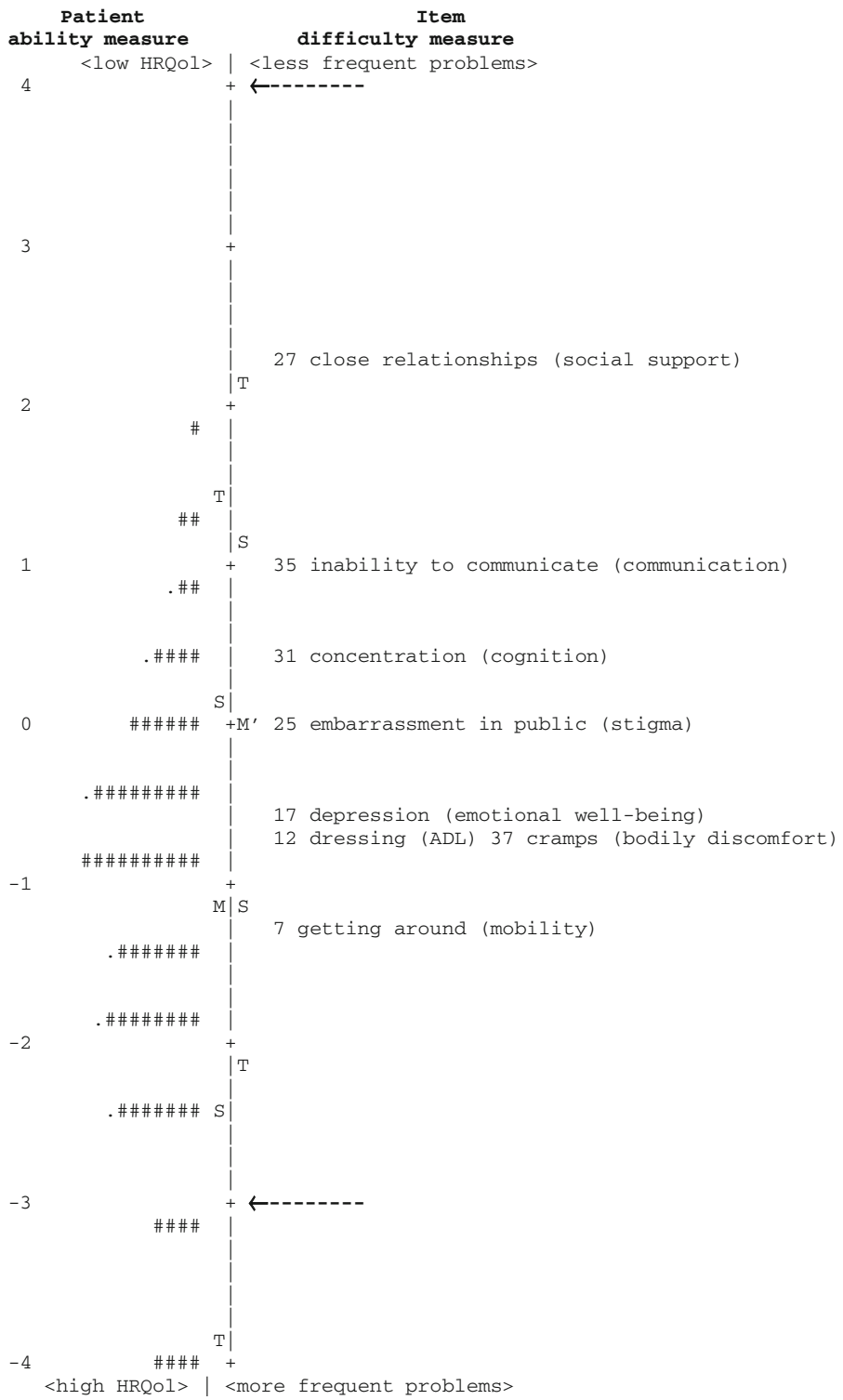

Fig. 2 Subject-ability and item-difficulty maps of the PDQ-8 ($n = 200$). The vertical line represents the measure of the variable, in linear logit units. The *left-hand column* locates the individual's ability along the variable (each '.' is one person and each '#' is three persons). The *right-hand column* locates the 8 item difficulty measures along the variable (for each item, the difficulty estimate represents the mean calibration of the threshold parameters according to the rating scale model). Alongside each item is also indicated its number (and dimension) in the PDQ-39. From bottom to top, measures indicate lower HRQoL (for patients) and lower difficulty/frequency (for items), respectively. By convention, the average difficulty of items in the test is set at 0 logits (and indicated with M′). Accordingly, a candidate with average ability is indicated with M. *Arrows* indicate the highest and the lowest item response category step calibrations

analysis. The absence of differential item functioning between the measures obtained using the stand-alone PDQ-8 compared with the PDQ-8 nested within PDQ-39 (PDQ-8/39) supports the pooling of these two groups and suggests that the studies using either of the two administration forms [6–9] are comparable from a psychometric point of view.

Cronbach alpha values ranged from 0.72 to 0.88 in previous studies on the PDQ-8/39 [6–8] and its value reported in the only study analysing PDQ-8 as a separate instrument was 0.75 [9]. In our group ($n = 200$) the PDQ-8 met the criteria of alpha values above 0.70 and an item-total correlation above 0.20. However, the PDQ-8 showed

an adequate internal consistency only for group decisions, but not for individual judgments [24, 31]. Convergent validity of the PDQ-8 in people with PD was acceptable, given that the predicted associations—and chiefly those with a measure of autonomy and participation—were confirmed in terms of direction and magnitude. In particular, the moderate correlation of the PDQ-8 with UPDRS-ADL, UPDRS-ME, and HY was very similar to that reported by previous studies [7, 9].

According to rating scale diagnostics performed using Rasch analysis [12, 14], the three central response categories ("rarely", "sometimes", and "often") did not comply with the set criteria for category functioning (average measures, thresholds, etc.); this suggests that respondents were unable to appreciably discern between them as indicating different levels of frequency [32]. These findings are in line with previous observations and indicate an inherent problem with the use of the 5-grade, retrospective frequency-related response scale of the PDQ-39 [10, 11]. Similarly, Hagell et al. found many disordered thresholds in these response categories [20]. Our collapsing procedure (producing a new 3-level rating scale: 0 = *never*; 1 = *sometimes*; 2 = *always*) improved the measure, minimizing irrelevant construct variance and ensuring that each rating category represents—for the target population of PD subjects—a clearly distinct level of ability [14, 33]. This adjustment improves the measurement qualities of the scale without substantially decreasing its reliability indexes.

After collapsing the categories, the data were reanalysed in order to calculate fit statistics, extract Rasch-modeled parameters of ability and difficulty, perform DIF and then—after pooling the data from the two groups—examine the validity and reliability issues. Rasch analysis confirmed the general adequacy of the item selection made by the original authors [6]. As an additional demonstration of the internal construct validity of the scale, the PDQ-8 item related to social support was found to be the easiest, whereas the item related to mobility was the hardest one. This general hierarchic arrangement found by Rasch analysis is consistent with clinical expectations. For example, in middle and late HY stages (as in our sample) the PDQ-39 dimension "mobility" usually presents the highest scores (i.e. greater problems with these items) while the PDQ-39 dimension "social support" often produces the lowest ratings (i.e. the lowest level of problems among different dimensions) [4, 8, 34, 35]. No ceiling effect was noted with PDQ-8 measures and the floor effect was negligible (1%).

The high item separation reliability indicated that great confidence can be placed in the replicability of item placement across future samples [14]. Indeed, the item hierarchy was very similar in our two independent samples, as demonstrated by DIF (Fig. 1). Unfortunately, the targeting and spread of item difficulty and the low person separation reliability showed that these eight items are only able to differentiate people with bad vs. good 'quality of life'/'self-perceived health'. Similar targeting problems were recently identified in PDQ-39 [35]. Overall, low Rasch reliability indexes and Cronbach's alpha levels of the PDQ-8 indicate that the instrument seems useful for group decisions but not for everyday clinical application in single patients [23, 24, 31]. If needed, the simplest way to obtain the required level of reliability for individual decisions would be to increase the number of items on the scale [23], e.g. the Spearman-Brown 'prophecy' formula indicates that adding eight more items would raise the alpha value from 0.72 to about 0.84.

Our sample size may cause some limitations to the generalisability of results; on the other hand, the sample represents a wide range of disease severity, duration, and ages, and shows similarities with previously reported international population-based studies using the PDQ-8 [6].

In summary, this study confirms the results of previous classical psychometric analyses of the PDQ-8 [6–9] and adds useful information on the meaningfulness of the PDQ-8 as a disease-specific instrument to measure the degree of perceived HRQoL in people with PD:

1. The PDQ-8 embedded in the PDQ-39 presented psychometric properties similar to the stand-alone PDQ-8.
2. As recently hypothesized [11], Rasch analysis showed that the response scale of the PDQ-8 could be simplified into a 3-category rating scale (0 = *never*; 1 = *sometimes*; 2 = *always*), a format likely to be experienced as less problematic by PD subjects, particularly with advanced impairment levels.
3. PDQ-8 items seem to tap different aspects of a single construct but we think that additional analyses should be performed in order to further check unidimensionality (e.g. through a confirmatory factor analysis using polychoric correlations [36]) and local dependence, and try to improve item selection and scaling properties of the questionnaire.
4. PDQ-8 seems to be a useful measure in studies where a short measure providing an overall index of self-perceived health in Parkinson's disease is required [6], but investigators should be aware that reliability is crucial in planning clinical studies, particularly when using rating scales as clinical trial endpoints [35]. Thus, for clinical decisions regarding individuals, it would be interesting to consider the development of a new short form of the PDQ-39 composed of more items (e.g. 16), that would give it higher reliability [23] with an acceptable respondent burden [11].

We think that the present findings need confirmation in different PD populations and countries, and represent a useful starting point for further psychometric studies, including an analysis of the actual performance of the

3-category response scale, and a study of the stability of item hierarchy across sub-samples defined according to potentially relevant clinical criteria.

References

1. Schrag, A., Jahanshahi, M., Quinn, N. (2000). How does Parkinson's disease affect quality of life? A comparison with quality of life in the general population. Movement Disorders, 15, 1112–1118.
2. Damiano, A. M., Snyder, C., Strausser, B., Willian, M. K. (1999). A review of health-related quality-of-life concepts and measures for Parkinson's disease. Quality of Life Research, 8, 235–243.
3. Marinus, J., Ramaker, C., van Hilten, J. J., Stiggelbout, A. M. (2002). Health related quality of life in Parkinson's disease: a systematic review of disease specific instruments. Journal of Neurology, Neurosurgery, and Psychiatry, 72, 241–248.
4. Jenkinson, C., Fitzpatrick, R., Peto, V., Greenhall, R., Hyman, N. (1997). The Parkinson's Disease Questionnaire (PDQ-39): development and validation of a Parkinson's disease summary index score. Age and Ageing, 26, 353–357.
5. Peto, V., Jenkinson, C., Fitzpatrick, R. (1998). PDQ-39: a review of the development, validation and application of a Parkinson's disease quality of life questionnaire and its associated measures. Journal of Neurology, 245(Suppl 1), S10–S14.
6. Jenkinson, C., Fitzpatrick, R. (2007). Cross-cultural evaluation of the short form 8-item Parkinson's Disease Questionnaire (PDQ-8): results from America, Canada, Japan, Italy and Spain. Parkinsonism & Related Disorders, 13, 22–28.
7. Katsarou, Z., Bostantjopoulou, S., Peto, V., Kafantari, A., Apostolidou, E., Peitsidou, E. (2004). Assessing quality of life in Parkinson's disease: can a short-form questionnaire be useful? Movement Disorders, 19, 308–312.
8. Tan, L. C., Luo, N., Nazri, M., Li, S. C., Thumboo, J. (2004). Validity and reliability of the PDQ-39 and the PDQ-8 in English-speaking Parkinson's disease patients in Singapore. Parkinsonism & Related Disorders, 10, 493–499.
9. Tan, L. C., Lau, P. N., Au, W. L., Luo, N. (2007). Validation of PDQ-8 as an independent instrument in English and Chinese. Journal of the Neurological Sciences, 255, 77–80.
10. Hagell, P., McKenna, S. P. (2003). International use of health status questionnaires in Parkinson's disease: translation is not enough. Parkinsonism & Related Disorders, 10, 89–92.
11. Kim, M. Y., Dahlberg, A., & Hagell, P. (2006). Respondent burden and patient-perceived validity of the PDQ-39. Acta Neurologica Scandinavica, 113, 132–137.
12. Tesio, L. (2003). Measuring behaviours and perceptions: Rasch analysis as a tool for rehabilitation. Journal of Rehabilitation Medicine, 35, 105–115.
13. Wright, B. D., & Masters, G. N. (1982). Rating scale analysis. Chicago: Mesa Press.
14. Bond, T. G., &, Fox, C. M. (2001). Applying the Rasch model: fundamental measurement in the human sciences. Mahwah, NJ: Lawrence Erlbaum Associates.
15. Andresen, E. M. (2000). Criteria for assessing the tools of disability outcomes research. Archives of Physical Medicine and Rehabilitation, 81(Suppl 2), S15–S20.
16. Hughes, A. J., Daniel, S. E., Kilford, L., & Lees, A. J. (1992). Accuracy of clinical diagnosis of idiopathic Parkinson's disease: a clinico-pathological study of 100 cases. Journal of Neurology, Neurosurgery, and Psychiatry, 55, 181–184.
17. Cardol, M., De Haan, R. J., De Jong, B. A., Van Den Bos, G. A., & De Groot, I. J. (2001). Psychometric properties of the Impact on Participation and Autonomy questionnaire. Archives of Physical Medicine and Rehabilitation, 82, 210–216.
18. Franchignoni, F., Ferriero, G., Giordano, A., Guglielmi, V., & Picco, D. (2007). Rasch psychometric validation of the Impact on Participation and Autonomy questionnaire in people with Parkinson's disease. Europa Medicophysica, 43, 451–461.
19. Fahn, S., Elton, R.L.; Members of the UPDRS Development Committee. (1987). Unified Parkinson's Disease Rating Scale. In S. Fahn, C. D. Marsden, D. Calne, M. Goldstein (eds.), Recent developments in Parkinson's Disease II (pp. 153–163). Florham Park, NJ: MacMillan Healthcare Information.
20. Guillemin, F., Bombardier, C., & Beaton, D. (1993). Cross-cultural adaptation of health-related quality of life measures: literature review and proposed guidelines. Journal of Clinical Epidemiology, 46, 1417–1432.
21. Hagell, P., Whalley, D., McKenna, S. P., & Lindvall, O. (2003). Health status measurement in Parkinson's disease: validity of the PDQ-39 and Nottingham Health Profile. Movement Disorders, 18, 773–783.
22. Bushnell, D. M., & Martin, M. L. (1999). Quality of life and Parkinson's disease: Translation and validation of the US Parkinson's Disease Questionnaire (PDQ-39). Quality of Life Research, 8, 345–350.
23. Streiner, D. L., & Norman, G. R. (Eds). (1995). Health measurement scales. A practical guide to their development and use, 2nd edn. Oxford: Oxford University Press.
24. Bland, J. M., & Altman, D. G. (1997). Cronbach's alpha. BMJ, 314 (7080), 572.
25. Portney, L. G., & Watkins, M. P. (2000). Foundations of clinical research: Applications to practice, 2nd edn. Upper Saddle River, NJ: Prentice-Hall Health.
26. Linacre, J. M. (2004). A user's guide to Winsteps. Rasch-model computer programs. Chicago, IL. http://www.winsteps.com/aftp/winsteps.pdf. Retrieved 10 March 2007.
27. Linacre, J. M., & Wright, B. D. (1987). Item bias: Mantel-Haenszel and the Rasch model. http://www.rasch.org/memo39.pdf. Retrieved 30 December 2007.
28. Linacre, J. M. (1999). Investigating rating scale category utility. Journal of Outcome Measurement, 3, 103–122.
29. Zhu, W., Updyke, W.F., & Lewandowski, C (1997). Post-hoc Rasch analysis of optimal categorization of an ordered-response scale. Journal of Outcome Measurement, 1, 286–304.
30. Linacre, J. M. (1994). Sample size and item calibration stability. Rasch Measurement Transactions, 7, 328.
31. Nunnally, J. C., & Bernstein, I. H. (1994). Psychometric theory, 3rd edn. New York: McGraw-Hill.
32. Wolfe, E. W., Smith, E. V. Jr. (2007). Instrument development tools and activities for measure validation using Rasch models: part I – instrument development tools. Journal of Applied Measurement, 8, 97–123.
33. Lopez, W. (1996). Communication validity and rating scales. Rasch Measurement Transactions, 10, 482–483.
34. Ma, H. I., Hwang, W. J., & Chen-Sea, M. J. (2005). Reliability and validity testing of a Chinese-translated version of the 39-item Parkinson's Disease Questionnaire (PDQ-39). Quality of Life Research, 14, 565–569.
35. Hagell, P., & Nygren, C. (2007). The 39-item Parkinson's disease questionnaire (PDQ-39) revisited: implications for evidence-based medicine. Journal of Neurology, Neurosurgery, and Psychiatry, 78, 1191–1198.
36. Reeve, B. B., Hays, R. D., Bjorner, J. B., et al. (2007). Psychometric evaluation and calibration of health-related quality of life item banks: plans for the Patient-Reported Outcomes Measurement Information System (PROMIS). Medical Care, 45(5 Suppl 1), S22–S31.